EIGHTH EDITION

TIETZ FUNDAMENTALS OF

CLINICAL CHEMISTRY and MOLECULAR DIAGNOSTICS

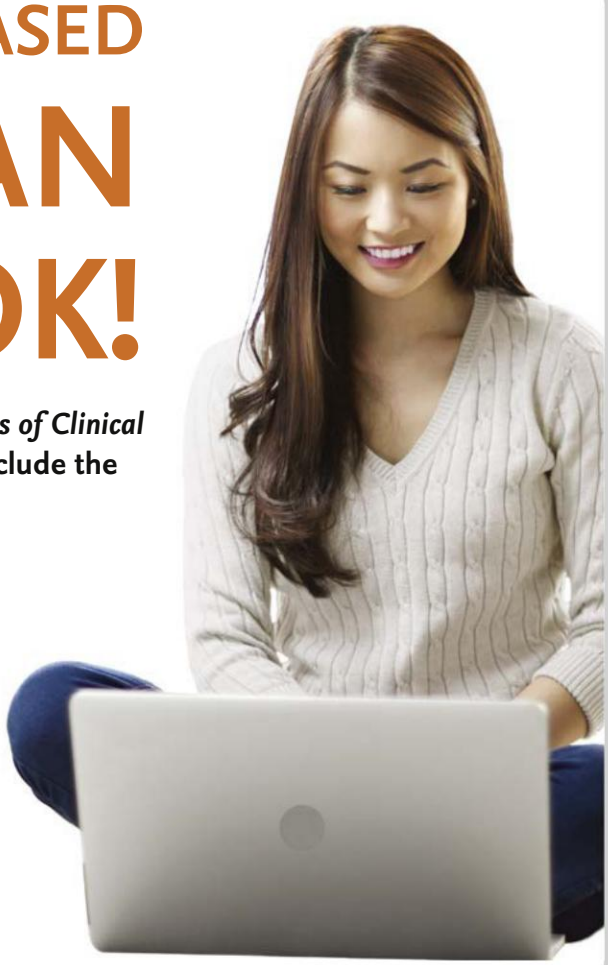
Nader Rifai
Andrea Rita Horvath
Carl T. Wittwer

ELSEVIER

EVOLVE STUDY RESOURCES
FREE WITH TEXTBOOK PURCHASE
EVOLVE.ELSEVIER.COM

Evolve®

YOU'VE JUST PURCHASED MORE THAN A TEXTBOOK!



Evolve Student Resources for *Rifai: Tietz Fundamentals of Clinical Chemistry and Molecular Diagnostics, Eighth Edition*, include the following:

- Answers to Review Questions
- Case Studies
- The Coakley Collection of Clinical Cases
- Deacon's Challenge Biochemical Calculations Collection
- References

Activate the complete learning experience that comes with each NEW textbook purchase by registering with your scratch-off access code at

http://evolve.elsevier.com/Tietz/fundamentals/

If you rented or purchased a used book and the scratch-off code at right has already been revealed, the code may have been used and cannot be re-used for registration. To purchase a new code to access these valuable study resources, simply follow the link above.

Place Sticker Here

REGISTER TODAY!

ELSEVIER

You can now purchase Elsevier products on Evolve!
Go to evolve.elsevier.com/shop to search and browse for products.

2018v1.0

TIETZ FUNDAMENTALS OF

CLINICAL CHEMISTRY and MOLECULAR DIAGNOSTICS

This page intentionally left blank

EIGHTH EDITION

TIETZ FUNDAMENTALS OF

CLINICAL CHEMISTRY and MOLECULAR DIAGNOSTICS

Nader Rifai, PhD
Professor of Pathology
Harvard Medical School
Louis Joseph Gay-Lussac Chair in Laboratory Medicine
Director of Clinical Chemistry
Boston Children's Hospital
Boston, Massachusetts

Andrea Rita Horvath, MD, PhD, FRCPath, FRCPA
Professor (Honorary), School of Public Health
University of Sydney
Professor (Conjoint), School of Medical Sciences
University of New South Wales
Clinical Director, SEALS Department of Clinical Chemistry and Endocrinology
New South Wales Health Pathology
Prince of Wales Hospital
Sydney, Australia

Carl T. Wittwer, MD, PhD
Professor of Pathology
University of Utah School of Medicine
Medical Director, Immunologic Flow Cytometry
ARUP Laboratories
Salt Lake City, Utah

ELSEVIER

ELSEVIER

3251 Riverport Lane
St. Louis, Missouri 63043

TIETZ FUNDAMENTALS OF CLINICAL CHEMISTRY AND
MOLECULAR DIAGNOSTICS, EIGHTH EDITION ISBN: 978-0-323-53044-6

Copyright © 2019 by Elsevier, Inc. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions. This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary. Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility. With respect to any drug or pharmaceutical products identified, readers are advised to check the most current information provided (i) on procedures featured or (ii) by the manufacturer of each product to be administered, to verify the recommended dose or formula, the method and duration of administration, and contraindications. It is the responsibility of practitioners, relying on their own experience and knowledge of their patients, to make diagnoses, to determine dosages and the best treatment for each individual patient, and to take all appropriate safety precautions. To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

International Standard Book Number: 978-0-323-53044-6

Director, Content Development: Laurie Gower
Executive Content Strategist: Kellie White
Publishing Services Manager: Catherine Jackson
Senior Project Manager: Amanda Mincher
Design Direction: Ryan Cook

Printed in the United States of America

Last digit is the print number: 9 8 7 6 5 4 3 2 1

ELSEVIER Book Aid International Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

ASSOCIATE EDITORS

Christina Ellervik, MD, PhD, DMSci
Associate Director of Clinical Chemistry
Boston Children's Hospital
Assistant Professor of Pathology
Harvard Medical School
Boston, Massachusetts
Associate Professor of Clinical Chemistry
Department of Clinical Medicine
Faculty of Health and Medical Sciences
University of Copenhagen
Copenhagen, Denmark

Mark D. Kellogg, PhD, MT(ASCP), DABCC, FAACC
Assistant Professor of Pathology
Harvard Medical School
Director of Quality Programs and Associate Director of Chemistry
Boston Children's Hospital
Department of Laboratory Medicine
Boston, Massachusetts

William Korzun, PhD, DABCC, MT(ASCP)
Associate Professor
Clinical Laboratory Sciences
Virginia Commonwealth University
Richmond, Virginia

Jason Y. Park, MD, PhD
Director
Advanced Diagnostics Laboratory
Department of Pathology
Children's Medical Center Dallas
Associate Professor of Pathology
The Eugene McDermott Center for Human Growth and Development
University of Texas Southwestern Medical Center
Dallas, Texas

Roy W.A. Peake, PhD, FRCPath D(ABCC), D(ABMGG)
Associate Director of Clinical Chemistry
Boston Children's Hospital
Instructor of Pathology
Harvard Medical School
Boston, Massachusetts

This page intentionally left blank

CONTRIBUTORS

Aasne K. Aarsand, MD, PhD
Consultant Medical Biochemist
Norwegian Porphyria Centre and Laboratory of Clinical Biochemistry
Haukeland University Hospital
Consultant Medical Biochemist
Norwegian Quality Improvement of Laboratory Examinations
Haraldplass Deaconess Hospital
Bergen, Norway

Khosrow Adeli, PhD, FCACB, DABCC
Head and Professor, Clinical Biochemistry
The Hospital for Sick Children
University of Toronto
Toronto, Canada

Fred S. Apple, PhD
Medical Director of Clinical Laboratories
Hennepin County Medical Center
Professor of Laboratory Medicine and Pathology
University of Minnesota
Minneapolis, Minnesota

Michael N. Badminton, BSc, MBChB, PhD, FRCPath
Clinical Senior Lecturer, School of Medicine
Cardiff University
Consultant, Medical Biochemistry
University Hospital of Wales
Cardiff, Wales, United Kingdom

Tony Badrick, BAppSc, BSc, BA, MLitSt, MBA, PhD, PhD, FAACB, FAIMS, FRCPA (Hon), FACB, FFScRCPA
Associate Professor, Faculty of Health Sciences and Medicine
Bond University, Robina
Queensland, Australia
Chief Executive Officer
Royal College of Pathologists of Australasia Quality Assurance Programs
Sydney, Australia

Renze Bais, BSc(Hons), PhD, FFSc(RCPA), DipT
Director
RBaisconsulting.com
Queensland, Australia

Lindsay A.L. Bazydlo, PhD, DABCC
Assistant Professor of Pathology
University of Virginia School of Medicine
Director of Toxicology Laboratory
Associate Director of Clinical Chemistry Laboratory
Co-Director of Coagulation Laboratory
University of Virginia Health System
Charlottesville, Virginia

Laura K. Bechtel, PhD, DABCC
CLIA Laboratory Director
Colorado Antiviral Pharmacology Laboratory
University of Colorado Denver-Skaggs School of Pharmacy and Pharmaceutical Sciences
Laboratory Director
Cordant Health Solutions
Denver, Colorado

Marc Berg, MD, FAAP
Clinical Professor of Pediatrics
Division of Pediatric Critical Care Medicine
Stanford University, Lucile Packard Children's Hospital
Palo Alto, California

Roger L. Bertholf, PhD
Professor of Clinical Pathology and Laboratory Medicine
Weill Cornell Medicine
Medical Director of Clinical Chemistry
Department of Pathology and Genomic Medicine
Houston Methodist Hospital
Houston, Texas

D. Hunter Best, PhD
Associate Professor of Pathology
University of Utah School of Medicine
Medical Director, Molecular Genetics & Genomics
ARUP Laboratories
Salt Lake City, Utah

Ingvar Bjarnason, MD, MSc, FRCPath, FRCP, DSc
Professor of Digestive Disease
Consultant Physician and Gastroenterologist
King's College Hospital
London, United Kingdom

Lee M. Blum, PhD
Assistant Laboratory Director/Toxicologist
Quality Assurance
NMS Labs, Inc.
Willow Grove, Pennsylvania

James C. Boyd, MD
Associate Professor (Emeritus)
Department of Pathology
University of Virginia Health System
Charlottesville, Virginia

Maria Domenica Cappellini, MD
Professor of Internal Medicine
Clinical Sciences and Community Health
University of Milan
Ca Granda Foundation-Policlinico Hospital
Milan, Italy

Ferruccio Ceriotti, MD
Laboratory Medicine Service
Ospedale San Raffaele
Milan, Italy

Mark A. Cervinski, PhD
Associate Professor, Pathology and Laboratory Medicine
The Geisel School of Medicine at Dartmouth
Hanover, New Hampshire
Director of Clinical Chemistry, Pathology and Laboratory Medicine
Dartmouth-Hitchcock Medical Center
Lebanon, New Hampshire

Rossa W.K. Chiu, MBBS, PhD, FHKAM, FRCPA
Choh-Ming Li Professor of Chemical Pathology
Department of Chemical Pathology
The Chinese University of Hong Kong
Shatin, New Territories
Hong Kong, China

Nigel J. Clarke, BSc (Hons), PhD
Vice President, Advanced Technology
Quest Diagnostics Nichols Institute
San Juan Capistrano, California

Timothy J. Cole, BSc(Hons), PhD
Associate Professor of Biochemistry and Molecular Biology
Monash University
Clayton, Melbourne
Victoria, Australia

Mark Cooper, BMBCh, PhD, FRCP, FRACP
Professor of Medicine, Endocrinology
Concord Repatriation General Hospital
Professor of Medicine
University of Sydney
Sydney, Australia

Michael P. Delaney, BSc, MD, FRCP, LLM
Consultant Nephrologist, Renal Medicine
East Kent Hospitals University NHS Foundation Trust
Canterbury, Kent, United Kingdom

Sridevi Devaraj, PhD, DABCC, FAACC
Professor of Pathology and Immunology
Baylor College of Medicine
Medical Director, Clinical Chemistry and Point of Care
Texas Children's Hospital
Houston, Texas

Dennis J. Dietzen, PhD
Professor of Pathology and Immunology
Washington University School of Medicine
Medical Director of Laboratory Services
St. Louis Children's Hospital
St. Louis, Missouri

Paul D'Orazio, PhD
Director, Critical Care Analytical Research and Development
Instrumentation Laboratory
Bedford, Massachusetts

Graeme Eisenhofer, PhD
Professor and Chief, Division of Clinical Neurochemistry
Department of Medicine III and Institute of Clinical Chemistry and Laboratory Medicine
University Hospital Carl Gustav Carus Dresden at the Techniche Universität Dresden
Dresden, Germany

Christina Ellervik, MD, PhD, DMSci
Associate Director of Clinical Chemistry
Assistant Professor of Pathology
Harvard Medical School
Boston Children's Hospital
Boston, Massachusetts
Associate Professor of Clinical Chemistry
Department of Clinical Medicine
Faculty of Health and Medical Sciences
University of Copenhagen
Copenhagen, Denmark

Callum G. Fraser, BSc, PhD
Professor, Centre for Research into Cancer Prevention and Screening
University of Dundee
Dundee, United Kingdom

William D. Fraser, BSc (Hons), MB ChB, MD (Hons), FRCP, FRCPath, Eur Clin Chem
Professor of Medicine, Dean of Norwich Medical School
University of East Anglia
Honorary Consultant in Metabolic Medicine/Chemical Pathology
Department of Clinical Biochemistry
Norfolk and Norwich University Hospital Trust
Norwich, Norfolk, United Kingdom

Dr. Danielle B. Freedman, MB, BS, FRCPath, EuSpLM
Consultant Chemical Pathologist and Associate Physician in Clinical Endocrinology
Director of Pathology
Luton and Dunstable University Hospital
Luton, United Kingdom

Jonathan R. Genzen, MD, PhD
Associate Professor, Department of Pathology
University of Utah
Section Chief, Clinical Chemistry
ARUP Laboratories
Salt Lake City, Utah

Jens Peter Goetze, MD, DMSc
Professor, Chief Physician
Department of Clinical Biochemistry
University of Copenhagen, Rigshospitalet
Copenhagen, Denmark

Ann Gronowski, PhD
Professor of Pathology and Immunology
Professor of Obstetrics and Gynecology
Washington University School of Medicine
St. Louis, Missouri

David S. Hage, PhD
Professor, Department of Chemistry
University of Nebraska
Lincoln, Nebraska

David Halsall, PhD
Clinical Director, Blood Sciences
Addenbrooke's Hospital
Cambridge, United Kingdom

Doris M. Haverstick, PhD
Associate Professor of Pathology
University of Virginia
Charlottesville, Virginia

Charles D. Hawker, PhD, MBA, FACSc, FAACC
Adjunct Professor of Pathology (Retired), School of Medicine
University of Utah
Scientific Director
Automation and Special Projects (Retired)
ARUP Laboratories
Salt Lake City, Utah

Victoria Higgins, PhD Candidate
Laboratory Medicine and Pathobiology
University of Toronto
Toronto, Canada

Christopher P. Holstege, MD
Professor, Emergency Medicine and Pediatrics
University of Virginia School of Medicine
Chief, Division of Medical Toxicology
University of Virginia School of Medicine
Medical Director
Blue Ridge Poison Center
University of Virginia Health System
Charlottesville, Virginia

Dave S.B. Hoon, MSc, PhD
Professor, Director
Translational Molecular Medicine
Sequencing Center, John Wayne Cancer Institute
Saint John's Health Center
Providence Health System
Santa Monica, California

Gary L. Horowitz, MD
Director of Informatics
Department of Pathology and Laboratory Medicine
Professor of Pathology
Tufts Medical Center
Boston, Massachusetts

Andrea Rita Horvath, MD, PhD, FRCPath, FRCPA
Professor (Honorary), School of Public Health
University of Sydney
Professor (Conjoint), School of Medical Sciences
University of New South Wales
Clinical Director, SEALS Department of Clinical Chemistry and Endocrinology
New South Wales Health Pathology
Prince of Wales Hospital
Sydney, Australia

John Greg Howe, PhD
Associate Professor of Laboratory Medicine
Yale University School of Medicine
New Haven, Connecticut

Allan S. Jaffe, MD
Consultant in Cardiology and Laboratory Medicine
Professor of Medicine
Chair
CCLS Division of Laboratory Medicine and Pathology
Mayo Clinic and Medical School
Rochester, Minnesota

Graham Ross Dallas Jones, MBBS, BSc(med), DPhil, FRCPA, FAACB
Conjoint Associate Professor, Department of Chemical Pathology, SydPath
St. Vincent's Hospital
Faculty of Medicine
University of New South Wales
Sydney, Australia

Patricia M. Jones, PhD, DABCC, FAACC
Professor of Pathology
University of Texas Southwestern Medical Center
Clinical Director, Chemistry
Department of Pathology
Children's Medical Center
Dallas, Texas

Emily Jungheim, MD, MSCI
Associate Professor of Obstetrics and Gynecology
Washington University School of Medicine
St. Louis, Missouri

Todd William Kelley, MD, MS
Laboratory Director
Navican Genomics, Inc.
Salt Lake City, Utah

Mark D. Kellogg, PhD, MT(ASCP), DABCC, FAACC
Assistant Professor of Pathology
Harvard Medical School
Director of Quality Programs and Associate Director of Chemistry
Boston Children's Hospital, Department of Laboratory Medicine
Boston, Massachusetts

Larry J. Kricka, DPhil, FACB, CChem, FRSC, FRCPath
Professor Emeritus, Department of Pathology and Laboratory Medicine
University of Pennsylvania Medical Center
Philadelphia, Pennsylvania

Mark M. Kushnir, PhD
Senior Scientist
ARUP Institute for Clinical and Experimental Pathology
ARUP Laboratories
Adjunct Assistant Professor of Pathology
University of Utah School of Medicine
Salt Lake City, Utah

Edmund J. Lamb, BSc, MSc, PhD, FRCPath
Consultant Clinical Scientist, Laboratory Medicine
East Kent Hospitals University NHS Foundation Trust
Canterbury, Kent, United Kingdom

James P. Landers, BS, PhD
Professor of Chemistry
University of Virginia
Charlottesville, Virginia

Loralie Langman, PhD
Professor of Laboratory Medicine and Pathology
Mayo Clinic College of Medicine
Rochester, Minnesota

Evi Lianidou, PhD
Professor of Analytical Chemistry–Clinical Chemistry
National and Kapodistrian University of Athens
Athens, Greece

Kristian Linnet, MD, PhD
Professor of Forensic Chemistry
University of Copenhagen
Copenhagen, Denmark

Stanley Lo, PhD
Associate Professor of Pathology
Medical College of Wisconsin
Technical Director, Clinical Chemistry
Pathology and Laboratory Medicine
Children's Hospital of Wisconsin
Milwaukee, Wisconsin

Y.M. Dennis Lo, MA, DM, DPhil, FRCP (Lond. & Edin.), FRCPath, FRS
Li Ka Shing Professor of Medicine
Department of Chemical Pathology
The Chinese University of Hong Kong
Shatin, New Territories
Hong Kong, China

Nicola Longo, MD, PhD
Professor and Chief, Division of Medical Genetics
Department of Pediatrics
Adjunct Professor of Pathology
University of Utah School of Medicine
Co-Director, Biochemical Genetics and Newborn Screening Laboratories
ARUP Laboratories
Salt Lake City, Utah

G. Mike Makrigiorgos, PhD
Professor and Director of Medical Physics and Biophysics
Radiation Oncology, Dana Farber Cancer Institute
Harvard Medical School
Boston, Massachusetts

Elaine R. Mardis, PhD
Professor of Pediatrics, College of Medicine
Ohio State University
Co-Director, Institute for Genomic Medicine, Research Institute
Nationwide Children's Hospital
Columbus, Ohio

William J. Marshall, MA, PhD, MSc, MB, BS, FRCP, FRCPath, FRCPEdin, FRSC, FLS
Emeritus Reader in Clinical Biochemistry
King's College London
London, United Kingdom

Ann McCormack, MBBS Hons I, FRACP, PhD
Head
Hormones and Cancer Group
Garvan Institute of Medical Research
Endocrinologist
St Vincent's Hospital
Conjoint Senior Lecturer of Medicine
University of New South Wales
Sydney, Australia

Christopher McCudden, PhD, DABCC, FACB, FCACB
Clinical Biochemist
The Ottawa Hospital, Department of Pathology and Lab
University of Ottawa, School of Medicine
Ottawa, Ontario, Canada

Gwendolyn A. McMillin, PhD
Professor of Pathology
University of Utah
Medical Director, Toxicology and Pharmacogenomics
ARUP Laboratories
Salt Lake City, Utah

W. Greg Miller, PhD
Professor of Pathology
Virginia Commonwealth University
Richmond, Virginia

Michael C. Milone, MD, PhD
Associate Professor of Pathology and Laboratory Medicine
Perelman School of Medicine of the University of Pennsylvania
Philadelphia, Pennsylvania

Karel G.M. Moons, PhD
Professor of Clinical Epidemiology
Julius Centre for Health Sciences and Primary Care
University Medical Center Utrecht
Utrecht University
Utrecht, The Netherlands

Robert D. Nerenz, PhD, DABCC
Assistant Professor of Pathology and Laboratory Medicine
Dartmouth-Hitchcock Medical Center
Lebanon, New Hampshire

Frederick S. Nolte, PhD, D(ABMM), F(AAM)
Professor and Vice-Chair
Department of Pathology and Laboratory Medicine
Medical University of South Carolina
Charleston, South Carolina

Maurice O'Kane, BSc(Hons), MB ChB, MD, FRCPath, FRCPEdin
Consultant Chemical Pathologist
Clinical Chemistry Laboratory
Altnagelvin Hospital, Western Health and Social Care Trust
Londonderry, Northern Ireland, United Kingdom

Mauro Panteghini, MD
Professor of Clinical Biochemistry and Clinical Molecular Biology
Department of Biomedical and Clinical Sciences 'Luigi Sacco'
University of Milan Medical School
Director, Clinical Pathology Laboratory
ASST Fatebenefratelli-Sacco, Milan
Scientific Coordinator, Research Centre for Metrological Traceability in Laboratory Medicine (CIRME)
University of Milan
Milan, Italy

Jason Y. Park, MD, PhD
Director, Advanced Diagnostics Laboratory
Department of Pathology
Children's Medical Center Dallas
Associate Professor of Pathology
The Eugene McDermott Center for Human Growth and Development
University of Texas Southwestern Medical Center
Dallas, Texas

Marzia Pasquali, PhD, FACMG
Professor of Pathology
University of Utah School of Medicine
Medical Director, Biochemical Genetics and Newborn Screening
ARUP Laboratories
Salt Lake City, Utah

Jay L. Patel, MD, MBA
Associate Professor of Pathology
University of Utah School of Medicine
Medical Director
Hematopathology and Molecular Oncology
ARUP Laboratories
Salt Lake City, Utah

Victoria M. Pratt, PhD
Director, Pharmacogenomics and Molecular Genetics Laboratories
Medical and Molecular Genetics
Indiana University School of Medicine
Indianapolis, Indiana

Christopher P. Price, PhD, FRCPath, FRSC, FACB
Honorary Senior Fellow
Nuffield Department of Primary Care Health Sciences
University of Oxford
Oxford, United Kingdom

Alan T. Remaley, MD, PhD
Senior Investigator
National Institutes of Health
National Heart, Lung and Blood Institute
Bethesda, Maryland

Nader Rifai, PhD
Professor of Pathology
Harvard Medical School
Louis Joseph Gay-Lussac Chair in Laboratory Medicine
Director of Clinical Chemistry
Boston Children's Hospital
Boston, Massachusetts

Alan L. Rockwood, PhD, DABCC
President
Rockwood Scientific Consulting
Professor of Pathology (Retired)
University of Utah School of Medicine
Salt Lake City, Utah

Thomas Røraas, PhD
Mathematician
Norwegian Quality Improvement of Laboratory Examinations
Haraldsplass Deaconess Hospital
Bergen, Norway

Thomas Rosano, PhD
Professor of Pathology
Director of Laboratory Medicine
Albany Medical Center
Albany, New York

William Rosenberg, MA, MB, BS, DPhil, FRCP
Peter Scheuer Chair of Liver Diseases
Institute for Liver and Digestive Health
Division of Medicine
University College London
London, United Kingdom

David B. Sacks, MB, ChB, FRCPath
Adjunct Professor of Medicine
Georgetown University
Clinical Professor of Pathology
George Washington University
Washington, DC
Honorary Professor of Clinical Laboratory Sciences
University of Cape Town
Cape Town, South Africa
Senior Investigator
Chief, Clinical Chemistry
National Institutes of Health
Bethesda, Maryland

Sverre Sandberg, MD, PhD
Director, Norwegian Quality Improvement of Laboratory Examinations
Haraldsplass Deaconess Hospital
Professor, Institute of Public Health and Primary Health Care
University of Bergen
Director, Norwegian Porphyria Centre
Haukeland University Hospital
Bergen, Norway

Mitchell G. Scott, PhD
Professor of Pathology and Immunology
Division of Laboratory and Genomic Medicine
Co-Medical Director, Clinical Chemistry
Washington University School of Medicine
St. Louis, Missouri

Leslie M. Shaw, BS, PhD
Professor of Pathology and Laboratory Medicine
Director, Toxicology Laboratory
Director, Biomarker Research Laboratory
Pathology and Laboratory Medicine
Perelman School of Medicine
University of Pennsylvania
Philadelphia, Pennsylvania

Roy A. Sherwood, BSc, MSc, DPhil
Professor of Clinical Biochemistry
King's College London
London, United Kingdom

Ana-Maria Simundic, PhD
Professor, Specialist in Laboratory Medicine
Department of Medical Laboratory Diagnostics
Clinical Hospital Sveti Duh
Zagreb, Croatia

Ravinder Sodi, PhD, CSci, EuSpLM, FRCPath
Consultant Clinical Biochemist
Department of Blood Sciences
Royal Lancaster Infirmary and Furness General Hospital
University Hospitals of Morecambe Bay NHS Foundation Trust
Honorary Lecturer
Lancaster Medical School
University of Lancaster
Lancaster, United Kingdom

Andrew St. John, PhD
Principal Consultant
ARC Consulting
Perth, Australia

Molly J. Stout, MD, MSCI
Assistant Professor of Obstetrics and Gynecology
Washington University School of Medicine
St. Louis, Missouri

Frederick G. Strathmann, PhD, MBA
Vice President of Quality Assurance
Director of New Technology and Innovation
Assistant Laboratory Director
NMS Labs
Willow Grove, Pennsylvania

Catharine Sturgeon, BSc, PhD, FRCPath
Honorary Lecturer
University of Edinburgh
Consultant Clinical Scientist
Department of Laboratory Medicine
Royal Infirmary of Edinburgh
Edinburgh, United Kingdom

Dorine W. Swinkels, MD, PhD
Professor of Experimental Clinical Chemistry
Department of Laboratory Medicine
Translational Metabolic Laboratory
Radboudumc, Nijmegen, The Netherlands

Sudeep Tanwar, MD, MRCP(UK)
Consultant Gastroenterologist and Hepatologist
Barts Health NHS Trust
London, United Kingdom

Andrew Taylor, PhD, FRSC, FRCPath
Consultant Clinical Biochemist
Royal Surrey County Hospital
Visiting Reader
University of Surrey
Guildford, Surrey, United Kingdom

Mia Wadelius, MD, PhD
Professor of Clinical Pharmacogenetics
Department of Medical Sciences, Clinical Pharmacology
Uppsala University
Uppsala, Sweden

Natalie E. Walsham, MBiochem, MSc, DipRCPath
Lead Clinical Scientist, Department of Clinical Biochemistry
University Hospital Lewisham
London, United Kingdom

Sharon D. Whatley, MSc, PhD, FRCPath
Clinical Scientist
All Wales Genetics Laboratory
Institute of Medical Genetics
University Hospital of Wales
Cardiff, Wales, United Kingdom

Ronald J. Whitley, PhD, DABCC, FAACC
Professor Emeritus, Department of Pathology and Laboratory Medicine
University of Kentucky
Lexington, Kentucky

William E. Winter, MD, DABCC, FAACC, FCAP
Professor of Pathology, Immunology, and Laboratory Medicine, Pediatrics, and Molecular Genetics and Microbiology
University of Florida
Gainesville, Florida

Carl T. Wittwer, MD, PhD
Professor of Pathology
University of Utah School of Medicine
Medical Director, Immunologic Flow Cytometry
ARUP Laboratories
Salt Lake City, Utah

Melanie L. Yarbrough, PhD
Clinical Chemistry Fellow
Department of Pathology and Immunology
Washington University School of Medicine
St. Louis, Missouri

CONTRIBUTORS TO EVOLVE

Mark D. Kellogg, PhD, MT(ASCP), DABCC, FAACC
Assistant Professor of Pathology
Harvard Medical School
Director of Quality Programs and Associate Director of Chemistry
Boston Children's Hospital, Department of Laboratory Medicine
Boston, Massachusetts

Roy W.A. Peake, PhD, FRCPath D(ABCC), D(ABMGG)
Associate Director of Clinical Chemistry
Boston Children's Hospital
Instructor
Harvard Medical School
Boston, Massachusetts

Wenfang S. Wu, PhD
Clinical Chemistry Postdoctoral Research Fellow
Harvard Medical School
Boston Children's Hospital, Department of Laboratory Medicine
Boston, Massachusetts

REVIEWERS

Daniel Bankson, SM, PhD, MBA
Associate Professor
Laboratory Medicine
University of Washington
Seattle, Washington

Janelle M. Chiasera, PhD
Professor and Chair
Senior Executive Associate Dean
School of Health Professions/Clinical and Diagnostic Sciences
The University of Alabama at Birmingham
Birmingham, Alabama

Katherine Davis, MS, MT(ASCP), CLS
Assistant Professor
Department of Clinical Laboratory Science
Senior Assessment Specialist
Academic Dean's Office
Loma Linda University
Loma Linda, California

Paula Deming, PhD
Associate Professor and Chair
Medical Laboratory and Radiation Sciences
Endowed Professor of Health Sciences
University of Vermont
Burlington, Vermont

Karen M. Escolas, EdD, MT(ASCP)
Chair
Department of Medical Laboratory Technology
Farmingdale State College
Farmingdale, New York

Vicky S. Freeman, PhD
Chair and Professor
Clinical Laboratory Sciences
University of Texas Medical Branch
Galveston, Texas

Shawn M. Froelich, MS, MLS(ASCP)CM
Clinical Coordinator
UnityPoint Health
Assistant Professor
Medical Laboratory Science
Allen College
Waterloo, Iowa

Neil Selwyn Harris, MBChB, MD
Clinical Associate Professor
Department of Pathology, Immunology and Laboratory Medicine
University of Florida, College of Medicine
Gainesville, Florida

Kathleen McLoughlin Kenwright, EdD, MLS(ASCP)MBCM
Chair and Associate Professor
Clinical Laboratory Science
University of Tennessee Health Science Center
Memphis, Tennessee

Nadine M. Lerret, PhD, MLS(ASCP)CM
Medical Laboratory Science
Rush University
Chicago, Illinois

Mary Ann McLane, BS, MS, PhD
Professor
Medical Laboratory Sciences
University of Delaware
Newark, Delaware

Nicholas M. Moore, MS, MLS(ASCP)CM
Assistant Professor and Assistant Director
Departments of Pathology and Medical Laboratory Science
Rush University Medical Center
Chicago, Illinois

Michelle Moy, MAdEd, MT(ASCP)SC
Assistant Professor and Program Director Biomedical Sciences
Department of Biological and Health Sciences
Madonna University
Livonia, Michigan

Kathleen Park, MA, MT(ASCP)
Assistant Department Chair
Medical Laboratory Technology
Austin Community College
Austin, Texas

Gerald DeWayne Redwine, PhD, MEd, MT(ASCP)
Associate Professor
Clinical Laboratory Science
Texas State University
San Marcos, Texas

Brady G. Rogers, MSBMS, MT
Medical Laboratory Technician Program Chair
Medical Laboratory Technology
Fortis College
Landover, Maryland

Robin Thompson-McAvoy, MLT, BSc, BEd-Adult
Coordinator of Medical Laboratory Science
Professor of Clinical Chemistry
St. Lawrence College
Kingston, Ontario

PREFACE

We are pleased to introduce the eighth edition of *Tietz Fundamentals of Clinical Chemistry and Molecular Diagnostics.* We built on the excellent work of our predecessors and used electronic tools to produce a state-of-the-art product for students, trainees, and practicing clinical laboratory scientists.

Chapters were extensively updated, and more than 60 new authors were recruited to present the most current and relevant information. We aimed to harmonize the presentation of information among chapters while retaining the personality and unique style of each author, hoping for a readable, educational text.

Unlike most other textbooks, all chapters in this edition were reviewed by three individuals: a reviewer, an associate editor, and a senior editor. We believe that these efforts have led to a better product. In addition, we made a concerted effort to create an *international* rather than an *American* product to reflect different practices from around the world; for example, all measurements are presented both in traditional and SI units.

In addition to the print format of *Fundamentals*, a wealth of supplementary educational materials including clinical case studies, biochemical calculations, multiple-choice questions, and references are available on the Elsevier Evolve platform for an enhanced learning experience.

This project has been a true group effort and represents the collective intellect, knowledge, and experience of approximately 120 leaders in laboratory medicine from 13 countries. We are in debt not only to the authors, reviewers, and editors of the chapters but also to the contributors of the supplementary materials that greatly enrich the product. We are grateful to Elsevier, and particularly to Laurie Gower, for supporting us throughout this project.

We sincerely hope that this product will be a valuable educational and reference resource for the clinical laboratory scientists' community worldwide.

Nader Rifai
Andrea Rita Horvath
Carl T. Wittwer

This page intentionally left blank

CONTENTS

PART I Principles of Laboratory Medicine

1 **Clinical Chemistry and Molecular Diagnostics,** 1
Nader Rifai, Andrea Rita Horvath, and Carl T. Wittwer
2 **Analytical and Clinical Evaluation of Methods,** 7
Kristian Linnet, Karel G.M. Moons, and James C. Boyd
3 **Preanalytical Variation,** 38
Ana-Maria Simundic
4 **Biological Variation,** 51
Sverre Sandberg, Thomas Røraas, Aasne K. Aarsand, and Callum G. Fraser
5 **Establishment and Use of Reference Values,** 64
Gary L. Horowitz and Graham R.D. Jones
6 **Specimen Collection, Processing, and Other Preanalytical Variables,** 77
Doris M. Haverstick and Patricia M. Jones
7 **Quality Management,** 90
W. Greg Miller and Sverre Sandberg
8 **Principles of Basic Techniques and Laboratory Safety,** 108
Stanley Lo

PART II Analytical Techniques and Instrumentation

9 **Optical Techniques,** 127
Larry J. Kricka and Jason Y. Park
10 **Electrochemistry and Chemical Sensors,** 145
Paul D'Orazio
11 **Electrophoresis,** 167
Lindsay A.L. Bazydlo and James P. Landers
12 **Chromatography,** 180
David S. Hage
13 **Mass Spectrometry,** 201
Mark M. Kushnir, Nigel J. Clarke, and Alan L. Rockwood
14 **Enzyme and Rate Analyses,** 215
Renze Bais and Mauro Panteghini
15 **Immunochemical Techniques,** 236
Larry J. Kricka and Jason Y. Park
16 **Automation,** 251
Charles D. Hawker, Jonathan R. Genzen, and Carl T. Wittwer
17 **Point-of-Care Testing,** 273
Andrew St. John, Maurice O'Kane, and Christopher P. Price

PART III Analytes

18 **Amino Acids, Peptides, and Proteins,** 288
Dennis J. Dietzen
19 **Serum Enzymes,** 313
Mauro Panteghini and Renze Bais
20 **Tumor Markers,** 331
Catharine Sturgeon
21 **Kidney Function Tests,** 359
Edmund J. Lamb and Graham R.D. Jones
22 **Carbohydrates,** 377
David B. Sacks
23 **Lipids, Lipoproteins, Apolipoproteins, and Other Cardiac Risk Factors,** 391
Sridevi Devaraj and Alan T. Remaley
24 **Electrolytes and Blood Gases,** 419
Mark A. Cervinski, Mark D. Kellogg, and Mitchell G. Scott
25 **Hormones,** 440
Timothy James Cole
26 **Catecholamines and Serotonin,** 450
Graeme Eisenhofer, Thomas Rosano, and Ronald J. Whitley
27 **Vitamins and Trace Elements,** 466
Ravinder Sodi and Andrew Taylor
28 **Hemoglobin, Iron, and Bilirubin,** 499
Christina Ellervik, Maria Domenica Cappellini, Stanley Lo, and Dorine W. Swinkels
29 **Porphyrins and Porphyrias,** 524
Michael N. Badminton, Sharon D. Whatley, and Aasne K. Aarsand
30 **Therapeutic Drug Monitoring,** 541
Michael C. Milone and Leslie M. Shaw
31 **Clinical Toxicology,** 562
Loralie Langman, Laura K. Bechtel, and Christopher P. Holstege
32 **Toxic Elements,** 590
Frederick G. Strathmann and Lee M. Blum

PART IV Pathophysiology

33 **Diabetes Mellitus,** 603
David B. Sacks
34 **Cardiovascular Disease,** 629
Fred S. Apple, Jens Peter Goetze, and Allan S. Jaffe
35 **Kidney Disease,** 651
Michael P. Delaney and Edmund J. Lamb
36 **Physiology and Disorders of Water, Electrolyte, and Acid-Base Metabolism,** 679
Mark A. Cervinski, Christopher McCudden, Marc Berg, and Mitchell G. Scott
37 **Liver Disease,** 699
William Rosenberg, Tony Badrick, and Sudeep Tanwar
38 **Gastric, Pancreatic, and Intestinal Function,** 723
Roy A. Sherwood, Natalie E. Walsham, and Ingvar Bjarnason
39 **Bone and Mineral Metabolism,** 744
William D. Fraser
40 **Disorders of the Pituitary Gland,** 773
Neil S. Harris, William E. Winter, Ann McCormack, and Roger L. Bertholf
41 **Adrenal Cortex,** 787
Roger L. Bertholf, Mark Cooper, and William E. Winter
42 **Thyroid Disorders,** 811
Danielle B. Freedman, David Halsall, William J. Marshall, and Christina Ellervik

43 Reproduction-Related Disorders, 834
Robert D. Nerenz, Emily Jungheim, and Ann Gronowski
44 Pregnancy and Prenatal Testing, 861
Melanie L. Yarbrough and Ann Gronowski
45 Newborn Screening and Inborn Errors of Metabolism, 882
Marzia Pasquali and Nicola Longo
46 Pharmacogenetics, 898
Gwendolyn A. McMillin, Mia Wadelius, and Victoria M. Pratt

PART V Molecular Diagnostics

47 Molecular Principles, 914
John Greg Howe, Elaine R. Mardis, Jason Y. Park, and Carl T. Wittwer
48 Molecular Techniques, 936
Carl T. Wittwer and G. Mike Makrigiorgos
49 Molecular Applications, 961
Frederick S. Nolte, D. Hunter Best, Todd William Kelley, Jay L. Patel, Evi Lianidou, Dave S.B. Hoon, Rossa W.K. Chiu, and Y.M. Dennis Lo

Appendix, 988
Index, 1029

1

Clinical Chemistry and Molecular Diagnostics

*Nader Rifai, Andrea Rita Horvath, Carl T. Wittwer**

OBJECTIVES

1. Define the following terms:
   - Ethics
   - Laboratory medicine
   - Molecular diagnostics
2. List and explain the reasons for performing a laboratory test.
3. Describe the field of laboratory medicine, including subdisciplines, information handling, and ethical issues.
4. Describe the role of the clinical chemist.
5. Describe the possible career paths for the clinical chemist.
6. State the applications of molecular diagnostics in laboratory medicine.
7. List and explain five ethical issues that confront laboratorians; describe the critical importance of maintaining confidentiality in the laboratory.
8. Evaluate a possible confidentiality or conflict of interest issue and determine whether it is an ethics violation.
9. State the roles of authors, editors, reviewers, and publishers in providing high quality scientific publications.

KEY WORDS AND DEFINITIONS

Ethics Rules or standards governing the conduct of an individual or the members of a profession.

Laboratory medicine A component of laboratory science that is involved in the selection, provision, and interpretation of diagnostic testing of individual specimens.

Laboratory testing A process conducted in a clinical laboratory to rule in or rule out a diagnosis, to select and monitor disease treatment, to provide a prognosis, to screen for a disease, or to determine the severity of and monitor a physiological disturbance.

Molecular diagnostics Use of molecular biology techniques to predict, prevent, diagnose, and monitor disease, including the selection and optimization of therapies.

The disciplines of *clinical chemistry* and *molecular diagnostics* elicit different images. For clinical chemistry, one thinks of pH measurements or large chemistry analyzers, whereas molecular diagnostics conjures up the human genome project, companion diagnostics, and personalized and precision medicine. Although clinical chemistry is at the core of laboratory medicine, molecular diagnostics is a more recent but explosive upstart. Clinical chemistry excels in random access testing, but molecular diagnostics has evolved massively parallel methods. On the surface, these disciplines appear clearly different.

However, consider the meaning behind the words that compose "clinical chemistry" and "molecular diagnostics." Chemistry by its very nature is molecular, and the study of molecules is chemistry. There is no difference here. Perhaps the "molecular" in "molecular diagnostics" suggests complex polymers with meaningful sequence, excluding simpler chemicals. DNA and RNA sequences largely define life, and powerful technologies for nucleic acids now eclipse those for other complex polymers such as proteins and carbohydrates. In common parlance, molecular diagnostics is dominated by nucleic acids. The words "clinical" and "diagnostics" are also similar, connecting both fields to human disease. "Clinical" is more generic than "diagnostics," but again in common use, "molecular diagnostics" includes not only diagnostics but prognosis and genetic predisposition as well. In each

* The authors gratefully acknowledge the contributions by David E. Bruns, François A. Rousseau, and Carl A. Burtis on which portions of this chapter are based.

two-word combination, the sum is greater than its parts, with combined meanings evolving to fit needs and interest. We believe that molecular diagnostics is best viewed as a subset of clinical chemistry.

According to the definition of the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC), "Clinical Chemistry is the largest subdiscipline of Laboratory Medicine which is a multidisciplinary medical and scientific specialty with several interacting subdisciplines, such as hematology, immunology, clinical biochemistry, and others. Through these activities clinical chemists influence the practice of medicine for the benefit of the public."

Clinical laboratories provide in vitro testing of chemical, biochemical, and genetic markers in various fluids or tissues of the human body to screen for a disease, confirm or exclude a diagnosis, help to select or monitor a treatment, or assess prognosis. Laboratory testing impacts health care delivery to virtually every patient.

LOOKING BACK

The examination of body fluids for the diagnosis of disease is certainly not a modern concept. The Greeks noticed before 400 BC that ants are attracted to "sweet urine." However, laboratory testing was not always appreciated by clinicians; the famous Dublin physician Robert James Graves (1796–1853) once remarked, "Few and scanty, indeed, are the rays of light which chemistry has flung on the vital mysteries," and the pioneer Max Josef von Pettenkofer (1818–1901) stated that clinicians use their chemistry laboratory services only when needed for "luxurious embellishment for a clinical lecture." Such views have changed throughout the years, and laboratory testing has proven to be a useful tool to clinicians who have grown to depend and rely on laboratory testing in the routine management of their patients Box 1.1.

Although it may be difficult to pinpoint the exact date when the concept of the clinical laboratory was born, all indications point to the mid-19th century. One such indication is an article titled "Hospital Construction" by Francis H. Brown that was published in the *Boston Medical and Surgical Journal*, the precursor of the *New England Journal of Medicine*, in 1861. Dr. Brown stated, "[Every hospital should have] a small room at the end of the ward to serve as a general laboratory … necessary small cooking might be accomplished here; dishes and other articles washed, etc.; and it would serve as a general store-room for brooms, pails, and other articles." Although Baron Justus von Liebig (1803–73) boasted that his clinical laboratory performed more than 400 tests per annum, the average mid- to large-sized laboratory nowadays performs several million tests yearly. The term *clinical chemistry* was purportedly coined by Charles Henry Ralfe (1842–96) of London Hospital when he used it as the title of his 1883 treatise. The first laboratory attached to a hospital was established in 1886 in Munich, Germany, by Hugo Wilhelm von Ziemssen. In the United States the first clinical laboratory was The William Pepper Laboratory of Clinical Medicine, established in 1895 at the University of Pennsylvania in Philadelphia.

Molecular diagnostics has more recent origins. "Molecular diagnosis" was first mentioned in 1968 as the title of a *New England Journal of Medicine* editorial, commenting on a new inborn error of metabolism that overproduced oxalic acid, resulting in kidney stones. "Molecular" referred to an enzymatic pathway and the substrates, not nucleic acid variants. Twenty years later, additional articles describing "molecular diagnostics" began to appear. In 1986, molecular diagnostics was defined as, "… the detection and quantification of specific genes by nucleic acid hybridization procedures," exemplified by speciation of plant nematodes. In 1987, molecular diagnostics was used to describe mapping of antigenic substances by affinity chromatography using immobilized antibodies. In 1988 the term was used to describe methods for detecting gene amplification and rearrangement using Southern blotting. With the advent of polymerase chain reaction (PCR), the term "molecular diagnostics" became more common, its use doubling in the medical literature every 6 to 7 years. By 1997, commercial real-time PCR instruments solidified "molecular diagnostics" as a branch of clinical chemistry and laboratory medicine.

BOX 1.1 Uses of Testing in the Clinical Laboratory

- Confirming a clinical suspicion (which could include making a diagnosis)
- Excluding a diagnosis
- Assisting in selection, optimization, and monitoring of treatment
- Providing a prognosis
- Screening for disease in the absence of clinical signs or symptoms
- Establishing and monitoring the severity of a physiological disturbance

EXPANDING BOUNDARIES DEFINED BY TECHNOLOGY

Unlike other specialties in laboratory medicine, clinical chemistry is very much influenced and shaped by technology. No discipline in laboratory medicine uses more technologies than clinical chemistry. Technologies that evolved over time not only changed practice but also remodeled the boundaries of the traditional clinical chemistry laboratory. For example, with the emergence of immunochemical techniques in the 1970s, the US Food and Drug Administration approved many tests for the measurement of proteins, small molecule hormones, and drugs, a development that profoundly changed clinical chemistry and its armamentarium of testing. Integrated automated platforms later enabled the measurement of hormones and therapeutic drugs by immunoassays simultaneously with electrolytes, glucose, and other general chemistry tests, thus subsuming the "endocrine lab" and the "drug lab."

Serologic tests for hepatitis and human immunodeficiency virus (HIV) and tests for autoimmune diseases also moved from their traditional home in microbiology and immunology to chemistry analyzers. Immunoglobulin analysis followed a similar path. The typical clinical chemistry laboratory includes testing for general chemistries, specific proteins and immunoglobulins, therapeutic and abused drugs, blood gases, hormones, biogenic amines, porphyrins, vitamins, and trace elements. Testing for inborn errors of metabolism (such as the measurements of amino acids and organic acids), measurements of coagulation factors, general hematologic testing, and serologic assays can belong either to the clinical chemistry laboratory or to another subspecialty, depending on the institution.

Clinical chemists have embraced technology over the years and used it effectively to derive answers to clinical questions. In modern clinical chemistry laboratories, technologies include spectrophotometry, atomic absorption, flame emission photometry, nephelometry, electrochemical and optical sensor technologies, electrophoresis, and chromatography. The influence of automation, information technology, and miniaturization is evident in current clinical chemistry laboratories. Mass spectrometry, once thought of as a research tool, is playing an ever-growing role in clinical chemistry for the measurement of both small molecules and peptides, and more recently proteins. Point-of-care testing is a disruptive innovation that decentralizes laboratory testing and presents the clinical chemist with many challenges and opportunities.

Molecular diagnostics has forever changed virology and microbiology, introducing faster and more sensitive methods based on nucleic acid amplification rather than microbial replication. Nanotechnology, microfluidics, electrical impedance, reflectance spectroscopy, and time-resolved fluorescence are only a few of the technologies used in point-of-care testing for proteins, drugs, DNA, and analysis of metabolites in small samples of whole blood. Molecular diagnostics in particular impacts diverse specialties, including infectious disease, genetics, and oncology, providing new tools for study at a molecular detail never before considered. In summary, the boundaries of clinical chemistry expand with technology, making the profession vibrant, interesting, and ever evolving.

The scope of the profession is constantly changing for the very same reasons. Scientific and technological developments, medical needs, patient demands, and economic pressures bring various disciplines of medicine closer together, and this integration results in more effective health care. For example, companion diagnostics, which help predict therapeutic responses and individualize patient treatment options, bring together pharmacy and medical laboratories. Point-of-care testing in real time with medical intervention breaks the walls of laboratories to bring the profession closer to clinicians and patients. New disruptive technologies (e.g., "lab on a chip," nanotechnology, home monitoring) as well as movement toward patient empowerment and direct-to-consumer testing bring laboratory testing closer to patients. All of these developments present special challenges to the future generations of laboratory professionals both in terms of how they should be trained and how they will practice.

Technology alone is not the answer to more effective and cost-effective clinical practice. The laboratory data obtained must be meaningful and support clinical management decisions. The generation of more data does not necessarily lead to better patient management and outcomes. In the 1960s and 1970s, with the advent of automated clinical analyzers, laboratories reported chemistry panels of 10 to 20 results. More recently, dense data from expression arrays, genome-wide association studies, epigenomics, and microRNA analyses excel in discovery research, but translation to clinical practice has been slower than anticipated. The promise of greater clinical significance with larger data sets seems intuitive, but history suggests caution. Clinical chemists in this world of "big data" translate high-quality measurement *data* into clinically relevant *information*. This information—when integrated with clinical history and presentation, clinical signs, and an understanding of pathophysiology—becomes *knowledge*. Knowledge, in the context of the experience and judgment of the clinician, is converted to *wisdom* that translates to clinical action for improved patient outcomes.

HOW IS CLINICAL CHEMISTRY PRACTICED?

Although the majority of clinical chemists choose a career in a clinical laboratory environment, many work in the in vitro diagnostics (IVD) and pharmaceutical industries. Clinical chemists, by virtue of their training, are translational researchers who are capable of developing, evaluating, and validating biochemical and genetic assays for clinical use; they develop skills that are essential for new biomarker assays, reagent kits, and companion diagnostics. Clinical chemists also provide interfaces between researchers, clinicians, the clinical laboratory, and the IVD industry to help translate biomarker research into clinically meaningful decisions and actions.

Clinical chemists practicing in the IVD or the pharmaceutical industry may not need to routinely interact with clinicians or interpret laboratory results, but they understand and appreciate the clinical utility and relevance of the assays and companion diagnostics they are developing and thus contribute more effectively to the development of diagnostics that improve health. The daily practice of the profession has changed over time. In the 1960s and 1970s, clinical chemists developed laboratory tests. At present, de novo assay development is still active only in certain areas such as chromatography, mass spectrometry, and molecular diagnostics.

However, as the profession matured and the instrumentation changed from open systems to "black boxes," the traditional analytical focus of the profession has significantly diminished. Clinical chemists are now more active in the preanalytical and postanalytical phases of testing and in establishing processes such as how best to select the right test for the right patient and to communicate test results to clinicians in a medically meaningful way, how to build laboratory processes that reduce error, and how to continuously improve the quality of laboratory practices (Box 1.2).

BOX 1.2 Functions of the Laboratory Professional

- Develop and validate de novo laboratory tests to meet clinical needs.
- Evaluate and characterize the analytical and clinical performance of laboratory tests.
- Present laboratory results to clinicians in an effective manner.
- Provide education and advice on the selection and interpretation of laboratory tests as part of the clinical team.
- Determine the cost effectiveness and intrinsic value of laboratory tests.
- Participate in the development of clinical testing algorithms and clinical practice guidelines.
- Ensure compliance with regulatory requirements.
- Participate in quality assurance and improvement of the laboratory service.
- Teach and train future generations of laboratory specialists.
- Participate in basic or clinical research.

In the current health care environment, there is increasing emphasis on clinical impact and cost effectiveness. Laboratories are expected to demonstrate evidence of improved measurable clinical outcomes and the usefulness and added value of tests to clinical decision making. Proving the fact that laboratory testing contributes to improved patient outcomes is challenging because the relationship between testing and clinical outcomes is mostly indirect. Nevertheless, clinical chemists should move away from being just providers of high-quality data. Transforming laboratory data to information and knowledge requires more skills in information and information management technology, evidence-based medicine, epidemiology, data mining, and translational research. It also requires a shift of thinking from essentialism to consequentialism and from technology-driven to customer-focused and patient-centered laboratory medicine.

To summarize, today's clinical chemists are laboratory professionals who are trained in pathophysiology and technology. The execution of their daily duties, which are more clinically or technology oriented, is influenced by their training (such as MD vs. PhD), interests, institutional needs, and the country where they practice. Clearly the practice of our profession has evolved over the past half a century, and there are even more challenges on the horizon that will expand and change its scope and role and enhance its diversity.

GUIDING PRINCIPLES OF PRACTICING THE PROFESSION

As in other branches of medicine, practitioners in the clinical laboratory are faced with ethical issues, often on a daily basis; examples are listed in Box 1.3.

CONFIDENTIALITY OF GENETIC INFORMATION

Confidentiality of genetic information has been prominent in the news in the first and second decades of this millennium.

BOX 1.3 Ethical Issues in Clinical Chemistry and Molecular Diagnostics

- Confidentiality of genetic information
- Confidentiality of patient medical information
- Allocation of resources
- Codes of conduct
- Publishing issues
- Conflicts of interest

Legislation was considered necessary to prevent denial of health insurance or employment to people found by DNA testing to be at risk of disease. Less appreciated is the fact that the issue of confidentiality of clinical laboratory data predated DNA testing. In fact, many non-DNA tests, old and new, also carry information about risks of illness and death. Clinical laboratory professionals have long been responsible for maintaining the confidentiality of all laboratory results, a situation made even more critical with the advent of increasingly powerful genetic testing.

CONFIDENTIALITY OF PATIENT MEDICAL INFORMATION

New test development requires the use of patient samples and may involve the use of patient medical information. Ethical judgments are required regarding the type of informed consent that is needed from patients for use of their samples and clinical information. Clinical laboratory physicians and scientists often serve on institutional review boards that examine proposed research on human subjects. In these discussions, ethical concepts such as clinical equipoise—which refers to the genuine uncertainty in the expert medical community over whether a particular treatment will be beneficial—and confidentiality are central to decisions.

ALLOCATION OF RESOURCES

Because resources are finite, laboratory professionals must make ethically responsible decisions about allocation of resources. There is often a trade-off between cost and quality. What is best for patients generally? How can the most good be done with the available resources? For laboratorians in business, creative accounting may tarnish the profession if patient care is not kept paramount.

CODES OF CONDUCT

Most professional organizations publish a code of conduct that requires adherence by their members. For example, the American Association for Clinical Chemistry (AACC) has published ethical guidelines that require AACC members to endorse principles of ethical conduct in their professional activities, including (1) selection and performance of clinical procedures, (2) research and development, (3) teaching, (4) management, (5) administration, and (6) other forms of professional service.

PUBLISHING ISSUES

Publication of documents having high scientific integrity depends on editors, authors, and reviewers all working in concert in an environment governed by high ethical standards.

Editors are responsible for the overall process, including identifying reviewers, evaluating the reviews and the authors' response to them, and making the final decision of whether to accept or reject a manuscript. Editors are also responsible for establishing policies and procedures to ensure consistency in the editorial process. Finally, the editor-in-chief is responsible for developing a conflict of interest policy and monitoring it among his or her editors. Publishers, being commercial or scientific societies, should monitor any conflicts of interest of the editor-in-chief.

Authors are responsible for honest and complete reporting of original data produced in ethically conducted research studies. Practices such as fraud, plagiarism (verbatim, mosaic), and falsification or fabrication of data are unacceptable. The International Committee of Medical Journal Editors and the Committee on Publication Ethics have published policies that address such behavior. Other practices to be avoided include duplicate publication, redundant publication, and inappropriate authorship credit. In addition, ethical policies require that factors that might influence the interpretation of study findings must be revealed, such as (1) the role of the commercial sponsor in the design and conduct of the study, (2) interpretation of results, and (3) preparation of the manuscript. Additional undesirable and harmful practices are publication bias and selective reporting, in which only studies with positive findings are reported and authors use "data dredging" and meaningless subanalyses to find a positive association rather than reporting the original hypothesis that was negative. These practices inflate the actual value of observations or utility of markers and diminish the quality of meta-analyses. As a result, a comprehensive registry of diagnostic and prognostic studies, similar to the registry of clinical trials, has been advocated.

Reviewers must provide a timely, fair, and impartial assessment of manuscripts. They must maintain confidentiality and never contact the authors until after the publication of the report. Finally, reviewers must excuse themselves from the review process if they perceive a conflict of interest.

Most journals now require authors to complete conflict of interest forms and delineate each author's contribution. Some journals, including *Clinical Chemistry,* publish this information along with the article for enhanced transparency.

CONFLICTS OF INTEREST

The interrelationships between practitioners in the medical field and commercial suppliers of drugs, devices, and equipment can be positive or negative. Concerns led the National Institutes of Health in 1995 to require official institutional review of financial disclosure by researchers and management in situations when disclosure indicates potential or actual conflicts of interest. In 2009 the Institute of Medicine issued a report that questioned inappropriate relationships between pharmaceutical device companies and physicians and other health care professionals. Similarly, the relationship between clinical laboratory professionals and manufacturers and providers of diagnostic equipment and supplies can be scrutinized.

As a consequence of these concerns and as a result of the enactment of various laws designed to prevent fraud, abuse, and waste in Medicare, Medicaid, and other health care reimbursement programs, professional organizations that represent manufacturers of IVD and other device and health care companies have published codes of ethics. For example, the Advanced Medical Technology Association has published a revised code of ethics that became effective on July 1, 2009. Topics discussed in this revised code include gifts and entertainment, consulting arrangements and royalties, reimbursement for testing, and education. Similarly, the European Diagnostic Manufacturers Association has published a code of ethics, including issues of member-sponsored product training and education, support for third-party educational conferences, sales and promotional meetings, consultants, gifts, provision of reimbursements, and donations for charitable and philanthropic purposes. Both documents address demands from regulators while nurturing the unique role that laboratory and other health care professionals play in developing and refining new technology.

WHAT IS IN THIS TEXTBOOK?

In this textbook, we have assembled what is essential to effectively practice clinical chemistry and molecular diagnostics. We begin with introductory chapters that describe the basics of laboratory medicine, including statistics, sample handling, preanalytical processes, reference intervals, and quality control. This is followed by a section on analytical techniques and applications, describing the main methods used in clinical chemistry, including immunoassays, mass spectrometry, and point-of-care testing. Next, all the major analytes are discussed, including enzymes, tumor markers, therapeutic drugs, and toxicology, among many others. This is followed by a section on pathophysiology that covers disease states and malfunction of different organ systems that correlate with abnormal laboratory findings. Finally, our last section is dedicated to molecular diagnostics, perhaps the fastest growing field in clinical chemistry. An appendix tabulates reference intervals for the clinical laboratory. The online version includes clinical cases, podcasts, and biochemical calculations. Our aim is to provide current scientific and practical knowledge to support laboratory professionals with knowledge resources that interface between science and technology on the one hand and the clinician and the patient on the other.

POINTS TO REMEMBER

- Clinical chemistry is the largest subdiscipline of laboratory medicine, and molecular diagnostics is a subset of clinical chemistry.
- Clinical chemistry is a profession that has been shaped and defined by technology.
- The role of clinical chemists evolved over time from analytically and technology focused to customer and patient centered.
- Clinical chemists are translational researchers who convert laboratory data to clinical knowledge.
- Career paths of clinical chemists are heterogeneous and include work in clinical laboratories and in vitro diagnostics and pharmaceutical industries.
- Clinical chemists must adhere to guiding principles of practicing the profession, which include maintaining confidentiality of genetic and medical information, using resources appropriately, abiding by codes of conduct, following ethical publishing rules, and managing and disclosing conflict of interest.

REVIEW QUESTIONS

1. The clinical laboratory discipline that is used most often to assess inherited disease through study of the constitutive genome is:
   a. transfusion services.
   b. clinical chemistry.
   c. molecular diagnostics.
   d. hematology.
2. When a practitioner in clinical chemistry has an inappropriate personal relationship with a commercial supplier of chemistry analyzers, there may be a potential issue with:
   a. publication ethics.
   b. confidentiality.
   c. selection of treatment.
   d. conflict of interest.
3. "Molecular testing" involves the clinical analysis of:
   a. atoms and molecules.
   b. nucleic acids.
   c. cellular components of blood.
   d. the physical structure of compounds.
4. Which one of the following is not considered an ethical issue facing a clinical laboratorian?
   a. Allocation of resources
   b. Conflicts of interest
   c. Discussion of one's salary
   d. Maintenance of confidentiality

SUGGESTED READINGS

Altman, D. G. (2014). The time has come to register diagnostic and prognostic research. *Clinical Chemistry*, *60*, 580–582.

Annesley, T. M., Boyd, J. C., & Rifai, N. (2009). Publication ethics: clinical chemistry editorial standards. *Clinical Chemistry*, *55*, 1–4.

Hallworth, M. J., Epner, P. L., Ebert, C., et al. (2015). Current evidence and future perspectives on the effective practice of patient-centered laboratory medicine. *Clinical Chemistry*, *61*(4), 589–599.

McMurray, J., Zerah, S., Hallworth, M., et al. (2010). The European Register of Specialists in Clinical Chemistry and Laboratory Medicine: guide to the Register, version 3-2010. *Clinical Chemistry and Laboratory Medicine*, *48*, 999–1008.

Rifai, N., Annesley, T., & Boyd, J. (2010). International year of chemistry 2011: *Clinical Chemistry* celebrates. *Clinical Chemistry*, *56*, 1783–1785.

2

Analytical and Clinical Evaluation of Methods

Kristian Linnet, Karel G.M. Moons, James C. Boyd

OBJECTIVES

1. Define the following:
   - Analytical measurement range
   - Analytical specificity
   - Bias
   - Coefficient of variation
   - Correlation coefficient
   - Difference curve
   - Error model
   - Frequency distribution
   - Gaussian probability distribution
   - Limit of detection
   - Linearity
   - Mean
   - Median
   - Population
   - Precision
   - Random error
   - Random sample
   - Regression analysis
   - Sample
   - Standard deviation
   - Systematic error
   - Student *t* distribution
   - Trueness
   - Uncertainty
2. List and describe three criteria that must be considerations in laboratory method selection, including the specific parameters involved in each criterion.
3. Compare population and sample mean, population parameter and sample statistic, and population standard deviation and sample standard deviation, including a description of each, symbols used to express these, how they are calculated, and the information they provide.
4. State the connection of the following concepts to analytical methods:
   - Accuracy
   - Analytical sensitivity
   - Analytical specificity
   - Calibration
   - Limit of detection
   - Linearity
   - Precision
   - Repeatability
   - Reproducibility
5. List two common approaches used to objectively analyze data in a methods comparison study.
6. Describe the components of a difference plot, including the plot's use in method comparison and how the plot is interpreted.
7. Discuss assessment of error in an objective analysis of data in method comparison, including how error occurrence relates to an assay's performance characteristics, the difference between random and systematic error, what causes error, and how error is evaluated in a difference plot.
8. For the following types of analyses, list the components of the analysis, its application in method comparison, how it is computed, how outliers affect it, and how the results are interpreted:
   - Deming regression
   - Nonparametric regression
   - Ordinary least-squares regression
   - Regression
9. Describe the calibration hierarchy, including the tracing of values of routine clinical chemistry measurements to a primary reference, how the values are obtained, and the methods involved; draw a calibration hierarchy given a specific analyte.
10. Discuss the concept of uncertainty in relation to clinical laboratory results, including the components of the standard uncertainty formula and two ways in which uncertainty is assessed.
11. Given appropriate values, state the formula and calculate the following:
    - Coefficient of variation
    - Coefficient of variation percent
    - Deming regression
    - Linear regression
    - Population mean
    - Precision analyses
    - Standard deviation
    - Standard uncertainty

- Likelihood ratio
- Odds ratio
- Predictive value
- Prevalence
- Receiver operating characteristic curve
- Sensitivity
- Specificity
- True/false positive
- True/false negative

12. State the formulas for and calculate, given appropriate information, the following: sensitivity, specificity, predictive value for positive/negative tests (posterior probabilities), odds ratio, and positive/negative likelihood ratio.
13. State the relationship between high sensitivity and false negatives; state the relationship between high specificity and false positives.
14. Compare dichotomous and continuous tests; include definition, sensitivity/specificity, and a clinical example of each type of test.
15. State how the predictive value of a laboratory test (posterior probability) is affected by prevalence.
16. Construct and interpret a receiver operating characteristic curve using data from a diagnostic test study.
17. Describe the added value of combination testing as it is used in the clinical laboratory; include examples, diagnostic usefulness, and associated problems.

KEY WORDS AND DEFINITIONS

Accuracy of analytical method Closeness of agreement of a single measurement with "true value."

Analyte The substance being analyzed in an analytical procedure.

Analytical measurement range The analyte concentration range over which measurements are within the declared tolerances for imprecision and bias; also referred to as *reportable range.*

Analytical sensitivity The ability of an analytical method to assess small variations in the concentration of analyte.

Analytical specificity The ability of an assay procedure to determine specifically the concentration of the target analyte in the presence of potentially interfering substances or factors in the sample matrix.

Assay comparison Comparison of measurements by two methods that is carried out objectively using statistical procedures and graphics displays.

Bias In an analytical method, the difference between the average value and the true value that is expressed numerically and is inversely related to the trueness.

Calibration In relation to analytical methods, a function that describes the relationship between instrument signal and concentration of analyte.

Coefficient of variation Relative standard deviation.

Commutability The equivalence of the mathematical relationships between the results of different measurement procedures for a reference material and for representative samples from healthy and diseased individuals.

Correlation coefficient Measure of association between two variables.

Deming regression Least-squares regression analysis taking measurement errors in both variables into account.

Difference plot A bias plot that shows the dispersion of observed differences between the measurements of two methods as a function of the average concentration of the measurements; also referred to as a *Bland-Altman plot.*

Error model A model of the error structure.

Frequency distribution A distribution of the frequency (ordinate) as a function of the variable value (abscissa), that is, a histogram of absolute or relative frequencies.

Gaussian probability distribution Bell shaped relative frequency distribution described under basic statistics.

Likelihood ratio The probability of occurrence of a specific test value given that the disease is present divided by the probability of the same test value if the disease was absent.

Linearity Range of values for which there is a linear relationship between concentration and signal.

Limit of detection An assay characteristic defined as the lowest value that significantly exceeds the measurements of a blank sample.

Matrix In relation to analytical methods, human serum that contains analytes.

Mean Arithmetic average of variables. See Basic Statistics.

Median Equal to the 50th percentile of a set of variables. See Basic Statistics.

Negative predictive value The proportion of subjects with a negative test who do not have the disease.

Odds ratio The probability of the presence of a specific disease divided by the probability of its absence.

Ordinary least-squares regression (OLR) analysis A method used to estimate the unknown parameters in a linear regression assessment performed to minimize the sum of squared vertical distances between observed responses and responses predicted by linear approximation.

Population In relation to analytical methods, the complete set of all observations that might occur as the result of performing a particular procedure according to specified conditions.

Positive predictive value The proportion of subjects with a positive test who have the disease.

Precision The closeness of agreement between independent results of measurements obtained under stipulated conditions. Usually expressed as the standard deviation.

Prevalence The frequency of disease in the population examined.

Random error Error that arises from imprecision of measurement of the type that is described by a Gaussian distribution (e.g., caused by pipetting variability, signal variability).

Random sample A random sample from a population is one in which each member has an equal chance of being selected.

Receiver operating characteristic curve A graph of sensitivity versus 1—specificity for all possible cutoff values of a diagnostic test; used to estimate sensitivity and specificity for various decision cutoffs.

Reference measurement procedure A procedure of highest analytical quality that has been shown to yield values having an uncertainty of measurement commensurate with its intended use, especially in assessing the trueness of other measurement procedures for the same quantity and in characterizing reference materials.

Regression analysis A statistical analysis that compares measurement relations between two analytical methods.

Repeatability Closeness of agreement between results of successive measurements carried out under the same conditions (i.e., corresponding to within-run precision).

Reproducibility Closeness of agreement between results of successive measurements carried out under changed conditions (e.g., corresponding to between-runs).

Sample A finite set of variables drawn from an infinite population of variables.

Sensitivity The proportion of subjects with disease who have a positive laboratory test result.

Specificity The proportion of subjects without disease who have a negative laboratory test result.

Standard deviation Square root of sum of squared deviations from the mean divided by number of variables minus one. See under Basic Statistics.

Student *t* distribution Distribution related to the Gaussian distribution given a limited sample size. See under Basic Statistics.

Systematic error Error in measurement that arises from calibration bias or nonspecificity of an assay and, in the course of a number of analyses of the same analyte, remains constant (y-intercept deviation from zero) or varies in a proportional way (slope deviation from unity) based on the analyte concentration.

Traceability In relation to analytical methods, a concept based on a chain of comparisons of measurements that lead to a known reference value done to ensure reasonable agreement between measurements of routine methods.

Trueness A qualitative term that describes the closeness of agreement between the average value obtained from a large series of results of measurements and a true value.

Uncertainty A parameter associated with the result of a measurement that characterizes the dispersion of the values that could reasonably be attributed to the measure; more briefly, uncertainty is a parameter characterizing the range of values within which the value of the quantity being measured is expected to lie.

ASSAY SELECTION OVERVIEW

The introduction of new or revised laboratory tests, markers, or assays is a common occurrence in the clinical laboratory. Test selection and evaluation are key steps in the process of implementing new measurement procedures (Fig. 2.1).

Evaluation of tests, markers, or assays in the clinical laboratory is influenced strongly by guidelines and accreditation or other regulatory standards. The Clinical and Laboratory Standards Institute (CLSI) has published a series of consensus protocols for clinical chemistry laboratories and manufacturers to follow when evaluating methods (http://www.clsi.org). The International Organization for Standardization (ISO) has also developed several documents related to method evaluation (ISOs). In addition, meeting laboratory accreditation requirements has become an important aspect in most countries. Abbreviations are listed in Box 2.1.

Analytical Performance Criteria

In evaluation of a laboratory test, (1) trueness (formerly termed accuracy), (2) precision, (3) analytical range, (4) detection limit, and (5) analytical specificity are of prime importance. The sections in this chapter on laboratory test evaluation and comparison contain detailed outlines of these concepts. Estimated test performance parameters should be related to analytical performance specifications that ensure acceptable clinical use of the test and its results. For more details related to the recommended models for setting analytical performance specifications, readers are referred to Chapters 4 and 7. From a practical point of view, the "ruggedness" of the test in routine use is of importance, and reliable performance, when used by different operators and with different batches of reagents over long time periods, is essential.

BASIC STATISTICS

In this section, fundamental statistical concepts and techniques are introduced in the context of typical analytical investigations. The basic concepts of (1) populations, (2) samples, (3) parameters, (4) statistics, and (5) probability distributions are defined and illustrated. Two important probability distributions—Gaussian and Student *t*—are introduced and discussed.

Frequency Distribution

A graphical device for displaying a large set of laboratory test results is the **frequency distribution**, also called a *histogram*.

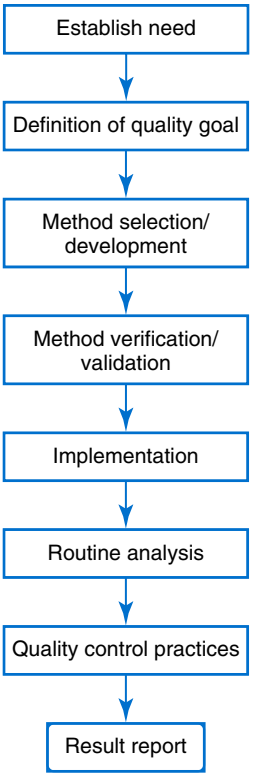


Fig. 2.1 A flow diagram that illustrates the process of introducing a new assay into routine use.

BOX 2.1 Abbreviations

CI	Confidence interval
CV	Coefficient of variation (= SD/x, where x is the concentration)
CV%	= CV × 100%
CV_A	Analytical coefficient of variation
CV_{RB}	Sample-related random bias coefficient of variation
ISO	International Organization for Standardization
OLR	Ordinary least-squares regression analysis
SD	Standard deviation
SEM	Standard error of the mean $\left(=SD/\sqrt{N}\right)$
SD_A	Analytical standard deviation
SD_{RB}	Sample-related random bias standard deviation
x_m	Mean
x_{mw}	Weighted mean

Fig. 2.2 shows a frequency distribution displaying the results of serum gamma-glutamyltransferase (GGT) measurements of 100 apparently healthy 20- to 29-year-old men. The frequency distribution is constructed by dividing the measurement scale into cells of equal width; counting the number, n_i, of values that fall within each cell; and drawing a rectangle above each cell whose area (and height because the cell widths are all equal) is proportional to n_i. In this example, the selected cells were 5 to 9, 10 to 14, 15 to 19, 20 to 24, 25 to 29, and so on, with 60 to 64 being the last cell (range of values, 5 to 64 U/L). The ordinate axis of the frequency distribution

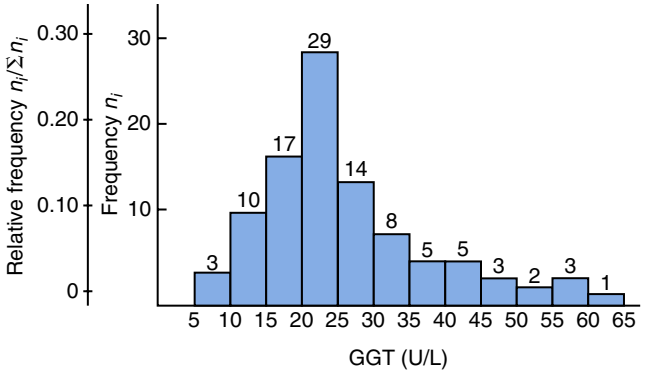


Fig. 2.2 Frequency distribution of 100 gamma-glutamyltransferase *(GGT)* values.

gives the number of values falling within each cell. When this number is divided by the total number of values in the data set, the relative frequency in each cell is obtained.

Often, the position of the value for an individual within a distribution of values is useful medically. The *nonparametric* approach can be used to determine directly the *percentile* of a given subject. Having ranked N subjects according to their values, the n-percentile, $Perc_n$, may be estimated as the value of the $[N(n/100) + 0.5]$ ordered observation. In the case of a non-integer value, interpolation is carried out between neighbor values. The 50th percentile is the median of the distribution.

Population and Sample

It is useful to obtain information and draw conclusions about the characteristics of the test results for one or more target populations. In the GGT example, interest is focused on the location and spread of the population of GGT values for 20- to 29-year-old healthy men. Thus a working definition of a **population** is the complete set of all observations that might occur as a result of performing a particular procedure according to specified conditions.

Most target populations of interest in clinical chemistry are in principle very large (millions of individuals) and so are impossible to study in their entirety. Usually a subgroup of observations is taken from the population as a basis for forming conclusions about population characteristics. The group of observations that has been selected from the population is called a **sample**. For example, the 100 GGT values make up a sample from a respective target population. However, a sample is used to study the characteristics of a population only if it has been properly selected. For instance, if the analyst is interested in the population of GGT values over various lots of materials and some time period, the sample must be selected to be representative of these factors, as well as of age, sex, and health factors of the individuals in the targeted population. Consequently, exact specification of the target population(s) is necessary before a plan for obtaining the sample(s) can be designed. In this chapter, a sample is also used as a specimen, depending on the context.

Probability and Probability Distributions

Consider again the frequency distribution in Fig. 2.2. In addition to the general location and spread of the GGT

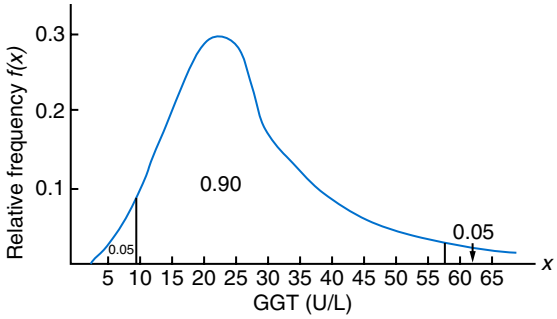


Fig. 2.3 Population frequency distribution of gamma-glutamyltransferase *(GGT)* values.

determinations, other useful information can be easily extracted from this frequency distribution. For instance, 96% (96 of 100) of the determinations are less than 55 U/L, and 91% (91 of 100) are greater than or equal to 10 but less than 50 U/L. Because the cell interval is 5 U/L in this example, statements such as these can be made only to the nearest 5 U/L. A larger sample would allow a smaller cell interval and more refined statements. For a sufficiently large sample, the cell interval can be made so small that the frequency distribution can be approximated by a continuous, smooth curve, similar to that shown in Fig. 2.3. In fact, if the sample is large enough, this can be considered a close representation of the "true" target *population frequency distribution*. In general, the functional form of the population frequency distribution curve of a variable x is denoted by $f(x)$.

The population frequency distribution allows us to make probability statements about the GGT of a randomly selected member of the population of healthy 20- to 29-year-old men. For example, the probability $\Pr(x > x_a)$ that the GGT value x of a randomly selected 20- to 29-year-old healthy man is greater than some particular value x_a is equal to the area under the population frequency distribution to the right of x_a. If $x_a = 58$, then from Fig. 2.3, $\Pr(x > 58) = 0.05$. Similarly, the probability $\Pr(x_a < x < x_b)$ that x is greater than x_a but less than x_b is equal to the area under the population frequency distribution between x_a and x_b. For example, if $x_a = 9$ and $x_b = 58$, then from Fig. 2.3, $\Pr(9 < x < 58) = 0.90$. Because the population frequency distribution provides all information related to probabilities of a randomly selected member of the population, it is called the probability distribution of the population. Although the true probability distribution is never exactly known in practice, it can be approximated with a large sample of observations, that is, test results.

Parameters: Descriptive Measures of a Population

Any population of values can be described by measures of its characteristics. A *parameter* is a constant that describes some particular characteristic of a population. Although most populations of interest in analytical work are infinite in size, for the following definitions, the population will be considered to be of finite size N, where N is very large.

One important characteristic of a population is its *central location*. The parameter most commonly used to describe the central location of a population of N values is the *population mean* (μ):

$$\mu = \frac{\sum x_i}{N}$$

An alternative parameter that indicates the central tendency of a population is the **median**, which is defined as the 50th percentile, Perc_{50}.

Another important characteristic is the *dispersion* of values about the population mean. A parameter very useful in describing this dispersion of a population of N values is the *population variance* σ^2 (sigma squared):

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

The *population standard deviation (SD)* σ, the positive square root of the population variance, is a parameter frequently used to describe the population dispersion in the same units (e.g., mg/dL) as the population values. For a Gaussian distribution, 95% of the population of values are located within the mean $\pm 1.96\ \sigma$. If a distribution is non-Gaussian (e.g., asymmetric), an alternative measure of dispersion based on the percentiles may be more appropriate, such as the distance between the 25th and 75th percentiles (the interquartile interval).

Statistics: Descriptive Measures of the Sample

As noted earlier, clinical chemists usually have at hand only a sample of observations (i.e., test results) from the overarching targeted population. A *statistic* is a value calculated from the observations in a sample to estimate a particular characteristic of the target population. As introduced earlier, the sample mean x_m is the arithmetical average of a sample, which is an estimate of μ. Likewise, the sample **standard deviation** (SD) is an estimate of σ, and the **coefficient of variation** (CV) is the ratio of the SD to the mean multiplied by 100%. The equations used to calculate x_m, SD, and CV, respectively, are as follows:

$$x_m = \frac{\sum x_i}{N}$$

$$\text{SD} = \sqrt{\frac{\sum (x_i - x_m)^2}{N-1}} = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{N}}{N-1}}$$

$$\text{CV} = \frac{\text{SD}}{x_m} \times 100\%$$

where x_i is an individual measurement and N is the number of sample measurements.

The SD is an estimate of the dispersion of the distribution. In addition, from the SD, an estimate of the uncertainty of x_m can be derived as an estimate of μ (see later discussion).

Random Sampling

A random sample of individuals from a target population is one in which each member of the population has an equal chance of being selected. A **random sample** is one in which each member of the sample can be considered to be a random

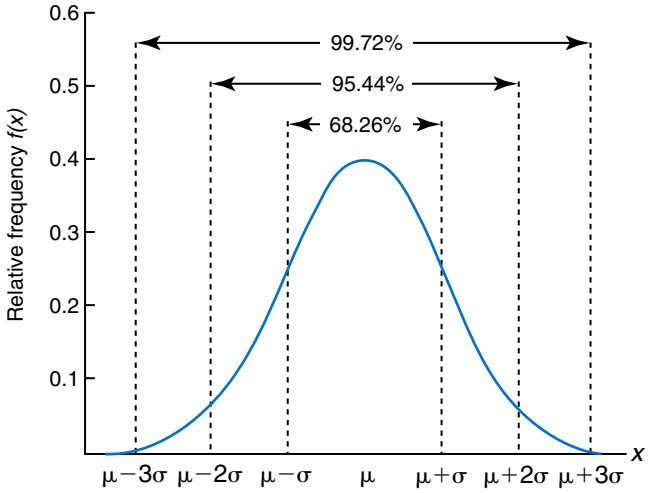


Fig. 2.4 The Gaussian probability distribution.

selection from the target population. Although much of statistical analysis and interpretation depends on the assumption of a random sample from some population, actual data collection often does not satisfy this assumption. In particular, for sequentially generated data, it is often true that observations adjacent to each other tend to be more alike than observations separated in time.

Gaussian Probability Distribution

The **Gaussian probability distribution**, illustrated in Fig. 2.4, is of fundamental importance in statistics for several reasons. As mentioned earlier, a particular test result x will not usually be equal to the true value μ of the specimen being measured. Rather, associated with this particular test result x will be a particular measurement error $\varepsilon = x - \mu$, which is the result of many contributing sources of error. Pure measurement errors tend to follow a probability distribution similar to that shown in Fig. 2.4, where the errors are symmetrically distributed, with smaller errors occurring more frequently than larger ones, and with an expected value of 0. This important fact is known as the central limit effect for distribution of errors: if a measurement error ε is the sum of many independent sources of error, such as $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k$, several of which are major contributors, the probability distribution of the measurement error ε will tend to be Gaussian as the number of sources of error becomes large.

Another reason for the importance of the Gaussian probability distribution is that many statistical procedures are based on the assumption of a Gaussian distribution of values; this approach is commonly referred to as *parametric*. Furthermore, these procedures usually are not seriously invalidated by departures from this assumption. Finally, the magnitude of the uncertainty associated with sample statistics can be ascertained based on the fact that many sample statistics computed from large samples have a Gaussian probability distribution.

The Gaussian probability distribution is completely characterized by its mean μ and its variance σ^2. The notation $N(\mu, \sigma^2)$ is often used for the distribution of a variable that is Gaussian with mean μ and variance σ^2. Probability statements

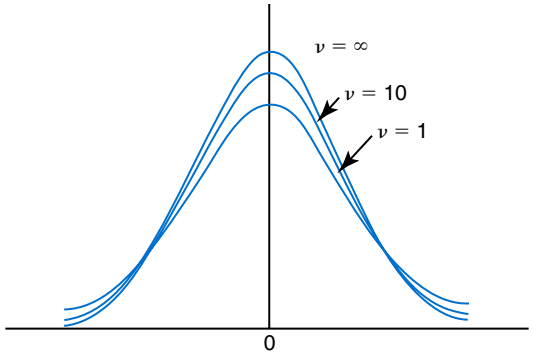


Fig. 2.5 The t distribution for ν = 1, 10, and ∞.

about a variable x that follows an $N(\mu, \sigma^2)$ distribution are usually made by considering the variable z,

$$z = \frac{x - \mu}{\sigma}$$

which is called the *standard Gaussian variable*. The variable z has a Gaussian probability distribution with $\mu = 0$ and $\sigma^2 = 1$, that is, z is $N(0, 1)$. The probably that x is within 2σ of μ [i.e., $\Pr(|x - \mu| < 2\sigma) =$] is 0.9544. Most computer spreadsheet programs can calculate probabilities for all values of z.

Student *t* Probability Distribution

To determine probabilities associated with a Gaussian distribution, it is necessary to know the population SD σ. In actual practice, σ is often unknown, so z cannot be calculated. However, if a random sample can be taken from the Gaussian population, the sample SD can be calculated, by substituting SD for σ, and computing the value t:

$$t = \frac{x - \mu}{\text{SD}}$$

Under these conditions, the variable t has a probability distribution called the **Student *t* distribution**. The t distribution is a family of distributions depending on the degrees of freedom ν (= $N - 1$) for the sample SD. Several t distributions from this family are shown in Fig. 2.5. When the size of the sample and the degrees of freedom for SD are infinite, there is no uncertainty in SD, so the t distribution is identical to the standard Gaussian distribution. However, when the sample size is small, the uncertainty in SD causes the t distribution to have greater dispersion and heavier tails than the standard Gaussian distribution, as illustrated in Fig. 2.5. At sample sizes above 30, the difference between the t distribution and the Gaussian distribution becomes relatively small and can usually be neglected. Most computer spreadsheet programs can calculate probabilities for all values of t, given the degrees of freedom for SD.

The Student t distribution is commonly used in significance tests, such as comparison of sample means, or in testing conducted if a regression slope differs significantly from 1. Descriptions of these tests can be found in statistics textbooks. Another important application is the estimation of confidence intervals (CIs). CIs are intervals that indicate the uncertainty of a given sample estimate. For example, it can be proved that $X_m \pm t_{alpha}$ (SD/$N^{0.5}$) provides an approximate

2*alpha*-CI for the mean. A common value for *alpha* is 0.025 or 2.5%, which thus results in a 0.95% or 95% CI. Given sample sizes of 30 or higher, t_{alpha} is about 2. ($SD/N^{0.5}$) is called the standard error (SE) of the mean. A CI should be interpreted as follows. Suppose a sampling experiment of drawing 30 observations from a Gaussian population of values is repeated 100 times, and in each case, the 95% CI of the mean is calculated as described. Then, in 95% of the drawings, the true mean μ is included in the 95% CI. According to the central limit theorem, distributions of mean values converge toward the Gaussian distribution irrespective of the primary type of distribution of x. This means that the 95% CI is a robust estimate only minimally influenced by deviations from the Gaussian distribution. In the same way, the t-test is robust toward deviations from normality.

Nonparametric Statistics

Distribution-free statistics, often called nonparametric statistics, provides an alternative to parametric statistical procedures that assume data to have Gaussian distributions. For example, distributions of reference values are often skewed and so do not conform to the Gaussian distribution (see Chapter 5 on reference intervals). Formally, to judge whether a distribution is Gaussian or not, a goodness of fit test should be conducted. A commonly used test is the Kolmogorov-Smirnov test, in which the shape of the sample distribution is compared with the shape presumed for a Gaussian distribution. If the difference exceeds a given critical value, the hypothesis of a Gaussian distribution is rejected, and it is then appropriate to apply nonparametric statistics. A special problem is the occurrence of outliers, (i.e., single measurements highly deviating from the remaining measurements). Outliers may rely on biological factors and so be of real significance (e.g., in the context of estimating reference intervals) or they may be related to clerical errors. Special tests exist for handling outliers.

Given that a distribution is non-Gaussian, it is appropriate to apply nonparametric descriptive statistics based on the percentile or quantile concept. The n-percentile, $Perc_n$, of a sample of N values may be estimated as the value of the $[N(n/100) + 0.5]$ ordered observation. In the case of a noninteger value, interpolation is carried out between neighbor values. The median is the 50th percentile, which is used as a measure of the center of the distribution. For the GGT example, we would order the $N = 100$ values according to size. The median or 50th percentile is then the value of the $[100(50/100) + 0.5 = 50.5]$ ordered observation (the interpolated value between the 50th and 51st ordered values). The 2.5th and 97.5th percentiles are values of the $[100(2.5/100) + 0.5 = 3]$ and $[100(97.5/100) + 0.5 = 98]$ ordered observations, respectively. When a 95% reference interval is estimated, a nonparametric approach is often preferable because many distributions of reference values are asymmetric. Generally, distributions based on the many biological sources of variation are often non-Gaussian compared with distributions of pure measurement errors that usually are Gaussian.

The nonparametric counterpart to the t-test is the Mann-Whitney test, which provides a significance test for the difference between median values of the two groups to be compared, given the same shape of the distributions. When there are more than two groups, the Kruskal-Wallis test can be applied.

Categorical Variables

When dealing with qualitative tests and in the context of evaluating diagnostic testing, categorical variables that only take the value positive or negative come into play. The performance is here given as proportions or percentages, which are proportions multiplied by 100. For example, the diagnostic **sensitivity** of a test is the proportion of diseased subjects who have a positive result. Having tested, for example, 100 patients, 80 might have had a positive test result. The sensitivity then is 0.8% or 80%. Exact estimates of the uncertainty can be derived from the so-called binomial distribution, but for practical purposes, an approximate expression for the 95% CI is usually applied as the estimated proportion $P \pm 2SE$, where the SE in this context is derived as:

$$SE = [P(1-P)/N]^{0.5}$$

where P is here a proportion and not a percentage. In the example, the SE equals 0.0016 and so the 95% CI is 0.77 to 0.83 or 77% to 83%. The applied approximate formula for the SE is regarded as reasonably valid when NP and $N(1-P)$ both are equal to or higher than 5.

TECHNICAL VALIDITY OF ANALYTICAL ASSAYS

This section defines the basic concepts used in this chapter: (1) **calibration**, (2) trueness and accuracy, (3) precision, (4) linearity, (5) **limit of detection** (LOD), (6) limit of quantification, (7) **specificity**, and (8) others.

Calibration

The calibration function is the relation between instrument signal *(y)* and concentration of **analyte** *(x)*, that is,

$$y = f(x)$$

The inverse of this function, also called the measuring function, yields the concentration from response:

$$x = f^{-1}(y)$$

This relationship is established by measurement of samples with known quantities of analyte (calibrators). Solutions of pure chemical standards should be distinguished from samples with known quantities of analyte present in the typical **matrix** that is to be measured (e.g., human serum). The first situation applies typically to a reference measurement procedure that is not influenced by matrix effects; the second case corresponds typically to a routine method that often is influenced by matrix components and so preferably is calibrated using the relevant matrix. Calibration functions may be linear or curved and, in

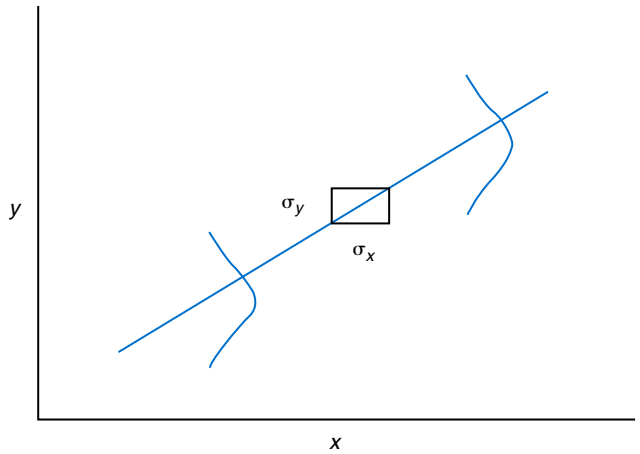


Fig. 2.6 Relation between concentration *(x)* and signal response *(y)* for a linear calibration function. The dispersion in signal response (σ_y) is projected onto the *x*-axis and is called assay imprecision [σ_x (=σ_A)].

TABLE 2.1 Overview of Qualitative Terms and Quantitative Measures Related to Method Performance

Qualitative Concept	Quantitative Measure
Trueness	*Bias*
Closeness of agreement of mean value with "true value"	A measure of the systematic error
Precision	*Imprecision (SD)*
Repeatability (within run)	A measure of the dispersion of random errors
Intermediate precision (long term)	
Reproducibility (interlaboratory)	
Accuracy	*Error of measurement*
Closeness of agreement of a single measurement with "true value"	Comprises both random and systematic influences

SD, Standard deviation.

the case of immunoassays, may often take a special form (e.g., modeled by the four-parameter logistic curve). An alternative, model-free approach is to estimate a smoothed spline curve, which often is performed for immunoassays. If the assumed calibration function does not correctly reflect the true relationship between instrument response and analyte concentration, a **systematic error** or **bias** is likely to be associated with the analytical method.

The precision of the analytical method depends on the stability of the instrument response for a given quantity of analyte. In principle, a random dispersion of instrument signal (vertical direction) at a given true concentration transforms into dispersion on the measurement scale (horizontal direction), as is shown schematically (Fig. 2.6). If the calibration function is linear and the imprecision of the signal response is the same over the analytical measurement range, the analytical SD (SD_A) of the method tends to be constant over the analytical measurement range (see Fig. 2.6). If the imprecision increases proportionally to the signal response, the analytical SD of the method tends to increase proportionally to the concentration *(x)*, which means that the *relative* imprecision (CV = SD/*x*) may be constant over the analytical measurement range assuming that the intercept of the calibration line is zero.

Trueness and Accuracy

Trueness of measurements is defined as closeness of agreement between the average value obtained from a large series of results of measurements and the true value.

The difference between the average value (strictly, the mathematical expectation) and the true value is the *bias*, which is expressed numerically and so is inversely related to the trueness. *Trueness* in itself is a qualitative term that can be expressed, for example, as low, medium, or high. From a theoretical point of view, the exact true value for a clinical sample is not available; instead, an "accepted reference value" is used, which is the "true" value that can be determined in practice. Trueness can be evaluated by comparison of measurements by the new test and by some preselected reference measurement procedure, both on the same sample or individuals.

The ISO has introduced the trueness expression as a replacement for the term *accuracy*, which now has gained a slightly different meaning. **Accuracy** is the closeness of agreement between the result of a measurement and a true concentration of the analyte. Accuracy thus is influenced by both bias and imprecision and in this way reflects the total error. Accuracy, which in itself is a qualitative term, is inversely related to the "uncertainty" of measurement, which can be quantified as described later (Table 2.1).

In relation to trueness, the concepts *recovery*, *drift*, and *carryover* may also be considered. *Recovery* is the fraction or percentage increase in concentration that is measured in relation to the amount added. Recovery experiments are typically carried out in the field of drug analysis. It is useful to distinguish between *extraction recovery*, which often is interpreted as the fraction of compound that is carried through an extraction process, and the recovery measured by the entire analytical procedure, in which the addition of an internal standard compensates for losses in the extraction procedure. A recovery close to 100% is a prerequisite for a high degree of trueness, but it does not ensure unbiased results because possible nonspecificity against matrix components (e.g., an interfering substance) is not detected in a recovery experiment. *Drift* is caused by instrument or reagent instability over time, so that calibration becomes gradually biased. Assay *carryover* also must be close to zero to ensure unbiased results.

Precision

Precision has been defined as the closeness of agreement between independent results of measurements obtained under stipulated conditions. The degree of precision is usually derived from statistical measures of imprecision, such as SD or CV (CV = SD/*x*, where *x* is the measurement concentration), which is inversely related to precision. Imprecision of measurements is solely related to the **random error** of measurements and has no relation to the trueness of measurements.

Precision is specified as follows:

Repeatability: **Repeatability** is the closeness of agreement between results of successive measurements carried out under the same conditions (i.e., corresponding to within-run precision).

Reproducibility: **Reproducibility** is the closeness of agreement between results of measurements performed under changed conditions of measurements (e.g., time, operators, calibrators, reagent lots). Two specifications of reproducibility are often used: total or between-run precision in the laboratory, often termed *intermediate precision*, and inter-laboratory precision (e.g., as observed in external quality assessment schemes [EQAS]) (see Table 2.1).

The total SD (σ_T) may be divided into within-run and between-run components using the principle of analysis of variance of components (variance is the squared SD):

$$\sigma_T^2 = \sigma_{\text{Within-run}}^2 + \sigma_{\text{Between-run}}^2$$

It is not always clear in clinical chemistry publications what is meant by "between-run" variation. Some authors use the term to refer to the total variation of an assay, but others apply the term *between-run variance component* as defined earlier.

In laboratory studies of analytical variation, estimates of imprecision are obtained. It is important to have an adequate number so that analytical variation is not underestimated. Commonly, the number 20 is given as a reasonable number of observations (e.g., suggested in the CLSI guideline for manufacturers). To verify method precision by users, it has been recommended to run internal QC samples for 5 consecutive days in 5 replicates.

To estimate both the within-run imprecision and the total imprecision, a common approach is to measure duplicate control samples in a series of runs. Suppose, for example, that a control is measured in duplicate for 20 runs, in which case 20 observations are present with respect to both components. The dispersion of the means (x_m) of the duplicates is given as follows:

$$\sigma_{x_m}^2 = \frac{\sigma_{\text{Within-run}}^2}{2} + \sigma_{\text{Between-run}}^2$$

From the 20 sets of duplicates, the within-run SD can be derived using the following formula:

$$\text{SD}_{\text{Within-run}} = \left[\Sigma \frac{d_i^2}{(2 \times 20)}\right]^{0.5}$$

where d_i refers to the difference between the ith set of duplicates. When SDs are estimated, the concept degrees of freedom (*df*) is used. In a simple situation, the number of degrees of freedom equals $N - 1$. For N duplicates, the number of degrees of freedom is $N(2 - 1) = N$. Thus both variance components are derived in this way. The advantage of this approach is that the within-run estimate is based on several runs, so that an average estimate is obtained rather than only an estimate for one particular run if all 20 observations had been obtained in the same run.

Generally, the estimate of the imprecision improves as more observations become available. Exact confidence limits for the SD can be derived from the χ^2 distribution. A graphical display of 95% CIs at various sample sizes is shown in Fig. 2.7. For example, suppose we have estimated the imprecision as an SD of 5.0 on the basis of $N = 20$ observations. From the figure, the 95% CI extends from about 0.75 × 5.0 to about 1.45 × 5.0, i.e., from 3.8 to 7.3.

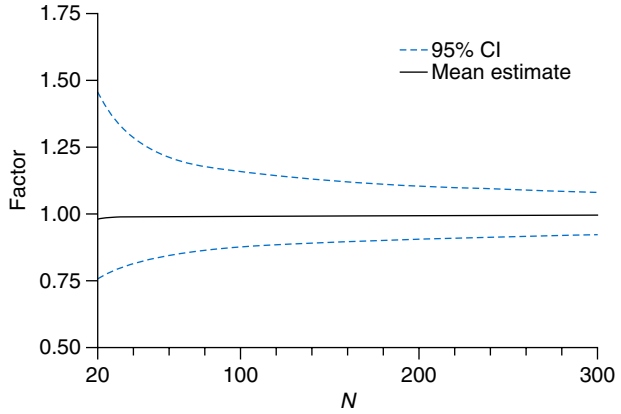


Fig. 2.7 Relation between factors indicating the 95% confidence intervals *(CIs)* of standard deviations (SDs) and the sample size. The true SD is 1, and the solid line indicates the mean estimate, which is slightly downward biased at small sample sizes.

Precision often depends on the concentration of analyte being considered. A presentation of precision as a function of analyte concentration is the precision profile, which usually is plotted in terms of the SD or the CV as a function of analyte concentration (Fig. 2.8).

Linearity

Linearity refers to the relationship between measured and expected values over the analytical measurement range. Linearity may be considered in relation to actual or relative analyte concentrations. In the latter case, a dilution series of a sample may be examined. This dilution series examines whether the measured concentration changes as expected according to the proportional relationship between samples introduced by the dilution factor. Dilution is usually carried out with an appropriate sample matrix (e.g., human serum [individual or pooled serum] or a verified sample diluent).

Evaluation of linearity can be performed visually or by statistical tests. When repeated measurements are available at each concentration, the random variation between measurements and the variation around an estimated regression line may be evaluated by an F-test. When significant nonlinearity is found, it may be useful to explore nonlinear alternatives to the linear regression line (i.e., polynomials of higher degrees). Another approach is to assess the residuals of an estimated regression line. An additional consideration for evaluating proportional concentration relationships is whether an estimated regression line passes through zero or not.

Analytical Measurement Range and Limits of Quantification

The **analytical measurement range** (measuring interval, reportable range) is the analyte concentration range over which measurements are within the declared tolerances for imprecision and bias of the method. Taking drug assays as an example, there exist (arbitrary) requirements of a CV% of less than 15% and a bias of less than 15%. The measurement range then extends from the lowest concentration (lower limit of quantification [LLOQ]) to the highest concentration (upper limit of

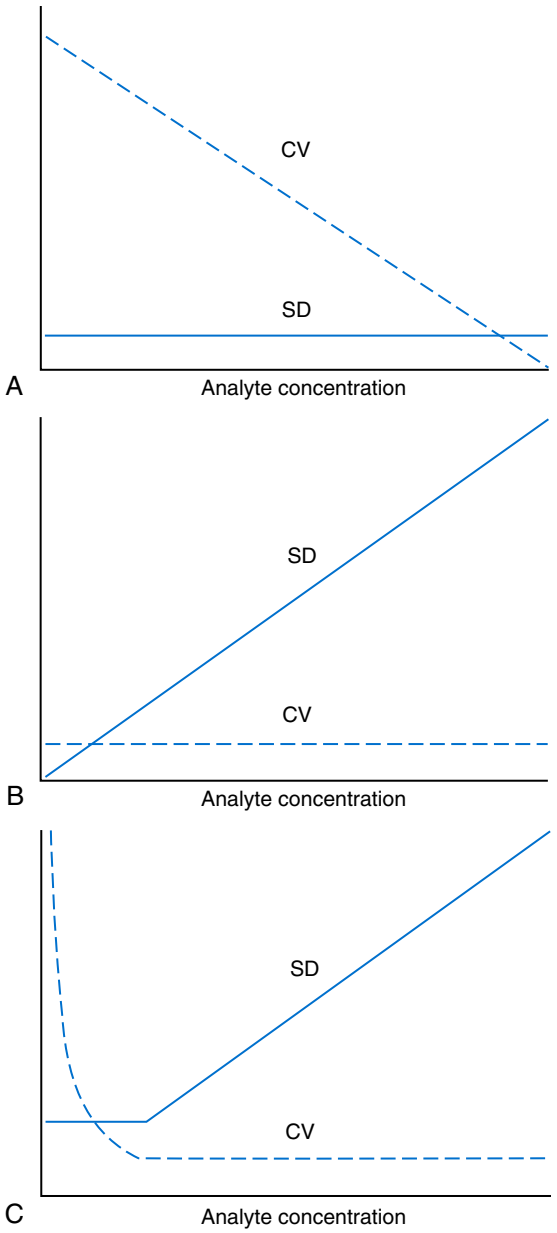


Fig. 2.8 Relations between analyte concentration and standard deviation *(SD)*/coefficient of variation *(CV)*. (A) The SD is constant, so that the CV varies inversely with the analyte concentration. (B) The CV is constant because of a proportional relationship between concentration and SD. (C) A mixed situation with constant SD in the low range and a proportional relationship in the rest of the analytical measurement range.

quantification [ULOQ]) for which these performance specifications are fulfilled. The LLOQ is medically important for many analytes, e.g., thyroid-stimulating hormone (TSH), where low TSH results are useful for the diagnosis of hyperthyroidism.

The LOD may be defined as the lowest value that significantly exceeds the measurements of a blank sample. Thus the limit has been estimated on the basis of repeated measurements of a blank sample and has been *reported* as the mean plus 2 or 3 SDs of the blank measurements. In the interval from LOD up to LLOQ, a result should be reported as "detected" but not provided as a quantitative result.

The LLOQ of an assay should not be confused with analytical sensitivity. The **analytical sensitivity** is defined as ability of an analytical method to assess small differences in the concentration of analyte. The analytical sensitivity depends on the precision of the method. The smallest difference that will be statistically significant equals $2 \times \sqrt{2 \times SD_A}$ at a 5% significance level.

Analytical Specificity and Interference

Analytical specificity is the ability of an assay procedure to determine the concentration of the target analyte without influence from potentially interfering substances or factors in the sample matrix (e.g., hyperlipemia, hemolysis, bilirubin, antibodies, other metabolic molecules, degradation products of the analyte, exogenous substances, anticoagulants).

POINTS TO REMEMBER

- Technical validation of analytical methods focuses on (1) calibration, (2) trueness and accuracy, (3) precision, (4) linearity, (5) LOD, (6) limit of quantification, (7) specificity, and (8) others.
- The difference between the average measured value and the true value is the bias, which can be evaluated by comparison of measurements by the new test and by some preselected reference measurement procedure, both on the same sample or individuals.
- The degree of precision is usually expressed on the basis of statistical measures of imprecision, such as SD or CV (CV = SD/x, where x is the measurement concentration).
- The measurement range extends from the lowest concentration (LLOQ) to the highest concentration (ULOQ) for which the analytical performance specifications (imprecision, bias) are fulfilled.
- Analytical specificity is the ability of an assay procedure to determine the concentration of the target analyte without influence from potentially interfering substances or factors in the sample matrix.

ASSAY COMPARISON

Comparison of measurements by two assays can be carried out by parallel measurements of a set of patient samples. To prevent artificial matrix-induced differences, fresh patient samples are the optimal material. A nearly even distribution of values over the analytical measurement range is also preferable. In an ordinary laboratory, comparison of two routine assays is the most frequently occurring situation. Less commonly, comparison of a routine assay with a reference measurement procedure is undertaken. When two routine assays are compared, it is not possible to establish that one set of measurements is the correct one. Rather, the question is whether the new assay can replace the existing one without a systematic change in result values. To address this question, a statistical procedure with graphics display should be applied. A difference (bias) plot, which shows differences as a function of the average concentration of measurements (Bland-Altman plot), or a **regression analysis**.

Basic Error Model

The occurrence of measurement errors is related to the performance characteristics of the assay. It is important to distinguish between pure, random measurement errors, which are present in all measurement procedures, and errors related to incorrect calibration and nonspecificity of the assay. Whereas a reference measurement procedure is associated only with pure random error, a routine method, additionally, is likely to have some bias related to errors in calibration and limitations with regard to specificity. Whereas an erroneous calibration function gives rise to a systematic error, nonspecificity gives an error that typically varies from sample to sample. The error related to nonspecificity thus has a random character, but in contrast to the pure measurement error, it cannot be reduced by repeated measurements of a sample. Although errors related to nonspecificity for a group of samples look like random errors, for the individual sample, this type of error is a bias. Because this bias varies from sample to sample, it has been called a *sample-related random bias.* In the following section, the various error components are incorporated into a formal **error model**.

Measured Value, Target Value, Modified Target Value, and True Value

Taking into account that an analytical method measures analyte concentrations with some random measurement error, it is necessary to distinguish between the actual, measured value and the average result obtained if the given sample was measured an infinite number of times. If the assay is a reference assay without bias and nonspecificity, the following simple relationship holds:

$$x_i = X_{\text{True}i} + \varepsilon_i$$

where x_i represents the measured value, $X_{\text{True}i}$ is the average value for an infinite number of measurements, and ε_i is the deviation of the measured value from the average value. If the sample was measured repeatedly, the average of ε_i would be zero and the SD would equal the analytical SD (σ_A) of the reference measurement procedure. Pure, random measurement error will usually be Gaussian distributed.

In the case of a routine assay, the relationship between the measured value for a sample and the true value becomes more complicated:

$$x_i = X_{\text{True}i} + \text{Cal-Bias} + \text{Random-Bias}_i + \varepsilon_i$$

The *Cal-Bias* term (calibration bias) is a systematic error related to the calibration of the method. This systematic error may be a constant for all measurements corresponding to an offset error, or it may be a function of the analyte concentration (e.g., corresponding to a slope deviation in the case of a linear calibration function). The *Random-Bias*$_i$ term is a bias that is specific for a given sample related to nonspecificity of the method. It may arise because of codetermination of substances that vary in concentration from sample to sample. For example, a chromogenic creatinine method codetermines some other components with creatinine in serum.

The final term in the equation above is the random measurement error term, ε_i. If an infinite number of measurements of a specific sample is performed by the routine method, the random measurement error term ε_i would be zero. The cal-bias and the random-bias$_i$, however, would be unchanged. Thus the average value of an infinite number of measurements would equal the sum of the true value and these bias terms. This average value may be regarded as the target value ($X_{\text{Target}i}$) of the given sample for the routine method:

$$X_{\text{Target}i} = X_{\text{True}i} + \text{Cal-Bias} + \text{Random-Bias}_i$$

As mentioned, the calibration bias represents a systematic error component in relation to the true values measured by a reference measurement procedure. In the context of regression analysis, this systematic error corresponds to the intercept and the slope deviation from unity when a routine method is compared with a reference measurement procedure (outlined in detail later). It is convenient to introduce a modified target value expression ($X'_{\text{Target}i}$) for the routine method to delineate this systematic calibration bias, so that:

$$X'_{\text{Target}i} = X_{\text{True}i} + \text{Cal-Bias}$$

Thus, for a set of samples measured by a routine method, the $X_{\text{Target}i}$ values are distributed around the respective $X'_{\text{Target}i}$ values with an SD, which is called σ_{RB}.

If the assay is a reference method without bias and nonspecificity, the target value and the modified target value equal the true value, that is,

$$X_{\text{Target}i} = X'_{\text{Target}i} = X_{\text{True}i}$$

The error model is outlined in Fig. 2.9.

Calibration Bias and Random Bias

For an individual measurement, the total error is the deviation of x_i from the true value, that is,

$$\text{Total error of } x_i = \text{Cal-Bias} + \text{Random-Bias}_i + \varepsilon_i$$

Estimation of the bias terms requires parallel measurements between the method in question and a reference method. With regard to calibration bias, the possibility of lot-to-lot variation in analytical kit sets should be recognized. Lot-to-lot variation shows up as a calibration bias that changes from lot to lot.

The previous exposition defines the total error in broader terms than is often seen. A traditional total error expression is:

$$\text{Total error} = \text{Bias} + 2\ \text{SD}_A$$

This is often interpreted as the calibration bias plus 2 SD_A. If a one-sided statistical perspective is taken, the expression is modified to Bias + 1.65 SD_A, indicating that 5% of results are located outside the limit. Interpreting the bias as identical with the calibration bias may lead to an underestimation of the total error.

Random bias related to sample-specific interferences may take several forms. It may be a regularly occurring additional random error component, perhaps of the same order of

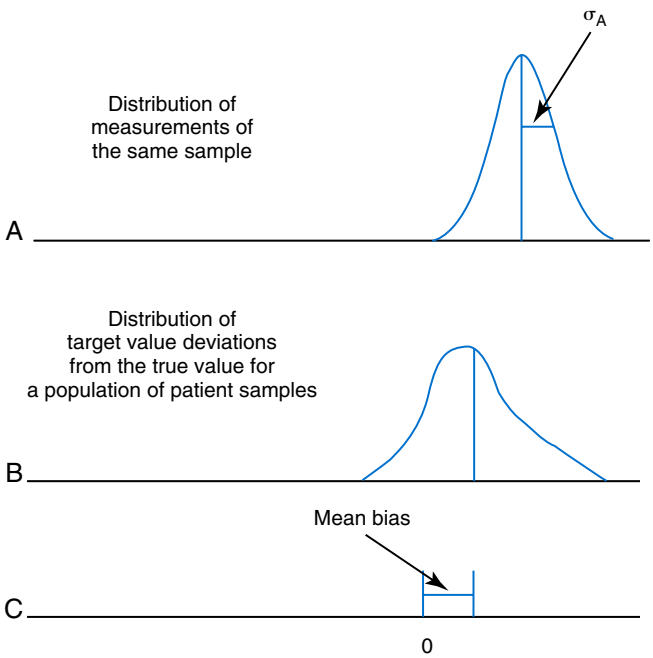


Fig. 2.9 Outline of basic error model for measurements by a routine assay. (A) The distribution of repeated measurements of the *same* sample, representing a normal distribution around the target value ($X_{Target\,i}$) *(vertical line)* of the sample with a dispersion corresponding to the analytical standard deviation, σ_A. (B) Schematic outline of the dispersion of target value deviations from the respective true values for a population of patient samples. A distribution of an arbitrary form is displayed. The standard deviation equals σ_{RB}. The vertical line indicates the mean of the distribution. (C) The distance from zero to the mean of the target value deviations from the true values represents the calibration bias (mean bias = cal-bias) of the assay.

magnitude as the analytical error. In this context, it is natural to quantify the error in the form of an SD or CV. The most straightforward procedure is to carry out a method comparison study based on a set of patient samples in which one of the methods is a reference method, as outlined later. Another form of sample-related random interference is more rarely occurring gross errors, which typically are seen in the context of immunoassays and are related to unexpected antibody interactions. Such an error usually shows up as an outlier in method comparison studies. Outliers should be investigated to identify their cause, which may be an important limitation in using a given assay.

Assay Comparison Data Model

Here we consider the error model described earlier in relation to the method comparison situation. For a given sample measured by two analytical methods, 1 and 2, the following equations apply:

$$x1_i = X1_{Target\,i} + \varepsilon 1_i = X_{True\,i} + \text{Cal-Bias1} + \text{Random-Bias1}_i + \varepsilon 1_i$$

$$x2_i = X2_{Target\,i} + \varepsilon 2_i = X_{True\,i} + \text{Cal-Bias2} + \text{Random-Bias2}_i + \varepsilon 2_i$$

In the following, some typical situations using this general model are described. First, comparison of a routine assay with a reference measurement procedure will be treated. Second, comparison of two routine assays is considered.

Assuming that method 1 is a reference method, the bias components disappear by definition, and the following situation can be described:

$$x1_i = X1_{Target\,i} + \varepsilon 1_i = X_{True\,i} + \varepsilon 1_i$$

$$x2_i = X2_{Target\,i} + \varepsilon 2_i = X_{True\,i} + \text{Cal-Bias2} + \text{Random-Bias2}_i + \varepsilon 2_i$$

The paired differences become

$$(x2_i - x1_i) = \text{Cal-Bias2} + \text{Random-Bias2}_i + (\varepsilon 2_i - \varepsilon 1_i)$$

We thus have an expression consisting of a systematic error term (calibration bias of method 2) and two random terms. The Random-Bias2 term is distributed around Cal-Bias2 according to an undefined distribution. ($\varepsilon 2_i - \varepsilon 1_i$) is a difference between two random measurement errors that are independent and, commonly, Gaussian distributed. However, we remind readers that the SD for analytical methods often depends on the concentration, as mentioned earlier. For analytes with a wide analytical measurement range (e.g., some hormones), both sample-related random interferences and analytical SDs are likely to depend on the measurement concentration, often in a roughly proportional manner. It may then be more useful to evaluate the *relative* differences—$(x2_i - x1_i)/[(x2_i + x1_i)/2]$—and accordingly express mean and random bias and analytical error as proportions.

In the comparison of two routine methods, the paired differences become

$$(x2_i - x1_i) = (\text{Cal-Bias2} - \text{Cal-Bias1}) + (\text{Random-Bias2}_i - \text{Random-Bias1}_i) + (\varepsilon 2_i - \varepsilon 1_i)$$

The expression again consists of a constant term, the difference between the two calibration biases, and two random terms. The first random term is a difference between two random-bias components that may or may not be independent. If the two field methods are based on the same measurement principle, the random bias terms are likely to be correlated. For example, two chromogenic methods for creatinine are likely to be subject to interference from the same chromogenic compounds present in a given serum sample. On the other hand, a chromogenic method and an enzymatic creatinine method are subject to different types of interfering compounds, and the random bias terms may be relatively independent. In the $\varepsilon 2_i - \varepsilon 1_i$ term, the same relationships as described previously are likely to apply. The general form of the expressed differences is the same in the two situations and the same statistical principles apply.

Planning a Method Comparison Study

Several points require attention, including the (1) number of samples necessary, (2) distribution of analyte concentrations (preferably uniform over the analytical measurement range), and (3) representativeness of the samples. To address the

latter point, samples from relevant patient categories should be included, so that possible interference phenomena can be discovered. Practical aspects related to storage and treatment of samples (e.g., container) and possible artifacts induced by storage (e.g., freezing of samples) and addition of anticoagulants should be considered. Comparison of measurements should preferably be undertaken over several days (e.g., at least 5 days) so that the comparison of methods does not become dependent on the performance of the methods in one particular analytical run. The CLSI guideline EP-09-A3, "Method Comparison and Bias Estimation Using Patient Samples," suggests measurement of 40 samples in duplicate by each method when a new method is introduced in the laboratory as a substitute for an established one. Additionally, 100 samples in duplicate is proposed for a vendor of an analytical test system. The EP15 guideline "User Verification of Manufacturer's Claims" suggests a more condensed approach based on a bias or **difference plot** for 20 samples.

Difference (Bland-Altman) Plot

The Bland-Altman plot is usually understood as a plot of the differences against the average results of the methods. Thus the difference plot in this version provides information on the relation between differences and concentration, which is useful in evaluating whether problems exist at certain ranges (e.g., in the high range) caused by nonlinearity of one of the methods. It may also be of interest to observe whether differences tend to increase proportionally with the concentration or whether they are independent of concentration.

The basic version of the difference plot requires plotting of the differences against the average of the measurements. Fig. 2.10 shows the plot for the drug **assay comparison** data. The interval ±2 SD of the differences is often delineated around the mean difference. To assess whether the bias is significantly different from zero, the SE of the mean difference is estimated as the SD divided by the square root of the number of paired measurements ($SE = SD/N^{0.5}$) and tested against zero by a t-test ($t = [\text{Mean} - 0]/SE$).

Nonparametric limits may also be considered. A constant bias over the analytical measurement range changes the average concentration away from zero. The presence of sample-related random interferences increases the width of the distribution. If the calibration bias depends on the concentration, if the dispersion varies with the concentration, or if both occur, the relations become more complex, and the interval mean ±2 SD of the differences may not fit very well as a 95% interval throughout the analytical measurement range.

The displayed Bland-Altman plot for the drug assay comparison data (see Fig. 2.10) shows a tendency toward increasing scatter with increasing concentration, which is a reflection of increasing random error with concentration. Thus a plot of the relative differences against the average concentration is of relevance (Fig. 2.11).

Regression Analysis

Regression analysis has the advantage that it allows the relation between the target values for the two compared methods

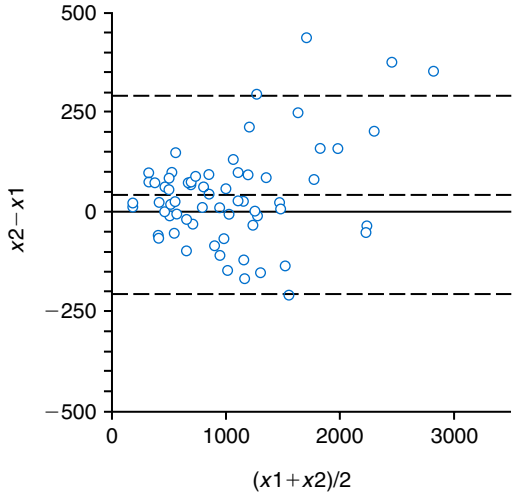


Fig. 2.10 Bland-Altman plot of differences for the drug comparison example. The differences are plotted against the average concentration. The mean difference (42 nmol/L) with ±2 standard deviation of differences is shown *(dashed lines)*.

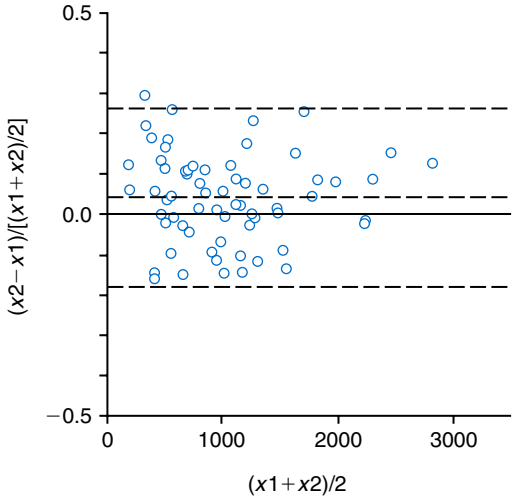


Fig. 2.11 Bland-Altman plot of *relative* differences for the drug comparison example. The differences are plotted against the average concentration. The mean relative difference (0.042) with ±2 standard deviation of relative differences is shown *(dashed lines)*.

to be studied over the full analytical measurement range. In linear regression analysis, it is assumed that the systematic difference between target values can be modeled as a constant systematic difference (intercept deviation from zero) combined with a proportional systematic difference (slope deviation from unity), usually related to a discrepancy with regard to calibration of the methods. In situations when random errors have a constant SD, unweighted regression procedures are used (e.g., **Deming regression** analysis). For cases with SDs that are proportional to the concentration, the weighted Deming regression procedure is preferred.

As outlined previously, we distinguish between the measured value (x_i) and the target value ($X_{\text{Target}i}$) of a sample subjected to analysis by a given method. In linear regression analysis, we assume a linear relationship between values devoid of random error of any kind. Thus to operate with a

linear relationship between values without random measurement error and sample-related random bias, we have to introduce modified target values:

$$X1_{\text{Target}i} = X1'_{\text{Target}i} + \text{Random-Bias1}_i >$$

$$X2_{\text{Target}i} = X2'_{\text{Target}i} + \text{Random-Bias2}_i$$

where we now assume a linear relationship between these modified target values:

$$X2'_{\text{Target}i} = \alpha_0 + \beta X1'_{\text{Target}i}$$

In this model, α_0 corresponds to a constant difference with regard to calibration, and $(\beta - 1)$ is a proportional deviation. Thus the systematic error or calibration difference between the measurements corresponds to

$$X2'_{\text{Target}i} - X1'_{\text{Target}i} = \alpha_0 + (\beta - 1)\, X1'_{\text{Target}i}$$

Because of sample-related random interferences and measurement imprecision (of the type that can be described by a Gaussian distribution, e.g., caused by pipetting variability, signal variability), individually measured pairs of values ($x1_i$, $x2_i$) will be scattered around the line expressing the relationship between $X1'_{\text{Target}i}$ and $X2'_{\text{Target}i}$. Fig. 2.12 outlines schematically how the random distribution of $x1$ and $x2$ values occurs around the regression line. We have

$$x1_i = X1_{\text{Target}i} + \varepsilon 1_i = X1'_{\text{Target}i} + \text{Random-Bias1}_i + \varepsilon 1_i$$

$$x2_i = X2_{\text{Target}i} + \varepsilon 2_i = X2'_{\text{Target}i} + \text{Random-Bias2}_i + \varepsilon 2_i$$

The random error components may be expressed as SDs, and generally we can assume that sample-related random bias (SD σ_{RB}) and analytical imprecision (SD σ_{A}) are independent for each analyte, yielding the relations

$$\sigma^2_{ex1} = \sigma^2_{\text{RB1}} + \sigma^2_{\text{A1}}$$

$$\sigma^2_{ex2} = \sigma^2_{\text{RB2}} + \sigma^2_{\text{A2}}$$

σ_{ex1} and σ_{ex2} are the total SDs of the distributions of $x1_i$ and $x2_i$ around their respective modified target values, $X1'_{\text{Target}i}$ and $X2'_{\text{Target}i}$. The sample-related random bias components for methods 1 and 2 may not necessarily be independent. They also may not be Gaussian distributed, contrary to the analytical components. Thus, when a regression procedure is applied, the explicit assumptions to take into account should be considered. In situations without random bias components of any significance, the relationships simplify to

$$\sigma^2_{ex1} = \sigma^2_{\text{A1}}$$

$$\sigma^2_{ex2} = \sigma^2_{\text{A2}}$$

In this situation, it usually can be assumed that the error distributions are Gaussian, and estimates of the analytical SDs may be available from QC data.

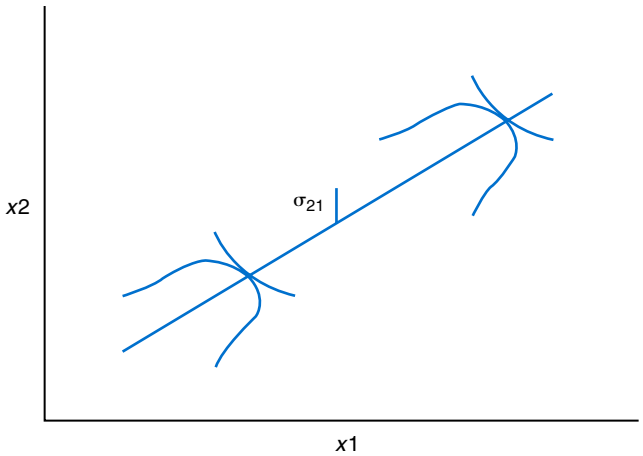


Fig. 2.12 Outline of the relation between $x1$ and $x2$ values measured by two assays subject to random errors with constant standard deviations over the analytical measurement range. A linear relationship between the modified target values ($X1'_{\text{Target}i}$, $X2'_{\text{Target}i}$) is presumed. The $x1_i$ and $x2_i$ values are Gaussian distributed around $X1'_{\text{Target}i}$ and $X2'_{\text{Target}i}$, respectively, as schematically shown. σ_{21} (σ_{yx}) is demarcated.

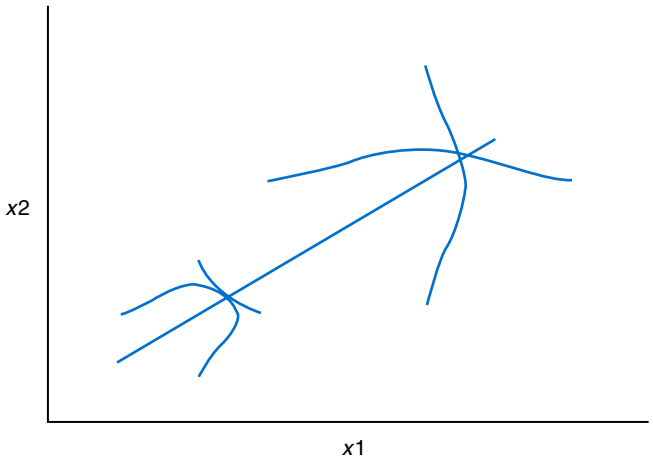


Fig. 2.13 Outline of the relation between $x1$ and $x2$ values measured by two assays subject to proportional random errors. A linear relationship between the modified target values is assumed. The $x1_i$ and $x2_i$ values are Gaussian distributed around $X1'_{\text{Target}i}$ and $X2'_{\text{Target}i}$, respectively, with increasing scatter at higher concentrations, as is shown schematically.

Another methodologic problem concerns the question of whether the dispersion of sample-related random bias and the analytical imprecision are constant or change with the analyte concentration. In cases with a considerable range (i.e., a decade or longer), this phenomenon should also be taken into account when a regression analysis is applied. Fig. 2.13 schematically shows how dispersions may increase proportionally with concentration.

Deming Regression Analysis and Ordinary Least-Squares Regression Analysis (Constant Standard Deviations)

To reliably estimate the relationship between modified target values (i.e., a_0 for α_0 and b for β), a regression procedure taking into account errors in both $x1$ and $x2$ is preferable (i.e., Deming approach) (see Fig. 2.12). Although the **ordinary**

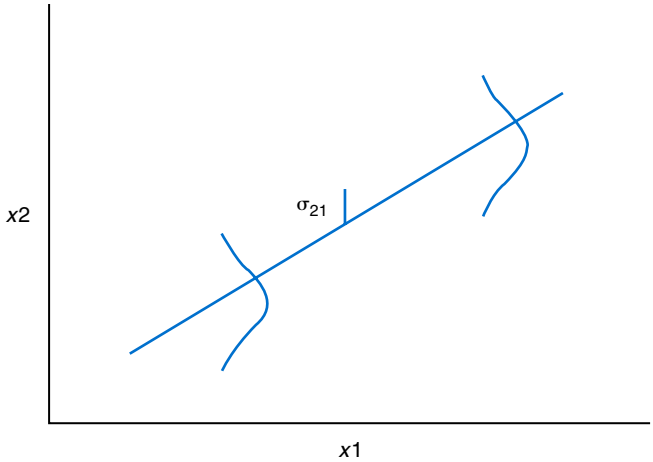


Fig. 2.14 The model assumed in ordinary least-squares regression. The $x2$ values are Gaussian distributed around the line with constant standard deviation over the analytical measurement range. The $x1$ values are assumed to be without random error. σ_{21} (σ_{yx}) is shown.

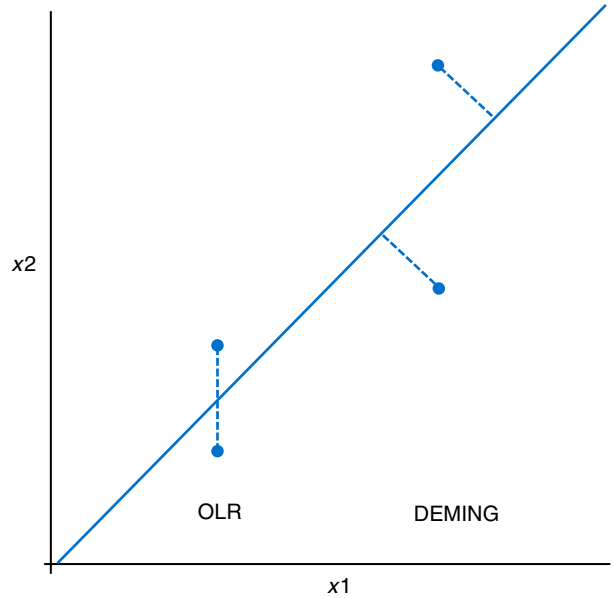


Fig. 2.15 In ordinary least-squares regression *(OLR)*, the sum of squared deviations from the line is minimized in the vertical direction. In Deming regression analysis, the sum of squared deviations is minimized at an angle to the line, depending on the random error ratio. Here the symmetric case is displayed with orthogonal deviations. (From Linnet, K. [1998]. The performance of Deming regression analysis in case of a misspecified analytical error ratio. *Clinical Chemistry, 44,* 1024–1031.)

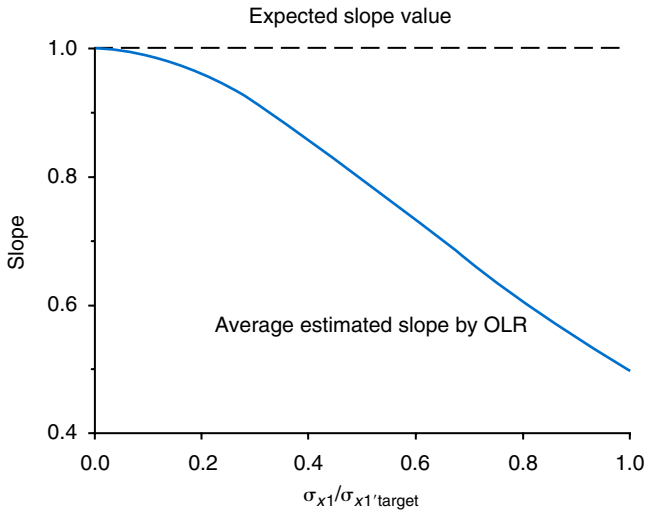


Fig. 2.16 Relations between the true (expected) slope value and the average estimated slope by ordinary least-squares regression *(OLR)*. The bias of the OLR slope estimate increases negatively for increasing ratios of the standard deviation (SD) random error in $x1$ to the SD of the $X1$ target value distribution.

least-squares regression (OLR) procedure is commonly used in method comparison studies, it does not take errors in $x1$ into account but is based on the assumption that only the $x2$ measurements are subject to random errors (Fig. 2.14). In the Deming procedure, the sum of squared distances from measured sets of values ($x1_i$, $x2_i$) to the regression line is minimized at an angle determined by the ratio between SDs for the random variations of $x1$ and $x2$. It can be proven theoretically that, given Gaussian *error* distributions, this estimation procedure is optimal. It should here be noted that it is the *error* distributions that should be Gaussian, not the dispersion of values over the measurement range. In Fig. 2.15, the symmetric case is illustrated with a regression slope of 1 and equal SDs for the random variations of $x1$ and $x2$, in which case the sum of squared distances is minimized orthogonally in relation to the line.

OLR regression is not recommended except in special situations. In OLR, the sum of squared distances is minimized in the vertical direction to the line (see Fig. 2.15). The neglect of the random error in $x1$ induces a downward biased slope estimate, which depends on the ratio between the SD for the random error in $x1$ and the SD of the $X1'$ target values. Fig. 2.16 shows the bias as a function of the ratio of the random error SD to the SD of the $X1'$ target value dispersion. For a ratio up to 0.1, the bias is less than 1%. At a ratio of 0.33, the bias amounts to 10%. In a given case, the analytical SD (e.g., from QC data) can be divided by the SD of the measured $x1$ values, which approximately equals the SD of $X1'$ target values. As an example, a typical comparison study for two serum sodium methods may be associated with a downward directed slope bias of about 10% (Fig. 2.17).

Computation Procedures for Ordinary Least-Squares Regression and Deming Regression

Assuming no errors in $x1$ and a Gaussian error distribution of $x2$ with constant SD throughout the analytical measurement range, OLR is the optimal estimation procedure. Given errors in both $x1$ and $x2$, the Deming approach is the method of choice. It should be noted for these parametric procedures that only the *error* distributions must be Gaussian or normal to ensure the nominal type I errors for associated statistical tests for slope and intercept hold true. The procedures are generally robust toward deviations from normality, but they are sensitive to outliers because of the squaring principle. Finally, the distribution of the $x1$ and $x2$ values over the measurement range does not have to be normal. A uniform distribution over the analytical

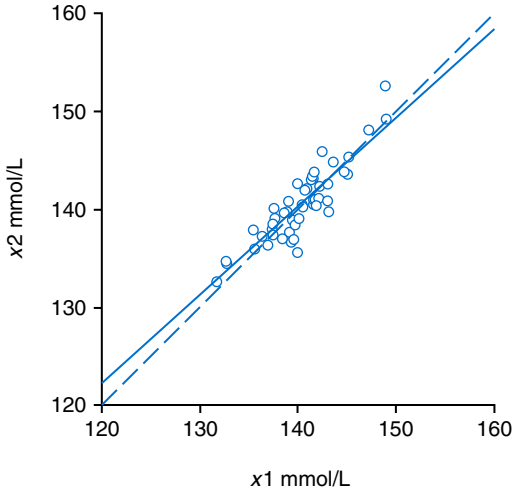


Fig. 2.17 Simulated comparison of two sodium methods. The *solid line* indicates the average estimated ordinary least-squares regression (OLR) line, and the *dotted line* is the identity line. Even though no systematic difference is evident between the two methods, the average OLR line deviates from the identity line corresponding to a downward slope bias of about 10%.

measurement range is generally of advantage, but the distribution in principle may take any form. For both procedures, we may evaluate the SD of the dispersion in the *vertical* direction around the line (commonly denoted $SD_{y \cdot x}$ and here given as SD_{21}). We have

$$SD_{21} = \left[\Sigma(x2_i - X2_{\text{Targetest}i})^2 / (N-2)\right]^{0.5}$$

Further discussion regarding the interpretation of SD_{21} will be given later.

To compute the slope in Deming regression analysis, the ratio between the SDs of the random errors of $x1$ and $x2$ is necessary, that is,

$$\lambda = \left(\sigma^2_{\text{RB1}} + \sigma^2_{\text{A1}}\right) / \left(\sigma^2_{\text{RB2}} + \sigma^2_{\text{A2}}\right)$$

SD_As can be estimated from duplicate sets of measurements as

$$SD^2_{\text{A1}} = (1/2N)\left[\Sigma(x1_{2i} - x1_{1i})^2\right]$$

$$SD^2_{\text{A2}} = (1/2N)\left[\Sigma(x2_{2i} - x2_{1i})^2\right]$$

If a specific value for λ is not available and the two routine methods that are compared are likely to be associated with random errors of the same order of magnitude, λ can be set to 1. The Deming procedure is generally relatively insensitive to a misspecification of the λ value.

Formulas for computing slope (β), intercept (α_0), and their SEs are available from other sources and are not provided here.

Evaluation of the Random Error Around an Estimated Regression Line

The estimated slope and intercept provide an estimate of the systematic difference or calibration bias between two methods over the analytical measurement range. An estimate of the random error is the dispersion around the line in the vertical direction, which is quantified as $SD_{y \cdot x}$ (here denoted SD_{21}). We have here without sample-related random interferences

$$\sigma^2_{21} = \beta^2 \sigma^2_{\text{A1}} + \sigma^2_{\text{A2}}$$

Thus σ_{21} reflects the random error both in $x1$ (with a rescaling) and in $x2$. Often β is close to unity, and in this case, σ^2_{21} becomes approximately the sum of the individual squared SDs. This relation holds true for both Deming and OLR analyses. Frequently, OLR is applied in situations associated with random measurement error in both $x1$ and $x2$, and in these situations, σ_{21} reflects the errors in both.

The presence of sample-related random interferences in both $x1$ and $x2$ gives the following expression:

$$\sigma^2_{21} = \left(\beta^2 \sigma^2_{\text{A1}} + \sigma^2_{\text{A2}}\right) + \left(\beta^2 \sigma^2_{\text{RB1}} + \sigma^2_{\text{RB2}}\right)$$

Thus the σ_{21} value is influenced by the slope value and the analytical error components σ_{A1} and σ_{A2} (grouped in the first bracket) and σ_{RB1} and σ_{RB2} (grouped in the second bracket). In many cases, the slope is close to unity, in which case we have simple addition of the components. Information on the analytical components is usually available from duplicate sets of measurements or from QC data. On this basis, the combined random bias term in the second bracket can be derived by subtracting the analytical components from σ_{21}. Overall, it can be judged whether the total random error is acceptable or not. The systematic difference can be adjusted for relatively easily by rescaling one of the sets of measurements. However, if the random error term is very large, such a rescaling does not ensure equivalency of measurements with regard to individual samples.

Assessment of Outliers

The distance from a suspected outlier to the line is recorded in SD units, and the outlier is rejected if the distance exceeds a predetermined limit (e.g., 3 or 4 SD units). In the case of OLR, the SD unit equals SD_{21}, and the vertical distance is considered. For Deming regression analysis, the unit is the SD of the deviation of the points from the line at an angle determined by the error variance ratio λ. A plot of these deviations, a so-called residuals plot, conveniently illustrates the occurrence of outliers. Fig. 2.18A illustrates an example of Deming regression analysis with occurrence of an outlier and the associated residuals plot (see Fig. 2.18B), which clearly shows the outlier pattern. In this example, the residuals plot was standardized to unit SD. Use of an outlier limit of 4 SD units in this example led to rejection of the outlier, and a reanalysis was undertaken. In this example, rejection of the outlier changed the slope from 1.14 to 1.03. With regard to outliers, the reason for their presence should be investigated as a method limitation (e.g., possibly a nonspecificity for the analyte).

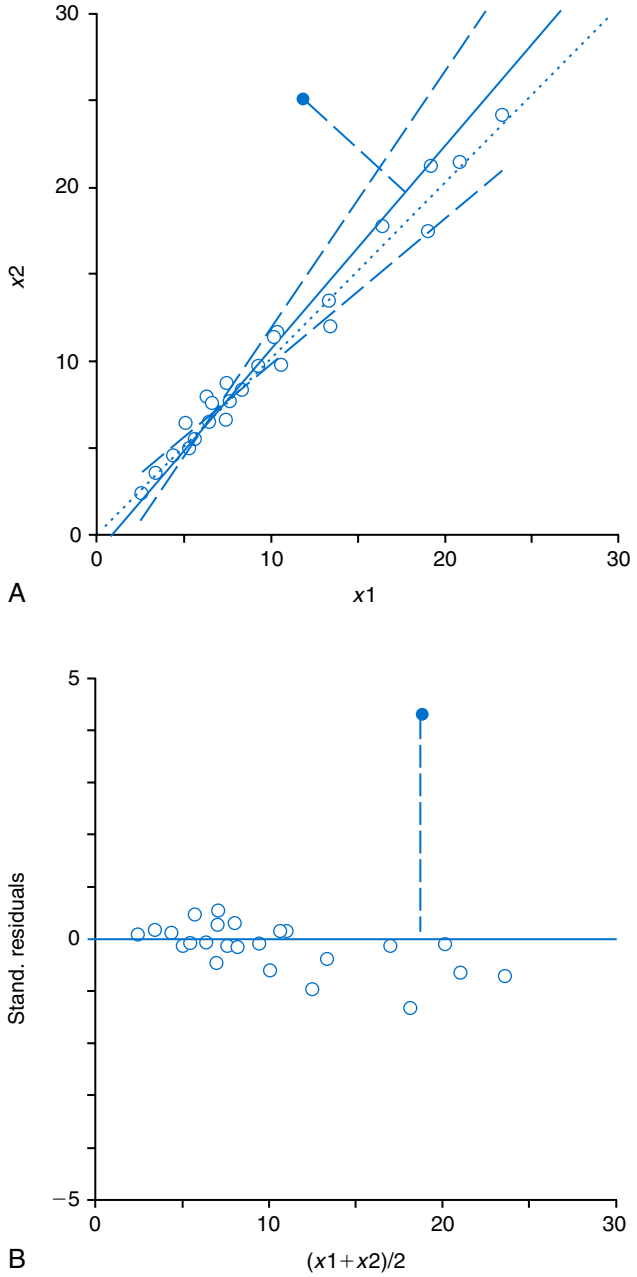


Fig. 2.18 (A) A scatter plot with the Deming regression line *(solid line)* with an outlier *(filled point)*. The *dotted straight line* is the diagonal, and the *curved dashed lines* demarcate the 95% confidence region. (B) Standardized residuals plot with indication of the outlier.

Correlation Coefficient

The ordinary **correlation coefficient**, ϱ, also called the Pearson product moment correlation coefficient, is estimated as r from sums of squared deviations for $x1$ and $x2$ values as follows:

$$r = p/(uq)^{0.5}$$

where

$$p = \Sigma\,(x1_i - x1_m)\,(x2_i - x2_m)$$

$$u = \Sigma(x1_i - x1_m)^2 \text{ and } q = \Sigma(x2_i - x2_m)^2$$

and

$$x1_m = \Sigma x1_i/N \text{ and } x2_m = \Sigma x2_i/N$$

The correlation coefficient r is a *relative* indicator of the amount of dispersion around the regression line. If the numeric interval of values is short, r tends to be low and vice versa for a long range of values. For example, consider simulated examples, where the random errors of $x1$ and $x2$ are the same but the width of the distributions of measured values differs (Fig. 2.19A and B). In A, the target values are uniformly distributed over the range 1 to 3, and in B, the range is 1 to 6. The random error SD is presumed constant, and it is set to 0.15 for both $x1$ and $x2$, corresponding to a CV of 5% at the value 3. Given sets of 50 paired measurements, the correlation coefficient is 0.93 in case *A* and 0.99 in case *B*. Furthermore, a single point located outside the range of the rest of the observations exerts a strong influence, resulting in a value of 0.97 (see Fig. 2.19C).

Although σ_{21} is the relevant measure for random error in method comparison studies, ϱ is still incorrectly used as a supposed measure of agreement between two methods. A systematic difference due to a difference with regard to calibration is not expressed through ϱ but solely in the form of an intercept (α_0) deviation from zero or a slope *(β)* deviation from unity.

Regression Analysis in Cases of Proportional Random Error

For analytes with extended ranges (e.g., 1 or several decades), the SD_A is seldom constant. Rather, a proportional relationship may apply. This may also be true for the random bias components. In this situation, the regression procedures described previously may still be used, but they are not optimal because the SEs of slope and intercept become larger than is the case when a weighted form of regression analysis is applied. Given a proportional relationship, a weighted procedure assigns larger weights to observations in the low range; low-range observations are more precise than measurements at higher concentrations that are subject to larger random errors. More specifically, weights are applied in the computations that are inversely proportional to the squared SDs (variances) that express the random error. In the weighted modification of the Deming procedure, distances from ($x1_i$, $x2_i$) to the line are inversely weighted according to the squared SDs at a given concentration (Fig. 2.20).

Testing for Linearity

Splitting of the systematic error into a constant and a proportional component depends on the assumption of linearity. A convenient test is a runs test, which in principle assesses whether negative and positive deviations from the points to the line are randomly distributed. The term *run* here relates to a sequence of deviations with the same sign. Consider, for example, the situation with a downward trend of $x2$ values at the upper end of the analytical measurement range (Fig. 2.21A). The SDs from the line (i.e.,

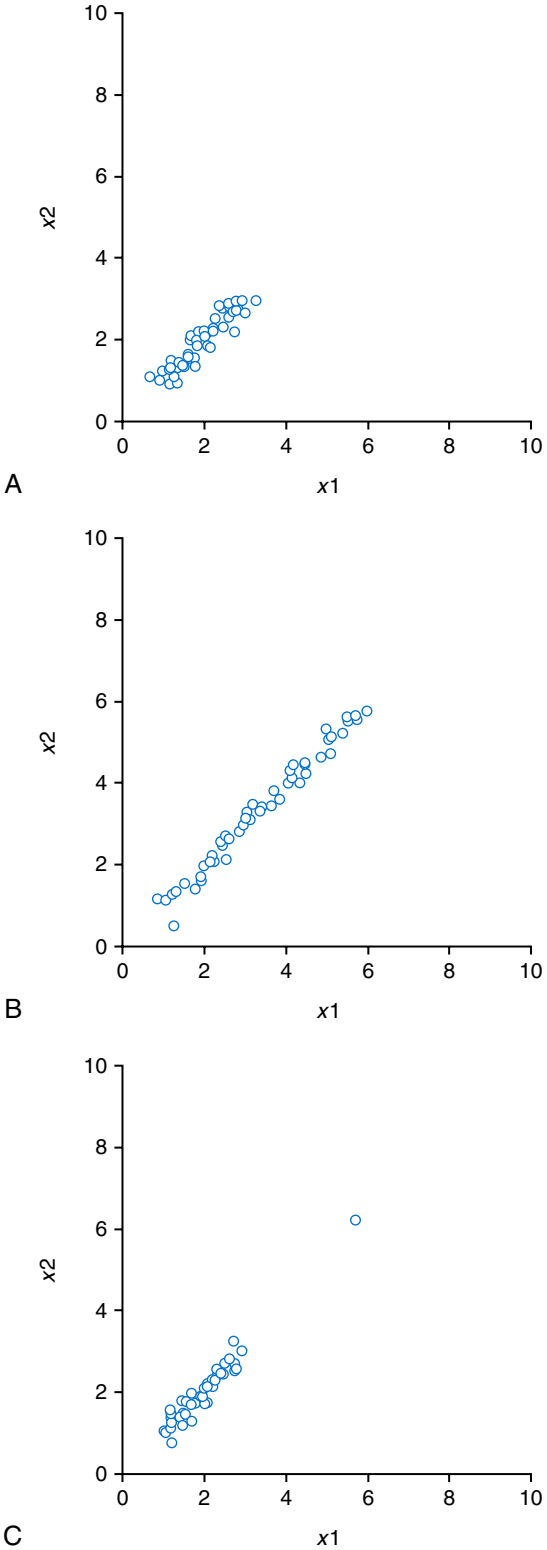


Fig. 2.19 Scatter plots illustrating the effect of the range on the value of the correlation coefficient ϱ. (A) Target values are uniformly distributed over the range 1 to 3 with random errors of both $x1$ and $x2$ corresponding to a standard deviation (SD) of 5% of the target value at 3 (constant error SDs). (B) The range is extended to 1 to 6 with the same random error levels. The correlation coefficient equals 0.93 in A and 0.99 in B. (C) The effect of a single aberrant point is shown. Forty-nine of the target values are distributed over the range 1 to 3, with a single point at 6. The correlation coefficient is 0.97.

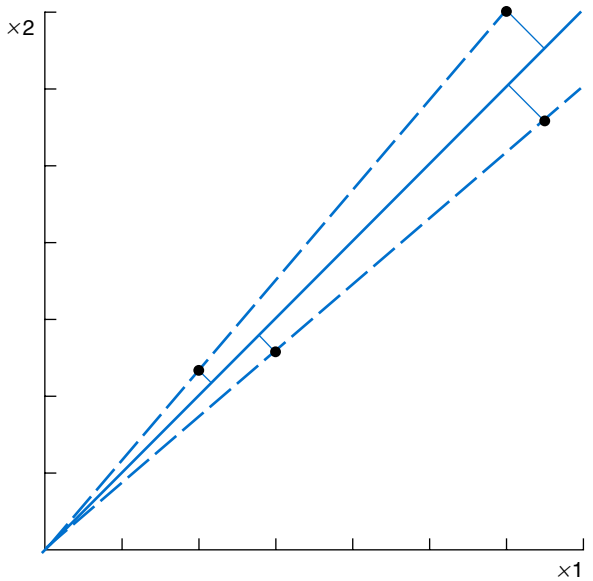


Fig. 2.20 Distances from data points to the line in weighted Deming regression assuming proportional random errors in $x1$ and $x2$. The symmetric case is illustrated with equal random errors and a slope of unity yielding orthogonal projections onto the line. (Modified from Linnet, K. [1999]. Necessary sample size for method comparison studies based on regression analysis. *Clinical Chemistry, 45,* 882–894. Used with permission.)

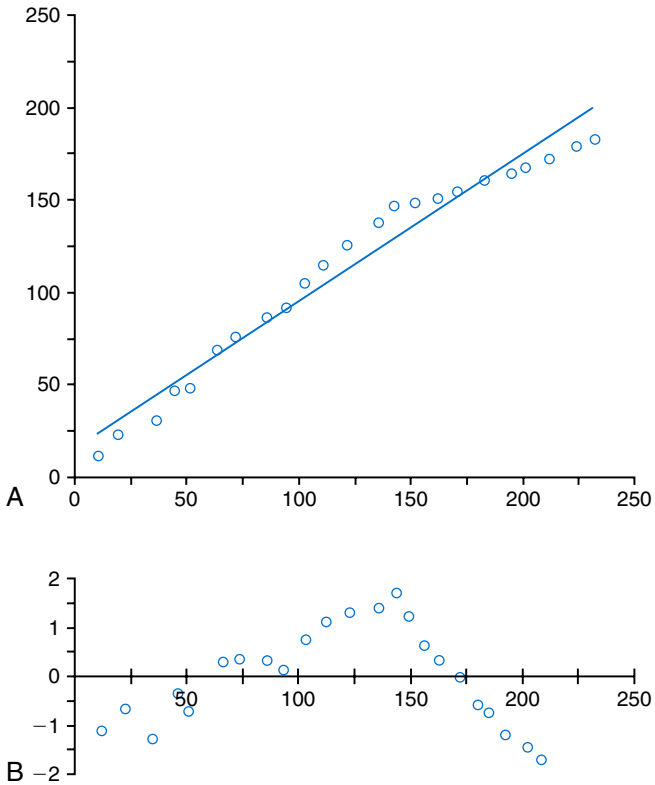


Fig. 2.21 (A) Scatter plot showing an example of nonlinearity in the form of downward-deviating $x2$ values at the upper part of the range. (B) Plot of residuals showing the effects of nonlinearity. At the upper end of the analytical measurement range, a sequence (run) of negative residuals is present.

the residuals) will tend to be negative in this area instead of being randomly distributed above and below the line (see Fig. 2.21B).

Nonparametric Regression Analysis (Passing-Bablok Procedure)

The slope and the intercept may be estimated by a nonparametric procedure, which is robust to outliers and requires no assumptions of Gaussian error distributions. The method takes measurement errors for both $x1$ and $x2$ into account, but it presumes that the ratio between random errors is related to the slope in a fixed manner:

$$\lambda = \left(SD^2_{RB1} + SD^2_{A1}\right) / \left(SD^2_{RB2} + SD^2_{A2}\right) = 1/\beta^2$$

Otherwise, a biased slope estimate is obtained. The procedure may be applied both in situations with random errors with constant SDs and in cases with proportional SDs. The method is not as efficient as the corresponding parametric procedures. Slope and intercept with CIs are provided, together with Spearman's rank correlation coefficient.

Interpretation of Systematic Differences Between Methods Obtained on the Basis of Regression Analysis

A systematic difference between two methods is identified if the estimated intercept differs significantly from zero or if the slope deviates significantly from 1. This is decided on the basis of t-tests:

$$t = (a_0 - 0)/SE(a_0)$$

$$t = (b - 1)/SE(b)$$

The t-tests can be supplemented with 95% CIs.

$SE(a_0)$ and $SE(b)$ are the SEs of the estimated intercept a_0 and the slope b, respectively. SEs can be derived by a computerized resampling principle called *the jackknife procedure*, which in practice can be carried out using appropriate software. Having estimated a_0 and b, we have the estimate of the systematic difference between the methods, D_c, at a selected concentration, $X1'_{Targetc}$:

$$D_c = X2'_{Targetestc} - X1'_{Targetc} = a_0 + (b - 1)\, X1'_{Targetc}$$

$X2'_{Targetestc}$ is the estimated $X2'$ target value at $X1'_c$. Note that D_c refers to the *systematic difference* (i.e., the difference between modified target values corresponding to a calibration difference). The SE of D_c can be derived by the jackknife procedure using a software program. By evaluating the SE throughout the analytical measurement range, a confidence region for the estimated line can be displayed. If method comparison is performed to assess the calibration to a reference measurement procedure, correction of a significant systematic difference $Delta_c$ will often be performed by recalibration [$x2_{rec} = (x1 - a_0)/b$]. The associated standard uncertainty is the SE of $Delta_c$.

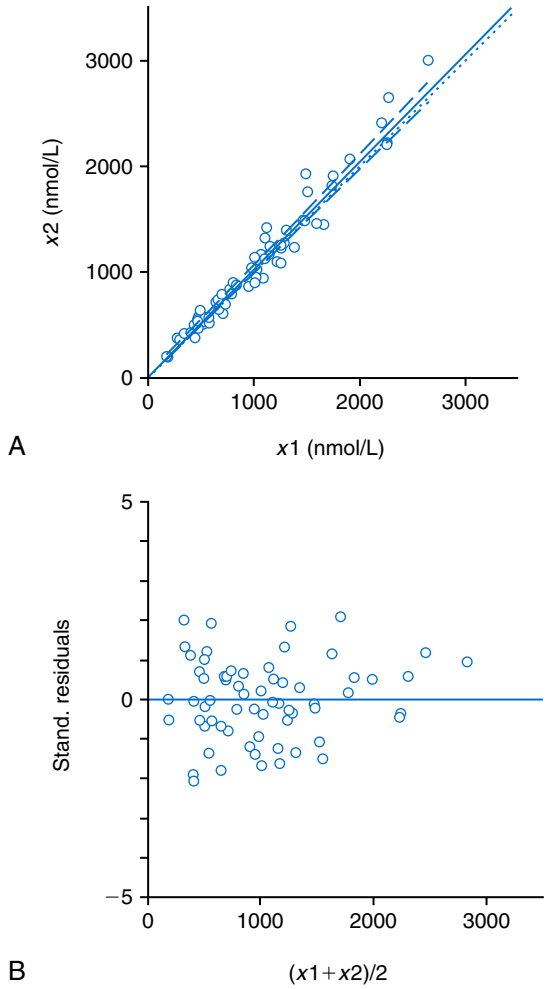


Fig. 2.22 An example of weighted Deming regression analysis for the comparison of drug assays. (A) The *solid line* is the estimated weighted Deming regression line, the *dashed curves* indicate the 95% confidence region, and the *dotted line* is the line of identity. (B) A plot of residuals standardized to unit standard deviation. The homogeneous scatter supports the assumed proportional error model and the assumption of linearity.

Example of Application of Regression Analysis (Weighted Deming Analysis)

Application of weighted Deming regression analysis may be illustrated by the comparison of drug assays example [N = 65 ($x1$, $x2$) single measurements]. As outlined in the section on the Bland-Altman plot (see Fig. 2.10), in this example the random error of the differences increases with the concentration, suggesting that the weighted form of Deming regression analysis is appropriate. Fig. 2.22 shows (A) the estimated regression line with 95% confidence bands and (B) a plot of normalized residuals. The nearly homogeneous scatter in the residuals plot supports the assumed proportional random error model and the assumption of linearity. The slope estimate (1.014) is not significantly different from 1 (95% CI: 0.97 to 1.06), and the intercept is not significantly different from zero (95% CI: −6.7 to 47.4) (Table 2.2). A runs test for linearity does not contradict the assumption of linearity. The amount of random error is quantified in the form

TABLE 2.2 Results of Weighted Deming Regression Analysis for the Comparison of Drug Assays Example, *n* = 65 Single (*x*1, *x*2) Measurements

	Estimate	SE	95% CI
Slope *(b)*	1.014	0.022	0.97–1.06
Intercept (a_0)	20.3	13.5	−6.7 to 47.4
Weighted correlation coefficient	0.98		
SD_{21} proportionality factor	0.11		
Runs test for linearity	NS		
$Delta_c = X_2 - X_1$ at $X_c = 300$	24.6	9.5	5.72–43.6
$Delta_c = X_2 - X_1$ at $X_c = 2000$	48.9	34.2	−19.3 to 117

CI, Confidence interval; *NS*, not significant; *SD*, standard deviation; *SE*, standard error.

of the SD_{21} proportionality factor equal to 0.11, or 11%. In the present example, with a slope close to unity and two routine methods with assumed random errors of about the same magnitude, we divide the random error by the square root of 2 and get $CV_{x1} = CV_{x2} = 7.8\%$. QC data in the laboratory have provided CV_As of 6.1% and 7.2% for methods 1 and 2, respectively. Thus in this example, the random error may be attributed largely to analytical error. The assay principle for both methods is high performance liquid chromatography (HPLC), which generally is a rather specific measurement principle; considerable random bias effects are not expected in this case.

In Table 2.2, estimated systematic differences at the limits of the therapeutic interval (300 and 2000 nmol/L) are displayed (24.6 and 48.9 nmol/L, respectively). This corresponds to percentage values of 8.2% and 2.4%, respectively. Estimated SEs by the jackknife procedure yield the 95% CIs, as shown in the table. At the low concentration, the difference is significant (95% CI: 5.7 to 44 nmol/L; does not include zero), which is not the case at the high level (95% CI: −19 to 117 nmol/L). Even though the intercept and slope estimates separately are not significantly different from the null hypothesis values of 0 and 1, respectively, the combined difference $Delta_c$ is significant at low concentrations in this example. If the difference is considered of medical importance and both methods are to be used simultaneously in the laboratory, recalibration of one of the methods might be considered.

MONITORING SERIAL RESULTS

An important aspect of clinical chemistry is monitoring of disease or treatment (e.g., tumor markers in cases of cancer, drug concentrations in cases of therapeutic drug monitoring). To assess changes in a rational way, various imprecision components have to be taken into account. Biologic within-subject variation (SD_I) and preanalytical (SD_{PA}) and analytical variation (SD_A) all have to be recognized. Assuming that preanalytical variation is already included in the estimated within-subject variation SD, a total SD (SD_T) can be estimated as follows:

$$SD_T^2 = SD_{WithinB}^2 + SD_A^2$$

The limit for statistically significant changes then is [*fx*2], where *k* depends on the desired probability level. Considering a two-sided 5% level, *k* is 1.96. The corresponding one-sided factor is 1.65. If a higher probability level is desired, *k* should be increased.

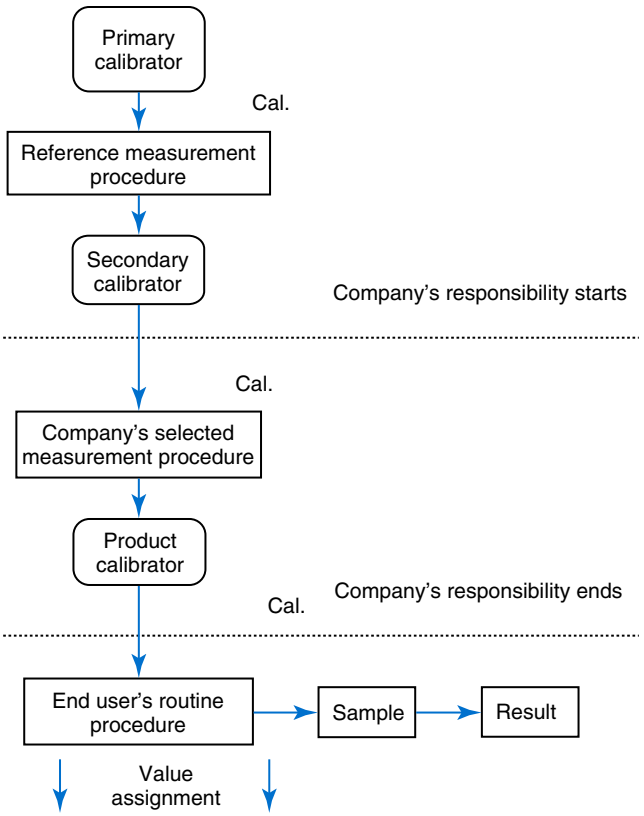


Fig. 2.23 The calibration hierarchy from a reference measurement procedure to a routine assay. The uncertainty increases from top to bottom. *Cal.*, Calibration.

TRACEABILITY

To ensure reasonable agreement between measurements of routine methods, the concept of traceability comes into focus. (See also Chapter 7 for further detail.) **Traceability** is based on an unbroken chain of comparisons of measurements leading to a known reference value (Fig. 2.23). For well-established analytes, a hierarchy of methods exists with *a reference measurement procedure* at the top, *selected measurement procedures* at an intermediate level, and finally *routine measurement procedures* at the bottom. A **reference measurement procedure** is a fully understood procedure of highest analytical quality containing a complete uncertainty budget given in Système Internationale (SI) units. Reference procedures are used to measure the analyte concentration in *secondary reference materials*, which typically have the same matrix as samples that are to be measured by routine procedures (e.g., human serum). Secondary reference materials are usually of high

analytical quality, and certified secondary reference materials must be validated for **commutability** with clinical samples if they are intended for use as trueness controls for routine methods. Otherwise, their use is restricted to selected measurement procedures for which they are intended. The certificate of analysis should state the methods for which the secondary reference materials have been validated to be commutable with clinical samples. When no information is given for commutability, it must be assumed that the reference material is not commutable with clinical samples, and the user has the responsibility to validate commutability for the methods of interest.

Using cortisol as an example, the primary reference material is crystalline cortisol with a chemical analysis for impurities (National Institute of Standards and Technology [NIST] standard reference material 921, cortisol [hydrocortisone]). A primary calibrator is then a cortisol preparation with a stated mass fraction (purity) (e.g., 0.998 and a 95% CI of ±0.001). The reference measurement procedure is an isotope-dilution gas chromatography–mass spectrometry method that is calibrated with the primary calibrator. A panel of individual frozen serum samples that have values assigned by the primary reference measurement procedure is available from the Institute for Reference Materials and Measurements (IRMM) as secondary reference materials (European Reference Material [ERM]-DA451/International Federation of Clinical Chemistry and Laboratory Medicine [IFCC]). A manufacturer's *selected measurement* procedure is calibrated with the secondary reference materials and is used for measurement of the quantity in the manufacturer's *product calibrator*, which is the calibrator used for the routine method in clinical laboratories.

In case a reference measurement procedure exists for an analyte (measurand), comparable results among measurement procedures can be achieved as described earlier, so-called standardization. When reference measurement procedures are not available, so-called harmonization refers to the process of establishing comparable results among measurement procedures for the given analyte. Harmonization is typically based on distribution among laboratories of commutable secondary reference materials with arbitrarily set target values (see Chapter 7).

Uncertainty Concept

According to the ISO's "Guide to the Expression of Uncertainty in Measurement" (GUM), **uncertainty** is formally defined as "a parameter associated with the result of a measurement that characterizes the dispersion of the values that could reasonably be attributed to the measurand." In practice, this means that the uncertainty is given as an interval around a reported laboratory result that specifies the location of the true value with a given probability (e.g., 95%). In general, the uncertainty of a result, which is traceable to a particular reference, is the uncertainty of that reference together with the overall uncertainty of the traceability chain.

In the outline of the uncertainty concept, it is assumed that any known systematic error components of a measurement method have been corrected, and the specified uncertainty includes uncertainty associated with correction of the systematic error(s). A distinction between type A and B uncertainties is made. Type A uncertainties are frequency-based estimates of SDs (e.g., an SD of the imprecision). Type B uncertainties are uncertainty components for which frequency-based SDs are not available. Instead, uncertainty is estimated by other approaches or by the opinion of experts. Finally, the total uncertainty is derived from a combination of all sources of uncertainty and can be expressed as a *standard uncertainty* (u_{st}), which is equivalent to an SD. By multiplication of a standard uncertainty with a *coverage factor (k)*, the uncertainty corresponding to a specified probability level is derived, e.g. multiplication with a coverage factor of 2 yields a probability level of ≈95%, given a Gaussian distribution. When the total uncertainty of an analytical result obtained by a routine method is considered (u_{st}), preanalytical variation (u_{PAst}), method imprecision (u_{Ast}), sample-related random interferences (u_{RBst}), and uncertainty related to calibration and bias corrections (traceability) (u_{Tracst}) should be taken into account. In expressing the uncertainty components as standard uncertainties, we have:

$$u_{st} = \left[u_{PAst}^2 + u_{Ast}^2 + u_{RBst}^2 + u_{Tracst}^2\right]^{0.5}$$

In principle, uncertainty can be judged *directly* from measurement comparisons ("top down") or *indirectly* from an analysis of individual error sources according to the law of error propagation ("error budget," "bottom up").

Example of Direct Assessment of Uncertainty on the Basis of Measurements of a Commutable Certified Reference Material

Suppose a CRM is available that was validated to be *commutable* with patient samples for a given routine method with a specified value 10.0 mmol/L and a standard uncertainty of 0.2 mmol/L. Ten repeated measurements in independent runs give a mean value of 10.3 mmol/L with SD 0.5 mmol/L. The SE of the mean is then [*fx*3]. The mean is not significantly different from the assigned value [$t = (10.3 - 10.0)/(0.2^2 + 0.16^2)^{0.5} = 1.17$]. The total standard uncertainty with regard to traceability is then $u_{Tracst} = (0.16^2 + 0.2^2)^{0.5} = 0.26$ mmol/L. If the bias had been significant, a correction to the method could have been carried out, and the standard uncertainty would then be the same at the given concentration. Thus measurements of the CRM provide an estimate of the uncertainty related to traceability, *given the assumption of commutability with patient samples*. The other components have to be estimated separately. Concerning method imprecision, long-term imprecision (e.g., observed from QC measurements) should be used rather than the short-term SD observed for CRM material. Here we suppose that the long-term SD_A is 0.8 mmol/L. Data on preanalytical variation can be obtained by sampling in duplicates from a series of patients or can be a matter of judgment (type B uncertainty) based on literature data or data on similar analytes. We here suppose that SD_{PA} equals half the analytical SD (i.e., 0.4 mmol/L). Finally, we lack data on a possible sample-related random bias component,

which we may choose to ignore in the present example. The standard uncertainty of the results then becomes

$$u_{st} = \left[u_{PAst}^2 + u_{Ast}^2 + u_{RBst}^2 + u_{Tracst}^2\right]^{0.5}$$
$$= \left(0.4^2 + 0.8^2 + 0.26^2\right)^{0.5}$$
$$= 0.93 \text{ (mmol/L)}$$

In this case, the major uncertainty component is the long-term imprecision in the laboratory. To attain a reasonably precise uncertainty estimate, estimated SDs should be based on an appropriate number of repetitions. In the subsection on method precision, it can be seen that $N = 30$ repetitions provides SD estimates with 95% CIs extending from about 20% below to 35% above an estimated value (see Fig. 2.7), which may be regarded as reasonable.

Indirect Evaluation of Uncertainty by Quantification of Individual Error Source Components

On the basis of a detailed quantitative model of the analytical procedure, the standard approach is to assess the standard uncertainties associated with individual input parameters and combine them according to the law of propagation of uncertainties. The relationship between the combined standard uncertainty $u_c(y)$ of a value y and the uncertainty of the *independent* parameters $x_1, x_2, \ldots x_n$, on which it depends, is

$$u_c\left[y(x_1, x_2, \ldots)\right] = \left[\Sigma c_i^2 u(x_1)^2\right]^{0.5}$$

where c_i is a sensitivity coefficient (the partial differential of y with respect to x_i). These sensitivity coefficients indicate how the value of y varies with changes in the input parameter x_i. If the variables are not independent, the relationship becomes

$$u_c\left[y(x_1, x_2, \ldots)\right] = \left[\Sigma c_i^2 u(x_1)^2 + \Sigma c_i c_k u(x_i, x_k)^2\right]^{0.5}$$

where $u(x_i, x_k)$ is the covariance between x_i and x_k, and c_i and c_k are the sensitivity coefficients. The covariance is related to the correlation coefficient ϱ_{ik} by

$$u(x_i, x_k) = u(x_i)\, u(x_k)\, \rho_{ik}$$

This is a complex relationship that usually will be difficult to evaluate in practice. In many situations, however, the contributing factors are independent, thus simplifying the picture. Below, some simple examples of combined expressions are shown. The rules are presented in the form of combining SDs or CVs given *independent* input components.

$$q = x + y \quad \text{SD}(q) = \left[\text{SD}(x)^2 + \text{SD}(y)^2\right]^{0.5}$$

$$q = x - y \quad \text{SD}(q) = \left[\text{SD}(x)^2 + \text{SD}(y)^2\right]^{0.5}$$

$$q = ax \quad \text{SD}(q) = a\text{SD}(x) \quad \text{and} \quad \text{CV}(q) = \text{CV}(x)$$

$$q = x^p \quad \text{CV}(q) = p\ \text{CV}(x)$$

$$q = xy \quad \text{CV}(q) = \left[\text{CV}(x)^2 + \text{CV}(y)^2\right]^{0.5}$$

$$q = x/y \quad \text{CV}(q) = \left[\text{CV}(x)^2 + \text{CV}(y)^2\right]^{0.5}$$

The formulas shown may be used, for example, to calculate the combined uncertainty of a calibrator solution from the uncertainties of the reference compound, the weighting, and dilution steps. In some situations, a Monte Carlo simulation model of a complex analytical method may be established to estimate the combined uncertainty of the method on the basis of input uncertainties.

POINTS TO REMEMBER

- For well-established analytes, a hierarchy of methods exists with *a reference measurement procedure* at the top, *selected measurement procedures* at an intermediate level, and finally *routine measurement procedures* at the bottom.
- The uncertainty is given as an interval around a reported laboratory result that specifies the location of the true value with a given probability (e.g., 95%).
- The uncertainty of a result, which is traceable to a particular reference, is the uncertainty of that reference together with the overall uncertainty of the traceability chain.
- The uncertainty can be judged *directly* from measurement comparisons ("top down") or *indirectly* from an analysis of individual error sources according to the law of error propagation ("error budget," "bottom up").

DIAGNOSTIC ACCURACY OF LABORATORY TESTS

We here consider the basic steps for evaluation of the clinical accuracy of laboratory tests. In diagnostic accuracy studies, the measurements or results of one (or more) laboratory test under evaluation (i.e., the so-called index test) are compared with the results of a reference standard or method. This reference is the best prevailing test or strategy that is used to establish the presence or absence of the disease of interest (i.e., the so-called target disease that is to be detected or excluded by the index tests). This reference standard is conducted and its results interpreted as blindly for and independently from the index test(s) results as possible. Test accuracy studies show the concordance in results of the index test(s) with the presence or absence of disease as defined by the reference standard results. These studies provide information regarding the frequency of types of errors (i.e., false positive and negative test results) by the index test in relation to the reference standard.

Diagnostic Accuracy, Sensitivity, and Specificity of a Test in Isolation

In a diagnostic accuracy study the results of the index test are compared with those of a reference test in the same

	Disease status	
Test result	**Diseased**	**Nondiseased**
Positive	TP	FP
Negative	FN	TN

Fig. 2.24 The basic 2-by-2 table for estimating the diagnostic accuracy of a dichotomized quantitative test result. Positive test results are divided into true positives *(TPs)* and false positives *(FPs)* and negative results into true negatives *(TNs)* and false negatives *(FNs)*. (From Linnet, K., Bossuyt, P. M., Moons, K. G., & Reitsma, J. B. [2012]. Quantifying the accuracy of a diagnostic test or marker. *Clinical Chemistry, 58,* 1292–1301.)

individuals, all of whom are suspected to have the target disease (the suspected disease cohort design). The simplest situation is a comparison of a single index test, with only two result categories (i.e., a dichotomous or binary index test) to a reference standard (i.e., a single-test accuracy study). The ideal dichotomous index test correctly identifies all individuals as diseased or nondiseased with an error rate of zero. A zero error rate is only possible when there is no overlap between index test results in the diseased and nondiseased individuals. However, when there is overlap in index test results, some individuals are classified wrongly as shown below in an example concerning the diagnosis of deep venous thrombosis (DVT) using a D-dimer index test. When using a quantitative (continuous) index test to classify individuals as diseased or nondiseased, a cutoff value needs to be chosen to estimate these error rates. This results in a so-called dichotomized index test.

Values of the dichotomous or dichotomized index test that exceed the cutoff in individuals having the target disease are classified as true positives (TP) (Fig. 2.24). Similarly, index test results lower than the cutoff in nondiseased individuals are true negatives (TN). Accordingly, index test results below the cutoff in truly diseased subjects are false negative (FN), and index test results exceeding the cutoff in truly nondiseased subjects are false positive (FP). Based on the frequencies of FN and FP results, an overall error rate or non-error rate can be derived. The overall diagnostic accuracy of an index test is then defined as the fraction of true classifications out of all classifications:

$$\text{Diagnostic accuracy} = (TN + TP) / (TN + TP + FP + FN)$$

This is an overall non-error rate that can be subdivided into the non-error rate of the nondiseased individuals, which is the specificity of the test and the non-error rate of diseased individuals which is the sensitivity of the test

$$\text{Specificity} = TN / (TN + FP)$$

$$\text{Sensitivity} = TP / (TP + FN)$$

Whereas a very specific test provides negative results for all or almost all subjects who are free of the target disease, a very sensitive test detects all or almost all diseased subjects. To assess the (im)precision of these estimates, CIs should be specified as described under *Categorical Variables*. Table 2.3 displays the widths of the 95% CIs at various sample sizes of 20 to 1000 for two selected proportions. The specificity and sensitivity of two tests applied in the same study subjects can be statistically compared using the McNemar's test.

TABLE 2.3 Relationship Between Sample Size and 95% Confidence Intervals of a Proportion (e.g., a Sensitivity or Specificity): Selected Examples of Proportions of 0.05 and 0.80

Sample Size	95% CI of a Proportion of 0.05	95% CI of a Proportion of 0.80
20	0.00–0.25	0.56–0.94
60	0.01–0.14	0.68–0.90
100	0.02–0.11	0.71–0.87
500	0.03–0.07	0.76–0.83
1000	0.04–0.07	0.77–0.82

CI, Confidence interval.

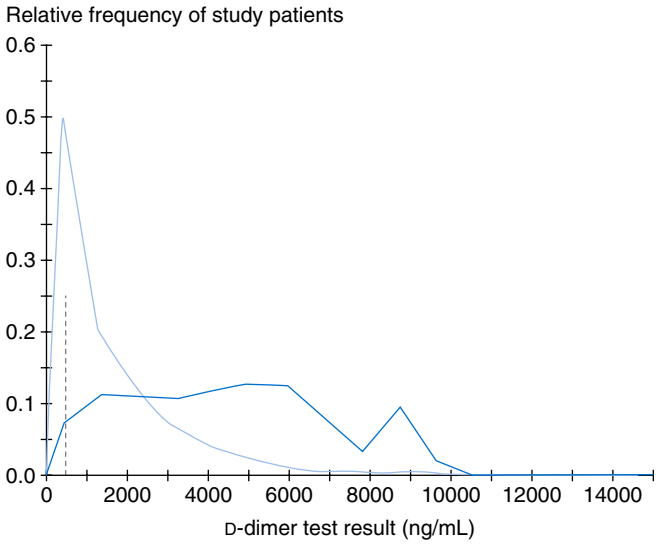


Fig. 2.25 Distribution of the quantitative D-dimer values for deep venous thrombosis (DVT) and non-DVT subjects in the example study. *Light blue line,* non-DVT; *blue line,* DVT. The *dashed line* indicates the commonly used cutoff value of 500 μg/L. (From Linnet, K., Bossuyt, P. M., Moons, K. G., & Reitsma, J. B. [2012]. Quantifying the accuracy of a diagnostic test or marker. *Clinical Chemistry, 58,* 1292–1301.)

Clinical Example: Accuracy of D-Dimer Test in Diagnosis of Deep Venous Thrombosis

We illustrate the concepts using some of the empirical data of a previously published study in primary care patients suspected of having DVT, the target disease (Fig. 2.25).

The study consisted of 2086 patients suspected of DVT, where DVT was defined as present in patients manifesting at least one of the following symptoms or signs: presence of swelling, redness, or pain in the leg. All patients were given a standardized diagnostic workup, including medical history; clinical examination; and testing for D-dimer, the (quantitative) index test. The reference procedure consisted of repeated compression ultrasonography tests and was performed in all patients, blinded to and independent of the index test results. A total of 416 (20%) of the 2086 included patients had DVT.

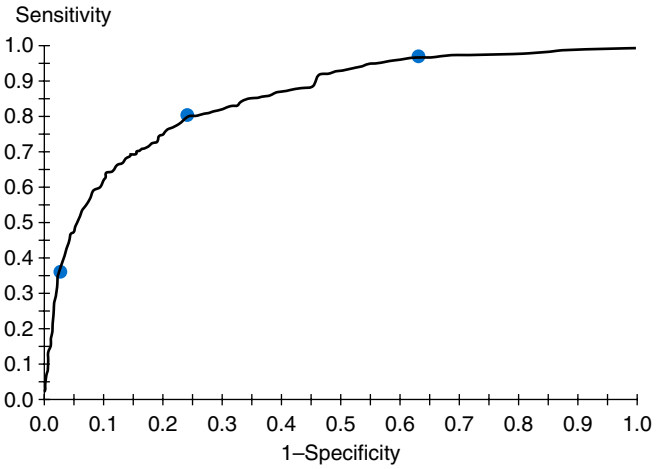


Fig. 2.26 Receiver operating characteristic curve of the D-dimer assay result for diagnosis of deep venous thrombosis in our example study. The *blue markers* correspond to various cutoff choices (from left to right, 5435 μg/L, 2133 μg/L, and 500 μg/L). (From Linnet, K., Bossuyt, P. M., Moons, K. G., & Reitsma, J. B. [2012]. Quantifying the accuracy of a diagnostic test or marker. *Clinical Chemistry, 58,* 1292–1301.)

Applying a commonly used cutoff of 500 μg/L or greater for the (originally) quantitative D-dimer assay (dashed line in Fig. 2.25), the sensitivity was 0.97 (i.e., 3% of the subjects with DVT had a value <500 μg/L). The specificity was only 0.37. The resulting overall diagnostic accuracy was 0.50. Whereas the test displayed good sensitivity at this threshold, detecting all but 3% of those having DVT, its specificity at this test threshold was relatively low, resulting in many FP results. The SEs were 0.012 for the specificity and 0.008 for the sensitivity, resulting in CIs of 0.356 to 0.402 and 0.955 to 0.987, respectively.

Receiver Operating Characteristic Curves

For a quantitative index test, the specificity and sensitivity depend on the selected cutoff point. A plot of the sensitivity and specificity pairs for all possible cutoff values over the measurement range provides the so-called **receiver operating characteristic (ROC) curve** (Fig. 2.26) Usually, sensitivity *(y)* is plotted against (1 – specificity) *(x)* at each possible cutoff value. The better the performance of the test, the higher the ROC curve is located in the left, upper region of the plot. With use of the ROC curve, an appropriate combination of specificity and sensitivity may be chosen, and the corresponding cutoff then selected.

An area under the ROC curve (i.e., the ROC area or so-called concordance or c-index) can be assessed by either parametric or nonparametric statistics. Given an SE of the ROC area or c-index, it is possible to test whether the area significantly exceeds 0.5, which would demonstrate that the index test performs better than chance. A worthless test has an area of 0.5. Furthermore, using the SE, also a 95% CI can be derived for the ROC area or c-index. For the D-dimer test example, the area under the ROC curve was 0.86 (SE, 0.011), with a 95% CI of 0.84 to 0.88.

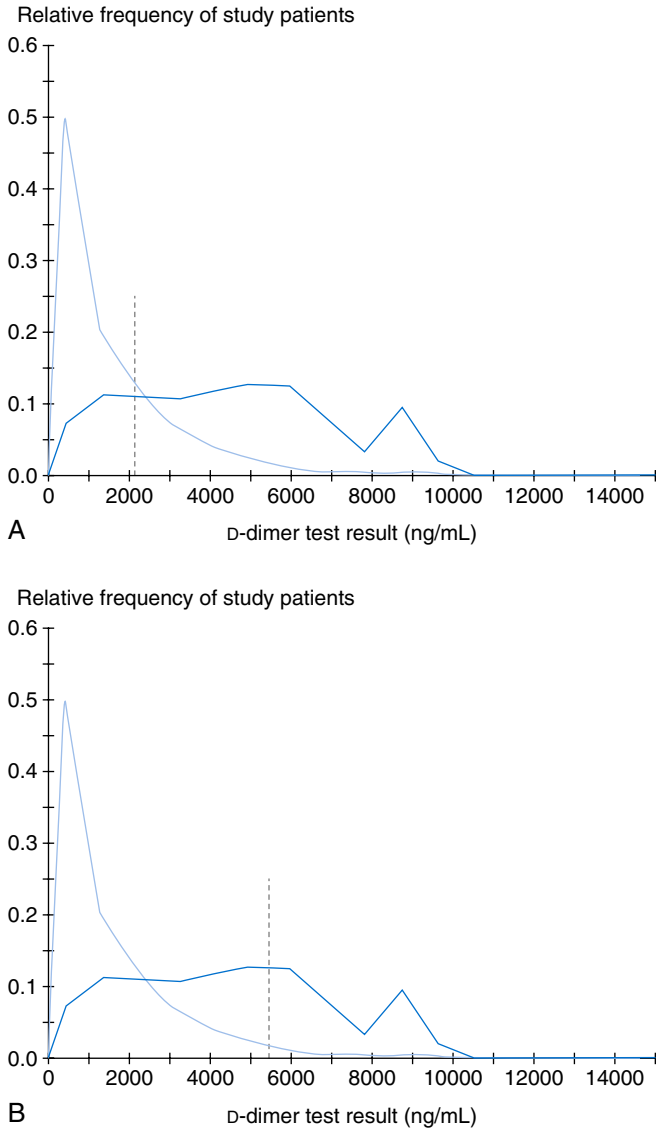


Fig. 2.27 Alternative cutoffs to 500 μg/L in the D-dimer example. (A) Cutoff (2133 μg/L) giving maximum value of the sum of the specificity and sensitivity. (B) Cutoff (5435 μg/L) providing a high specificity (0.975). *Light blue line,* non–deep venous thrombosis (DVT); *blue line,* DVT. The *dashed line* indicates the cutoff value. (From Linnet, K., Bossuyt, P. M., Moons, K. G., & Reitsma, J. B. [2012]. Quantifying the accuracy of a diagnostic test or marker. *Clinical Chemistry, 58,* 1292–1301.)

Selection of Cutoff Value in Case of Quantitative Index Tests

The specificity and sensitivity determined for an index test almost always vary inversely over the range of possible cutoffs. The cutoff point that provides the maximum of the sum of the specificity and sensitivity *could* be selected. In the D-dimer example, this cutoff would be close to 2000 μg/L, yielding a specificity of 0.76 and a sensitivity of 0.80 (Fig. 2.27A). However, this method of cutoff selection is commonly not recommended. The selection should rather be based on the intended purpose of the index test. If an index test is applied primarily to rule out the presence of disease (e.g., in the case of the D-dimer assay for exclusion of DVT), the cutoff point

should be at the lower end of the distribution of values of diseased individuals (see Fig. 2.25) (e.g., a cutoff of 500 μg/L). At this cutoff, the sensitivity approaches 1.0. But attaining such a high sensitivity is at the cost of a loss of specificity because of overlap of test values in the diseased and nondiseased individuals. Conversely, when FP results are judged unacceptable, the cutoff should be toward the upper limit of the distribution of values for the nondiseased group. For the D-dimer test example, a cutoff value corresponding to the 97.5 percentile of the distribution of values for those not having DVT (5435 μg/L) resulted in a specificity of 0.975, but now the sensitivity was only 0.36 (i.e., nearly the opposite of the situation with a cutoff of 500 μg/L) (see Fig. 2.27B).

Posterior Probabilities (Predictive Values)

Unlike sensitivity and specificity, the **positive predictive value** assesses the probability of having the disease given a positive test result, P (D | Tpos), whereas the **negative predictive value** assesses the probability of not having the disease given a negative test result P (Non – D | Tneg). The probability of presence of target disease given the index test result is an example of a so-called posterior disease probability, where the prior probability corresponds to the prevalence of the disease in the given situation. The **prevalence** of disease (P[D]) in the study sample is the a priori (pretest) probability of disease.

Given a positive test result (Tpos), the posterior disease probability is estimated as the fraction of TP out of all test result positives:

$$P(D \mid Tpos) = TP/(TP + FP)$$

Analogously for a negative result (Tneg), the probability that the given disease is absent is

$$P(Non - D \mid Tneg) = TN/(TN + FN)$$

Just as with sensitivity and specificity values, these posterior disease probabilities depend on the selected cutoff point for a quantitative test. In case of a dichotomous or dichotomized index test, these posterior probabilities are also called predictive values. They are highly dependent on the disease prevalence.

From the Bayes rule, the following relations exist:

$$P(D \mid Tpos) = [Sensitivity \times P(D)] / [Sensitivity \times P(D) + (1 - Specificity)(1 - P(D))]$$

$$P(Non - D \mid Tneg) = [Sensitivity \times (1 - P(D))] / [Specificity \times (1 - P(D)) + P(D) \times (1 - Specificity)]$$

Likelihood Ratios and Odds Ratios

From relative frequency distributions for results of the index test in the nondiseased and diseased groups, the so-called diagnostic **likelihood ratio** (LR) of an index test result *(X)* can be calculated as the ratio between the heights of the relative frequency *(f)* distributions at that specific test value. We get:

$$LR(X) = f_D(X) / f_{Non-D}(X)$$

In case the relative frequency of the distribution of diseased individuals is higher than that of the nondiseased individuals, the ratio exceeds 1. This indicates that disease is more likely than nondisease given this particular index test result. More formally, the ratio can be used to calculate posterior disease probabilities given specific values of the index test *(X)* and the disease prevalence (D):

$$P(D \mid X) = P(D) \times LR(X) / [P(D) \times LR(X) + (1 - P(D))]$$

or a more simple calculation can be carried out using odds instead of probabilities:

$$Odds(D \mid X) = Odds(D) \times LR(X)$$

based on the relation:

$$Odds = P/(1 - P)$$

Odds is an alternative way of expressing probabilities commonly used in betting games in Anglo-Saxon countries. For example, a probability of 0.80, or 80%, corresponds to an odds value of 4 according to the formula above. The higher the odds, the closer a probability is to one. From the equation, the posterior odds are equal to the prior odds multiplied by the diagnostic LR for the result *X*.

For a dichotomous or dichotomized index test, the following relationships apply:

$$LR(pos) = Sensitivity/(1 - Specificity)$$

$$LR(neg) = (1 - Sensitivity)/Specificity$$

A simple way of achieving the posttest probability of disease from the prevalence (pretest probability of disease) and the diagnostic LR is to use the Fagan nomogram. A recent example is the estimation of the probability of DVT from testing for D-dimer.

Comparison of Diagnostic Accuracy of Two Tests in Isolation

The diagnostic accuracy—that is, the ability to detect or exclude the target disease as determined by the reference method—of a new diagnostic index test is usually compared with another, established, index test. We here focus on the pure performances of the tests without consideration of other tests (i.e., we consider each test in isolation). When comparing the accuracy of two or more diagnostic index tests, a paired design is generally preferable for reasons of both validity and efficiency. In the target disease-suspected patients, the two index tests under comparison and the reference standard are performed on all subjects, again independently and blinded with regard to each other's test results. An example of a paired comparison is displayed in Fig. 2.28. Overall, the index test having the largest area under the ROC curve represents the best test. Preferably, CIs of areas and differences of areas should be provided.

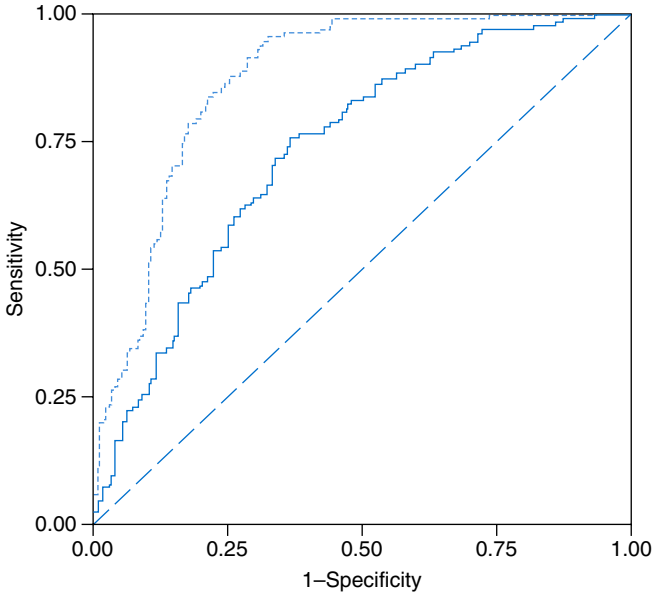


Fig. 2.28 Comparison of the receiver operating characteristic curves of two hypothetical index tests for the same target disease undertaken in the same individuals. The *dotted blue curve* represents a superior diagnostic test, both with regard to sensitivity and specificity over all possible cutoff points. The *dashed diagonal* represents a worthless test, with equal probability of a false-positive (1 – Specificity) and false-negative (1 – Sensitivity) result across all cutoff values (i.e., flipping a coin test). (From Linnet, K., Bossuyt, P. M., Moons, K. G., & Reitsma, J. B. [2012]. Quantifying the accuracy of a diagnostic test or marker. *Clinical Chemistry, 58*, 1292–1301.)

The STARD (Standards for Reporting of Diagnostic Accuracy Studies) initiative (Box 2.2) and the so-called QUADAS-2 (Quality Assessment Tool for Diagnostic Accuracy Studies) aim at improving diagnostic accuracy studies (http://www.quadas.org).

Diagnostic Accuracy of a Test in the Clinical Context

The diagnostic process commonly consists of a series of sequential steps in which much diagnostic information (i.e., diagnostic test results) is acquired. After each step, the physician intuitively judges the probability of the target disease being present. The initial step always consists of patient history and physical signs. If uncertainty about the presence and type of disease remains, subsequent tests are performed, often in another stepwise fashion. These supplementary tests may consist of simple blood or urine tests or be imaging, electrophysiology, or genetic tests or even later in the process more invasive testing such as biopsy, angiography, or arthroscopy. The supplementary information of each subsequent test is implicitly added to the already collected diagnostic information, and the target disease probability is constantly updated. This process continues until the target disease can be included or excluded with sufficient certainty.

Diagnostic test studies should reflect the steps in the diagnostic process so that the added value of such tests in excess of the information that is already present can be assessed. Depending on the situation, studies may reveal that the diagnostic information of any subsequent test is already supplied by the simpler previous test results. When regarded in isolation, such subsequent test or marker may indeed show diagnostic accuracy or value, but when assessed in the overall diagnostic workup, it does not. Such a case can arise because different tests may gauge the same underlying pathologic processes to varying degrees and thus provide related diagnostic information.

Clinical Example: Added Value of D-Dimer Testing in the Diagnosis of Suspected Deep Venous Thrombosis

The same DVT case study described earlier is considered. A total of 2086 patients were suspected of DVT, having at least one of the following symptoms: swelling, redness, or pain in the leg. All patients had a standardized diagnostic workup consisting of index tests from medical history taking, physical examination, and quantitative D-dimer testing. The reference standard was repeated compression ultrasonography, according to current clinical practice, carried out in all patients independent of the results of the index tests and blinded with regard to all preceding collected index test results. In total, 416 of the 2068 included patients (20%) had DVT confirmed by ultrasonography. We focus on estimating the added value of D-dimer testing to the information provided by history taking and physical examination.

A multivariable statistical approach is needed to assess the diagnostic accuracy of combined index test results. Logistic regression models express the probability of DVT (on the logit scale) as a linear function of the included index test results. Note that index test results may be included as binary, categorical, or even continuous results. Table 2.4 (model 1) shows the results from history and physical examination test results that were significantly related to DVT in the multivariable analysis, here defined as a multivariable **odds ratio** significantly ($P < .05$) different from 1 (no association).

To quantify whether the quantitative D-dimer assay result has added diagnostic value beyond the history and physical examination results combined, the basic model 1 was simply extended by including the index test D-dimer value, resulting in model 2 (see Table 2.4). After the inclusion of the D-dimer assay result, the regression coefficients of most history and physical tests in model 2 are found to be different from those in model 1: They now express the contribution of the corresponding test results, given a specific D-dimer result. This change reveals that the history and physical and the D-dimer results are indeed correlated and partly provide the same diagnostic information regarding whether DVT is present or not. The trend of lower regression coefficients of most findings can be interpreted as follows: A portion of the information supplied by the history and physical items is now replaced by the D-dimer assay result.

BOX 2.2 STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies

Title or Abstract

1. Identification as a study of diagnostic accuracy using at least one measure of accuracy (e.g., sensitivity, specificity, predictive values, AUC)

Abstract

2. Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)

Introduction

3. Scientific and clinical background, including the intended use and clinical role of the index test
4. Study objectives and hypotheses

Methods

Study Design

5. Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)

Participants

6. Eligibility criteria
7. On what basis potentially eligible participants were identified (e.g., symptoms, results from previous tests, inclusion in registry)
8. Where and when potentially eligible participants were identified (setting, location, and dates)
9. Whether participants formed a consecutive, random, or convenience series

Test Methods

10a. Index test, in sufficient detail to allow replication
10b. Reference standard, in sufficient detail to allow replication
11. Rationale for choosing the reference standard (if alternatives exist)
12a. Definition of and rationale for test positivity cutoffs or result categories of the index test, distinguishing prespecified from exploratory
12b. Definition of and rationale for test positivity cutoffs or result categories of the reference standard, distinguishing prespecified from exploratory
13a. Whether clinical information and reference standard results were available to the performers or readers of the index test
13b. Whether clinical information and index test results were available to the assessors of the reference standard

Analysis

14. Methods for estimating or comparing measures of diagnostic accuracy
15. How indeterminate index test or reference standard results were handled
16. How missing data on the index test and reference standard were handled
17. Any analyses of variability in diagnostic accuracy, distinguishing prespecified from exploratory
18. Intended sample size and how it was determined

Results

Participants

19. Flow of participants, using a diagram
20. Baseline demographic and clinical characteristics of participants
21a. Distribution of severity of disease in those with the target condition
21b. Distribution of alternative diagnoses in those without the target condition
22. Time interval and any clinical interventions between index test and reference standard

Test Results

23. Cross-tabulation of the index test results (or their distribution) by the results of the reference standard
24. Estimates of diagnostic accuracy and their precision (e.g., 95% CIs)
25. Any adverse events from performing the index test or the reference standard

Discussion

26. Study limitations, including sources of potential bias, statistical uncertainty, and generalizability
27. Implications for practice, including the intended use and clinical role of the index test

Other Information

28. Registration number and name of registry
29. Where the full study protocol can be accessed
30. Sources of funding and other support; role of funders

AUC, Area under the curve; *CI,* confidence interval; *STARD,* Standards for Reporting of Diagnostic Accuracy Studies.
From Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., et al. (2015). STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Clinical Chemistry, 61*(12), 1446–1452.

Diagnostic Accuracy of Combinations of Diagnostic Tests: Receiver Operating Characteristic Area

The multivariable diagnostic model, which is based on a combination of diagnostic index tests, as exemplified in models 1 and 2 in Table 2.4, can be considered as a single (overall or combined) quantitative index test, consisting of a composite of individual index tests. The test result of this "combined index test model" for each study patient is simply the calculated posterior probability of DVT presence given the observed pattern of the individual index test results in that patient. (See the footnote to Table 2.4 on how to calculate this probability of disease presence.)

As for single continuous index tests described earlier, also for "test combinations" combined into a single multivariable model, the ROC area (*c*-statistic) can be calculated to indicate the ability of this "test combination" to discriminate between

TABLE 2.4 Basic and Extended Multivariable Diagnostic Model to Discriminate Between Deep Venous Thrombosis Presence Versus Absence[a]

	MODEL 1 (BASIC MODEL)			MODEL 2 (BASIC MODEL + D-DIMER)		
	Regression Coefficient (SE)	OR (95% CI)	*P* value	Regression Coefficient (SE)	OR (95% CI)	*P* value
(Intercept)	−3.70 (0.26)	—	<.01	−4.94 (0.32)	—	<.01
Presence of malignancy	0.62 (0.22)	1.9 (1.2–2.9)	<.01	0.22 (0.26)	1.2 (0.7–2.1)	0.41
Recent surgery	0.44 (0.16)	1.6 (1.1–2.1)	<.01	0.003 (0.19)	1.0 (0.7–1.5)	0.99
Absence of leg trauma	0.75 (0.18)	2.1 (1.5–3.0)	<.01	0.67 (0.20)	2.0 (1.3–2.9)	<.01
Vein distension	0.48 (0.13)	1.6 (1.1—2.1)	<.01	0.25 (0.16)	1.3 (0.9–1.8)	0.12
Pain on walking	0.41 (0.15)	1.5 (1.1–2.0)	<.01	0.46 (0.18)	1.6 (1.1–2.3)	0.01
Swelling whole leg	0.36 (0.12)	1.4 (1.1–1.8)	<.01	0.47 (0.14)	1.6 (1.2–2.1)	<.01
Difference in calf circumference (per cm)	0.36 (0.04)	1.4 (1.3–1.5)	<.01	0.29 (0.04)	1.3 (1.2–1.4)	<.01
D-Dimer (per 500 ng/mL)	NA	NA	NA	0.29 (0.02)	1.3 (1.3–1.4)	<.01

[a]Exp (regression coefficient) is the odds ratio (OR) of a diagnostic test result. For example, an odds ratio of 2 for absence of leg trauma (model 2) means that a suspected patient without a recent leg trauma has a two times higher chance of having deep venous thrombosis (DVT) than a patient with a recent leg trauma (because in the latter the leg trauma would more likely be the cause of the presenting symptoms and signs). Similarly, an odds ratio of 1.3 for calf difference in cm (model 2) means that for every centimeter increase in calf circumference difference, a patient has a 1.3 times (or 30%) higher chance of having DVT.

A diagnostic model can be considered as a single overall or combined test consisting of different test results, with the probability of DVT presence as its test result. For example, for a male subject without malignancy, recent surgery, or leg trauma but with vein distension and a painful not swollen leg when walking with a calf difference of 6 cm the formula is (model 1):

$Z = -3.70 + 0.62*0 + 0.44*0 + 0.75*0 + 0.48*1 + 0.41*1 + 0.36*0 + 0.36*6 = -0.65$

The probability for this patient of the presence of DVT based on the basic model then is exp(−0.65)/(1 + exp[−0.65]) = 34%.

CI, Confidence interval; *NA,* not applicable; *SE,* standard error.

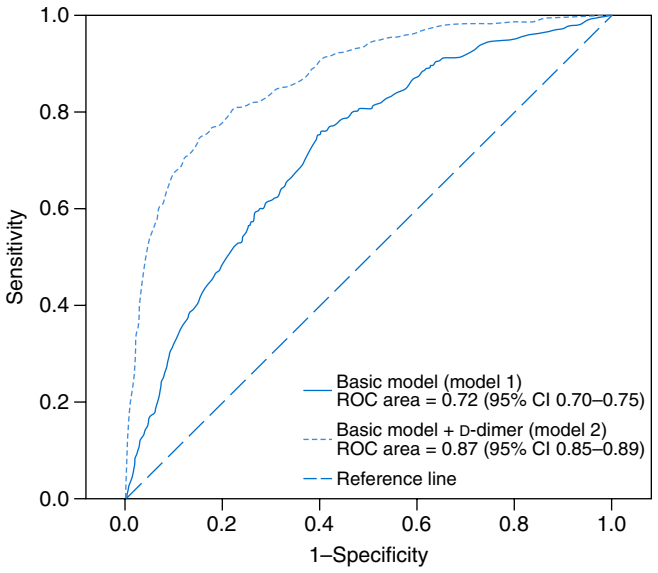


Fig. 2.29 Receiver operating characteristic *(ROC)* curves for the combination of history and physical examination tests before and after addition of the D-dimer assay result. (From Moons, K. G., de Groot, J. A., Linnet, K., Reitsma, J. B., & Bossuyt, P. M. [2012]. Quantifying the added value of a diagnostic test or marker. *Clinical Chemistry, 58,* 1408–1417.)

the presence versus absence of the target disease (here DVT). Fig. 2.29 shows the ROC curves and areas for models 1 and 2. Adding the quantitative D-dimer assay to model 1 mediated an increase in the ROC area from 0.72 to 0.87, a considerable and statistically significant gain ($P < .01$).

Impact of Diagnostic Tests

When thinking about approaches to evaluate the impact of diagnostic tests on medical decision making, patient outcomes, and healthcare at large, it is useful to describe the pathways through which benefits (and risks) of using the test are likely to occur. This so-called *working pathway* provides a framework (Fig. 2.30) to explain how a given test leads to benefits or risks for patients' health or healthcare. Such working pathways include:

1. The anticipated technical or analytical capabilities of the test
2. The unintended and intended results and effects of the test (e.g., benefits of diagnosis and treatment) when applied in the targeted context
3. Those individuals in whom these effects are likely to occur (e.g., in the targeted patients or in the care providers)
4. The anticipated mechanisms through which these potential effects will occur
5. Existing care in the targeted context and individuals
6. The expected time frame in which potential risks and benefits might occur

A clear description of the working pathway of a new test can determine the current benefits (and risks) of prevailing care in the intended medical context. It also helps determine what added value or benefits the new test must provide to improve existing care and what evidence is necessary to quantify whether these (added) benefits are indeed achieved at what risks or costs.

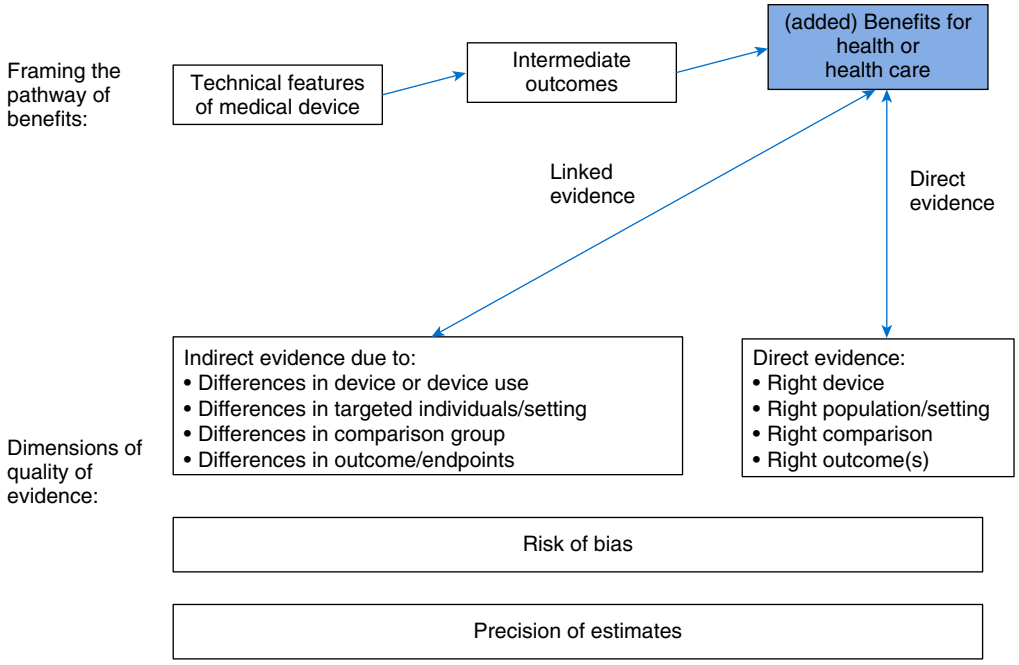


Fig. 2.30 Relationship between the pathway through which devices may lead to benefits or added benefits for health or health care and the three dimensions of quality for evidence (indirectness of evidence, risk of bias, precision of estimates). (From KNAW. [2014]. *Evaluation of new technology in health care. In need of guidance for relevant evidence*. Amsterdam: KNAW.)

POINTS TO REMEMBER

- The diagnostic accuracy of a test indicates the frequency and type of errors that a test will produce when differentiating between patients with and without the target disease.
- The cohort design based on patients suspected of the diseases targeted by the index test is generally preferable for evaluating diagnostic accuracy.
- It is not meaningful to regard estimates of diagnostic performance as properties of the test itself but rather to interpret them as depending on the setting in which the index test was applied and dependent on other tests that are commonly used in that setting.
- Focus has been on approaches and measures for quantification of the diagnostic accuracy of combinations of index tests and of the added value of a new diagnostic test beyond existing diagnostic tests.

REVIEW QUESTIONS

1. Which statement applies to a sample of values drawn from a Gaussian distribution?
   a. The central location is best described by the median.
   b. The dispersion is best described by the interquartile range.
   c. The distribution of the values is likely to be asymmetric.
   d. The *t* distribution is useful for estimation of the 95% CI for the mean value.
2. The analytical specificity of an assay is:
   a. the ability of an assay procedure to determine the concentration of a target analyte in the presence of interfering substances in the sample matrix.
   b. the detection limit of a method.
   c. the ability of an analytical method to assess small variations in the concentration of analyte.
   d. the analyte concentration range over which measurements are within the declared tolerances for imprecision and bias of the method.
3. Two analytical methods are to be compared by analysis in parallel of a suitable number of patient samples. Which of the following is correct?
   a. Ordinary least-squares regression analysis is the most appropriate data analysis approach.
   b. It is generally recommended that the manufacturer use 40 samples for comparison and the user laboratory 100 samples.
   c. A calibration difference is most typically disclosed by an intercept estimate significantly different from zero obtained by regression analysis.
   d. In case of constant CV%s, the Bland-Altman difference plot shows an increasing scatter of the measured differences at increasing measurement values.

4. In a regression analysis comparing results of two methods, the *y*-intercept is calculated to be 2.0 and the slope is 3. This indicates a(n):
   a. calibration error.
   b. uncertainty.
   c. systematic difference.
   d. interference in one method.
5. Which one of the following, when stated as an interval around a reported laboratory result, will specify the location of the true value with a given probability?
   a. Traceability
   b. Coefficient of variation
   c. Trueness
   d. Uncertainty
6. The traceability chain extends downwards from the reference measurement procedure to the routine analytical method. Which of the following is correct?
   a. A reference measurement procedure is sensitive to matrix effects.
   b. The standard uncertainty indicates a 95% uncertainty interval.
   c. Harmonization of laboratory measurements do not presuppose traceability to a reference measurement procedure.
   d. The reference measurement procedure is always more precise than the routine analytical method.
7. The diagnostic accuracy of a test is assessed on a number of subjects suspected of having a given target disease. Which of the following is correct?
   a. The diagnostic accuracy is characteristic for the test and is not influenced by the actual setting in which it is evaluated.
   b. The ROC area provides a measure of the diagnostic accuracy, which is not dependent on a selected cut-off value.
   c. When the cut-off value of a quantitative test is increased, the specificity declines and the sensitivity increases.
   d. In order to rule out the presence of disease, it is important that the specificity is high.
8. The probability of the presence of a specific disease divided by the probability of its absence is the:
   a. likelihood ratio.
   b. odds ratio.
   c. prevalence.
   d. predictive value.
9. When a receiver operating characteristic curve is plotted, the *x*-axis represents the:
   a. false-positive rate.
   b. true-positive rate.
   c. false-negative rate.
   d. true-negative rate.
10. A new test is added to an existing set of diagnostic procedures. Which of the following is correct?
   a. The diagnostic accuracy of the new test is the most important point to consider.
   b. A multivariate data treatment based on logistic regression analysis presupposes quantitative test results.
   c. It is unlikely that results from several tests are correlated.
   d. The difference between the ROC curve area after addition of the new test and the area of the ROC curve of the original diagnostic procedure expresses the added value of the new test.

SUGGESTED READINGS

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, *i*, 307–310.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., et al. (2015). An updated list of essential items for reporting diagnostic accuracy studies. *Clinical Chemistry*, *61*, 1446–1452.

Bossuyt, P. M. M., Reitsma, J. B., Linnet, K., et al. (2012). Beyond diagnostic accuracy: the clinical utility of diagnostic tests. *Clinical Chemistry*, *58*, 1636–1643.

Dybkær, R. (1997). Vocabulary for use in measurement procedures and description of reference materials in laboratory medicine. *European Journal of Clinical Chemistry and Clinical Biochemistry*, *35*, 141–173.

Hendriksen, J. M. T., Geersing, G. J., van Voorthuizen, S. C., et al. (2015). The cost–effectiveness of point-of-care D-dimer tests compared with a laboratory test to rule out deep venous thrombosis in primary care. *Expert Review of Molecular Diagnostics*, *15*, 125–136.

Horvath, A. R., Lord, S. J., StJohn, A., et al. (2014). From biomarkers to medical tests: the changing landscape of test evaluation. *Clinical Chemistry Acta*, *2014*(427), 49–57.

Krouwer, J. S. (2002). Setting performance goals and evaluating total analytical error for diagnostic assays. *Clinical Chemistry*, *48*, 919–927.

Linnet, K., Bossuyt, P. M., Moons, K. G., et al. (2012). Quantifying the accuracy of a diagnostic test or marker. *Clinical Chemistry*, *58*, 1292–1301.

Linnet, K. (1990). Estimation of the linear relationship between the measurements of two methods with proportional errors. *Statistics in Medicine*, *9*, 1463–1473.

Linnet, K. (1993). Evaluation of regression procedures for methods comparison studies. *Clinical Chemistry*, *39*, 424–432.

Linnet, K. (1999). Limitations of the paired t-test for evaluation of method comparison data. *Clinical Chemistry*, *45*, 314–315.

Linnet, K. (2000). Nonparametric estimation of reference intervals by simple and bootstrap-based procedures. *Clinical Chemistry*, *46*, 867–869.

Moons, K. G., de Groot, J. A., Linnet, K., et al. (2012). Quantifying the added value of a diagnostic test or marker. *Clinical Chemistry*, *58*, 1408–1417.

Moons, K. G. (2010). Criteria for scientific evaluation of novel markers: a perspective. *Clinical Chemistry*, *56*, 537–541.

Obuchowski, N. A., Lieber, M. L., & Wians, F. H., Jr. (2004). ROC curves in clinical chemistry: uses, misuses, and possible solutions. *Clinical Chemistry*, *50*, 1118–1125.

Oudega, R., Moons, K. G., & Hoes, A. W. (2005). Ruling out deep venous thrombosis in primary care. A simple diagnostic algorithm including D-dimer testing. *Thrombosis and Haemostasis*, *94*, 200–205.

Passing, H., & Bablok, W. (1984). Comparison of several regression procedures for method comparison studies and determination of sample sizes. *Journal of Clinical Chemistry and Clinical Biochemistry*, *22*, 431–445.

Pencina, M. J., D'Agostino, R. B., & Vasan, R. S. (2010). Statistical methods for assessment of added usefulness of new biomarkers. *Clinical Chemistry and Laboratory Medicine*, *48*, 1703–1711.

Snedecor, G. W., & Cochran, W. G. (1989). *Statistical methods* (8th ed.). Ames, Iowa: Iowa State University Press, 75. 121, 140–142, 170–4, 177, 237–8, 279.

Vesper, H. W., & Thienpont, L. M. (2009). Traceability in laboratory medicine. *Clinical Chemistry*, *55*, 1067–1075.

3

Preanalytical Variation

Ana-Maria Simundic

OBJECTIVES

1. List major sources of variability in preanalytical phase.
2. Define the following terms:
   - Modifiable and unmodifiable influencing factors
   - Endogenous and exogenous interference factors
   - In vivo and in vitro hemolysis
   - Lipemia
   - Icterus
   - First morning urine
   - Timed urine specimen
   - First void urine
   - Midstream urine
   - Suprapubic urine aspiration
   - Unstimulated whole saliva
   - Stimulated saliva
3. State how to deal with modifiable and nonmodifiable influencing factors.
4. Summarize the requirements for a proper definition of the fasting state.
5. Describe mechanisms through which hemolysis, icterus, and lipemia affect various assays and parameters.
6. Describe ways of removal of lipids from the sample.
7. Identify major preanalytical issues related to some specific laboratory tests or sample types.
8. Define optimal sample type for blood gas analysis.
9. Describe proper procedure for saliva collection.
10. Explain the difference between stimulated and unstimulated saliva collection.
11. List advantages and challenges of using saliva as a diagnostic specimen.
12. Describe how cold agglutinins, cryoglobulins, and ethylenediaminetetraacetic acid "(EDTA)-dependent antibodies" interfere with hematologic analytes:

KEY WORDS AND DEFINITIONS

Cold agglutinins Antibodies specific for erythrocyte surface carbohydrate antigens, which bind to the erythrocyte surface at 0°C to 4°C.

Cryoglobulins Immunoglobulins with temperature-dependent solubility that precipitate at temperatures below 37°C.

Endogenous interferences Interferences that originate from the substances present in the patient sample.

Exogenous interferences The effect of various substances added to the patient sample (anticoagulants, gel particles, infusions, etc.).

First morning urine A urine specimen that is collected immediately after arising from bed after an overnight sleep (also called *overnight* or *early morning* urine).

First void urine The first portion of the urine (usually first 15 to 50 mL) passed at any time of the day.

Hemolysis A process of membrane disruption of erythrocytes and other blood cells, accompanied by the subsequent release of cell components into the plasma and red coloration of the serum or plasma after centrifugation.

Icterus A clinical condition that is associated with a change of the color of the serum due to the increase of bilirubin concentration above 5.9 mg/dL or 100 μmol/L.

In vitro hemolysis Hemolysis that occurs outside of the patient at various steps of the preanalytical phase (during blood sampling, sample handling, delivery to the laboratory, and sample storage).

In vivo hemolysis Hemolysis that occurs within the body, before the blood has been drawn and is a result of a pathologic condition.

Influencing factors In vivo and in vitro effects on laboratory results of biological origin that modify the concentration of the affected analyte in a method independent way.

Interference factors Compounds and mechanisms in a biological specimen that alter the result of an analyte thus leading to falsely increased or decreased laboratory test results for that analyte.

Laboratory error Any error that occurs throughout the total testing process (i.e., preanalytical phase, analytical phase, and postanalytical phase).

Lipemia Turbidity of the sample visible to the naked eye, which is caused by light scattering due to the presence of large lipoprotein particles.

Midstream urine A urine specimen collected during the middle of a urine flow after the urinary opening has been carefully cleaned (also called *clean catch urine specimen*).

Modifiable influencing factors Influencing factors that can be controlled by patient action.

Preanalytical phase Part of the total testing process that encompasses test requesting, patient preparation, sampling, sample transport, and delivery to the laboratory, and sample handling before analysis.

Stimulated saliva A saliva sample obtained by the stimulation of salivation with citric acid or by oral movements (yawning, chewing paraffin or gum) and collected through adsorbing saliva by cotton rolls, cotton swabs, or filter paper.

Suprapubic aspiration A specific way of urine collection performed by aspirating urine from the distended bladder through the abdominal wall.

Timed urine specimen A urine specimen that is collected at a specified time or in relation to some activity (e.g., before a meal, after a meal, before therapy, exercise, etc.) during a 24-hour period.

Unmodifiable influencing factors Factors that lead to unavoidable changes in analyte concentration and cannot be controlled by patient action.

Unstimulated whole saliva A saliva sample collected either passively or by spitting into a collector vial.

PREANALYTICAL PHASE

The annual incidence of premature patient deaths associated with some form of preventable medical error in the United States has been estimated to be 98,000 per year. More recent data indicate that the actual mortality caused by preventable medical errors might even be fourfold higher. **Laboratory errors** significantly contribute to the overall error frequency in healthcare. They can lead to diagnostic errors (i.e., missed diagnosis, misdiagnosis, and delayed diagnosis) and consequent patient harm, and increased healthcare expenditure. The **preanalytical phase** is recognized as the most vulnerable part of the total testing process and it accounts for two-thirds of all laboratory errors and may be attributed to various causes. Many preanalytical steps are performed outside the laboratory and are not under the direct supervision of laboratory staff. Furthermore, many individuals are involved in various preanalytical steps and those individuals have different levels of education and professional background. Finally, safe practice standards for many activities and procedures are: (1) either not available, or (2) are available but are not evidence based, or (3) there is low level of compliance with those standards.

The ISO 15189 accreditation standard clearly defines that medical laboratories are responsible for the management and quality of the preexamination phase. It is the role of the laboratorian to ensure that the right sample is taken from the right patient, at the right time, and that correct test results are provided to the requesting physician in a timely manner. If the quality of the specimen is compromised to a degree where the expected effect is larger than the allowable error, causing clinically significant bias, the sample should be rejected for analysis. The guiding principle should be: "No result is always better than a bad result."

Influencing and Interference Factors

Influencing Factors

Influencing factors are effects on laboratory results of biological origin which can occur in vivo (most common) and in vitro (during sample transport and storage). Biological factors modify the concentration of the analyte in a method independent way.

These factors are either present in the healthy individual, like circadian rhythms, or appear as an effect of a disease and its treatment. Examples of **modifiable influencing factors** are diet, time of the day, or time of the year (season), whereas **unmodifiable influencing factors** are gender, ethnicity, genetic background, etc.

Modifiable Influencing Factors

Changes occurring due to the time of sampling may affect the concentrations of various analytes. These changes can be linear (in a chronological order) and cyclic (of a repetitive nature, such as seasonal changes, or changes due to the menstrual cycle).

Several analytes tend to fluctuate in terms of their plasma concentration over the course of a day (i.e., circadian rhythm). For example, the concentration of cortisol is highest in the morning and lowest at night. To avoid the potential effect of analyte changes during the day, reference intervals are defined for blood sampling between 7 and 9 a.m.

Besides the circadian rhythm, some analytes can exhibit significant changes due to the biological changes that occur in hormone patterns during menstruation. For example, aldosterone concentration in plasma is twice as high before ovulation as in the follicular phase.

Diet also may substantially affect the composition of plasma. Differences in serum composition may occur respective to the source of nutrients, the number of meals, and the proportion of nutrients in a diet. Effects from diet can be either long term or acute. Long-term effects of diet may occur over a couple of days, or longer. For example, a diet rich in fat leads to increased serum triglyceride concentration, whereas diet rich in carbohydrates decreases serum protein and lipid concentrations. Vegetarians tend to have lower concentrations of plasma cholesterol, triglycerides, and creatinine with reduced urinary excretion of creatinine, and a higher urinary pH, as a result of reduced intake of precursors of acid metabolites. Most of the changes described are corrected following the restoration of good nutrition.

Acute effects of diet and other influencing factors occur immediately or shortly after ingestion of food or fluid intake. For example, the decrease of plasma glucose and increase of lactate are the acute effects of ethanol consumption, which

occur as early as 2 to 4 hours after the ingestion. Tobacco smoking leads to a number of acute changes in the concentration of fatty acids, epinephrine, free glycerol, aldosterone, and cortisol. These changes occur within 1 hour of smoking one to five cigarettes. Moreover, within only 10 minutes of smoking a single cigarette there is an increase of glucose concentration by up to 10 mg/dL (0.56 mmol/L). This increase may persist for 1 hour.

Physical activity of varying duration and intensity may lead to substantial changes in the plasma composition, and the extent of this change depends on several factors, such as training status, fluid intake, electrolytes, and carbohydrates and even the ambient temperature, etc. For example, even a mild physical effort, like clenching a fist during venous blood sampling, can increase the concentration of potassium and should therefore be avoided. Intensive exercise is associated with transient elevations of cardiac biomarkers, markers of muscle damage, enhanced platelet aggregation, leukocytosis, elevation of tissue-plasminogen activator levels, and activation of the fibrinolytic system.

Some modifiable biological influence factors, such as diet, can be controlled by patient action, whereas some others (e.g., age) are noncontrollable. Proper standardization of controllable preanalytical variables will lead to the significant reduction of preanalytical variability.

To ensure standardization of patient preparation for blood sampling, the following general requirements should be applied to all blood tests in the fasting state:

1. Blood should be drawn preferably in the morning from 7 to 9 a.m.
2. Fasting should last for 12 hours, during which water consumption is permitted.
3. Alcohol should be avoided for 24 hours before blood sampling.
4. In the morning before blood sampling, patients should refrain from cigarette smoking and caffeine-containing drinks (tea, coffee, etc.).

Unmodifiable Influencing Factors

Various biological factors, known as **unmodifiable influencing factors**, lead to changes in analyte concentration, which are unavoidable (Table 3.1). These factors should be considered when interpreting laboratory results because their influence cannot be prevented by preanalytical standardization. Possible ways to address these factors are to partition reference intervals for age, race, gender, and month of pregnancy, and so on (see Chapter 5).

Interference Factors

Interference factors have the ability to alter the result of any analyte after the specimen has been collected, thus leading to falsely increased or decreased laboratory test results for that analyte. Their effect is assay dependent.

Interferences can be endogenous and exogenous. **Endogenous interferences** are biological constituents of the sample. They originate from the substances present in the patient sample whereas exogenous interferences are either present in the sample (drugs, herbal components, etc.) or are added to the sample during or after sampling. **Exogenous interferences** relate to the effect of various substances added to the patient sample, such as separator gels, anticoagulants, surfactants, etc., all of which may cause significant interference.

Possible endogenous interference factors include:

1. Glucose
2. Paraproteins
3. Free hemoglobin (i.e., hemolysis)
4. Lipids (i.e., lipemia)
5. Bilirubin (i.e., icterus)

Possible exogenous interference factors include:

1. Drugs
2. Herbal components
3. Anticoagulants
4. Tube additives
5. Tube gels
6. Intravenous infusions

TABLE 3.1 Unavoidable Influences on Laboratory Results

Influence	Examples of Analyte Concentrations Changed	Remarks
Age	Alkaline phosphatase, LDL-cholesterol, hormones, creatinine	Provide age-dependent reference intervals
Race	Creatine kinase higher in black than in white males Granulocytes higher in white than in black Creatinine higher in black than in white	Provide race-specific reference intervals
Gender	Alanine aminotransferase, γ-GT, creatinine	Provide gender-specific reference intervals
Pregnancy	Triglycerides ↑, homocysteine ↓ during pregnancy	Document months of pregnancy with laboratory results
Altitude	CRP, hemoglobin ↑, transferrin↓	Consider weeks of adaptation, when coming from or going to high altitude

CRP, C-reactive protein.

POINTS TO REMEMBER

Influencing and Interference Factors

- Influencing factors lead to changes in the quantity of the analyte in a method-independent way
- Influencing factors may be changeable (e.g., diet, time of the day) or unchangeable (e.g., gender, ethnicity, genetic background)
- The effect of influencing factors may be reduced through the standardization of preanalytical conditions
- Interferences are mechanisms and factors that lead to falsely increased or decreased results of laboratory tests
- Interference factors and their mechanisms differ with respect to the intended analyte and analytical method, and may be reduced or eliminated by selecting a more specific method

ENDOGENOUS INTERFERENCE FACTORS

Hemolysis

Definition and Background

Hemolysis is a process of membrane disruption of erythrocytes and other blood cells, accompanied by the subsequent release of cell components into the plasma and red coloration of the serum (or plasma) to various degrees, after centrifugation. Though hemoglobin is the most abundant protein in red blood cells, hemolysis is not necessarily always associated with the release of hemoglobin into the surrounding extracellular fluid. For example, if a blood sample is stored at a low temperature, low molecular intracellular components (e.g., electrolytes) diffuse from the cells, but hemoglobin will not. Furthermore, efflux of cell components due to cell lysis affects all blood cells (i.e., platelets and white blood cells) and not only erythrocytes. Therefore it is important to remember that red coloration of the serum or plasma can never accurately predict the concentration of analytes released from blood cell components.

Hemolysis is the most common preanalytical error and cause of sample rejection. Its frequency is up to 30% and accounts for almost 60% of unsuitable specimens. The frequency of hemolysis largely depends on the collection facility, the phlebotomy personnel, and the characteristics of the patient population. The highest frequency of hemolysis has been observed in samples from emergency departments, pediatric departments, and intensive care units, whereas hemolysis has proven to be the least frequent in outpatient phlebotomy centers where blood sampling is done by specialized laboratory staff. These differences are due to the level of knowledge and skills of the staff performing blood collection procedures.

There are two major sources of hemolysis: (1) in vivo and (2) in vitro hemolysis. **In vivo hemolysis** occurs within the body, before the blood has been drawn and is a result of a pathologic condition. Hemolysis in vivo may occur as a result of numerous biochemical (enzyme deficiencies, erythrocyte membrane defects, hemoglobinopathies), physical (prolonged marching, drumming, prosthetic heart valves), chemical (ethanol, drug overdose, toxins, snake venom), immunologic (autoantibodies) mechanisms, and infections (babesiosis, malaria). The most common causes of in vivo hemolysis are a reaction to incompatible transfusion and autoimmune hemolytic anemia.

In vivo hemolysis is not very common and accounts for only 3% of all hemolyzed samples. Laboratories should have a procedure in place for distinguishing in vivo and in vitro hemolysis. In vivo hemolysis should always be suspected when patient blood is hemolyzed over a longer period of time, after repeated blood sampling, even after special care has been taken to avoid hemolysis. Decreased concentration of haptoglobin in serum and free hemoglobin in urine are the most pronounced and specific laboratory signs of in vivo hemolysis. Haptoglobin is a protein that binds free hemoglobin in the circulation to prevent oxidative damage induced by hemoglobin. Once released from the erythrocyte into the plasma, hemoglobin forms complexes with haptoglobin, and those complexes are removed from the circulation by macrophages.

TABLE 3.2 Ratio Between Intracellular and Extracellular Concentration of Various Parameters in Red Cells

Analyte	Intracellular Concentration (Compared to Extracellular)
LDH (lactate dehydrogenase)	↑ 160×
Inorganic phosphate	↑ 100×
Potassium	↑ 40×
AST (aspartate aminotransferase)	↑ 40×
Folic acid	↑ 30×
ALT (alanine aminotransferase)	↑ 7×
Magnesium	↑ 3×

Note: The most pronounced effect of hemolysis is seen for LDH. LDH activity may be increased by >20% in mildly hemolyzed samples and as much as >350% in grossly hemolyzed samples.

In more pronounced cases of in vivo hemolysis, haptoglobin in serum can be undetectable (i.e., below the limit of detection), whereas its concentration in cases of in vitro hemolysis remains unchanged. When in vivo hemolysis is confirmed, the laboratory should not reject hemolyzed samples for analysis, since parameters in hemolyzed samples reflect the actual patient condition and are relevant for adequate patient care (diagnosis, therapy management, monitoring).

In vitro hemolysis occurs outside of the patient at various steps of the preanalytical phase (e.g., blood sampling, sample handling, delivery to the laboratory, and sample storage).

Mechanisms of Hemolysis Interference

Hemolysis is an endogenous interference that causes clinically relevant bias of patient results through several distinct mechanisms:

Spectrophotometric interference. Spectrophotometric interference of hemolysis occurs due to the ability of hemoglobin to absorb light at the following wavelengths: 415, 540, and 570 nm. This characteristic of hemoglobin causes optical interference, which can lead to either falsely increased or decreased concentrations of the measured parameters. The direction and degree of the interference largely depends on the analyte and the method.

Release of cell components into the sample. Some components are present in blood cells in concentrations that are several times higher than those in the extracellular space (i.e., plasma or serum). Table 3.2 shows some of the most pronounced differences between intracellular and extracellular concentrations in red cells.

From this it follows that there is a dramatic increase of the concentration of the listed analytes measured in hemolyzed plasma (or serum) due to the efflux of those substances from erythrocytes.

Since intracellular components may also escape from platelets during clotting, there is a marked difference in the potassium concentration between serum and plasma. The mean estimated difference in the concentration of potassium

in serum and plasma is 0.36 ± 0.18 mmol/L and this difference is positively associated with the platelet count. Plasma is therefore the recommended sample type for the accurate measurement of potassium.

Sample dilution. Some analytes, such as albumin, bilirubin, glucose, and sodium, are present in much higher concentrations in plasma than in blood cells. For those parameters hemolysis will cause a dilution effect and their concentrations will be lower in hemolyzed samples. The effect of sample dilution causes clinically significant bias only at a higher degree of hemolysis. For example, glucose is negatively affected by severe hemolysis (−8.3%) only when the concentration of free Hb is greater than 3 g/L.

Chemical interference. Various blood cell components may affect the analyte measurement procedure by directly or indirectly competing for molecules in the reagents, inhibiting indicator reactions or modifying the analyte by complexation, proteolysis, or precipitation. One such effect is caused by the enzyme, adenylate kinase, present in erythrocytes, as well as in platelets. When released from the cells during hemolysis, adenylate kinase may compete for ADP with creatine kinase in a creatine kinase assay if inhibitors are not supplied in the reaction mixture. Furthermore, hemoglobin released from erythrocytes during hemolysis may interfere with various assays through its pseudo-peroxidase activity. Pseudo-peroxidase activity of free hemoglobin released from erythrocytes interferes in the assay for measurement of bilirubin concentration through the inhibition of the formation of diazonium salt.

Hemolysis may cause a clinically significant interference on a wide range of analytes in immunochemistry assays. This interference is caused by modifying the reaction analytes (antigens and antibodies) by the proteolytic action of cathepsin E, the major proteolytic enzyme in mature erythrocytes. Proteolytic enzymes released from erythrocytes may mask or potentially enhance epitope recognition in various immunoassays. Interference caused by proteolytic activity may cause measurement bias of various degrees and directions, depending on the assay.

Lipemia

Lipemia is defined as turbidity of the sample visible to the naked eye. This turbidity is caused by light scattering due to the presence of large lipoprotein particles. The increase in concentration of lipoproteins in blood most commonly occurs due to postprandial triglyceride increase, parenteral lipid infusions, or some lipid disorders. Not all lipoproteins have equal contribution to the sample turbidity. The effect of lipoprotein particles on sample turbidity depends on the size of the particles. Chylomicrons and very-low-density lipoproteins (VLDLs), the largest lipoprotein particles in the circulation, have the greatest contribution to sample turbidity. To avoid postprandial lipemia, patients are requested to fast for 12 hours before blood sampling.

Mechanisms of Interference Caused by Lipemia

Lipemia is an important endogenous interference that may cause clinically relevant bias of patient results through the following mechanisms:

Spectrophotometric interference. Lipemia causes interference by light absorbance and light scattering. A lipemic sample absorbs light, decreasing the intensity of the light passing through the sample. This absorption occurs in the 300 to 700 nm wavelengths. Sample absorbance rises with the decreasing wavelengths and is maximal in the ultraviolet range. That is why many enzymatic methods in which the end product is measured at 340 nm are strongly affected by lipemia.

Lipemic samples also cause light scattering. Light scattering occurs in all directions and its intensity depends on the number and size of lipoprotein particles and the wavelength of measurement. For this reason, light scattering of lipoprotein particles causes significant interference with turbidimetry and nephelometry. In methods where the transmittance of light is inversely proportional with the concentration of the analyte, in the absence of the sample blank, sample turbidity causes positive bias. However, in some competitive assays where the transmittance of light is directly proportional to the concentration of the analyte, sample turbidity will cause negative bias.

Interference caused by the volume depletion effect. Plasma in healthy individuals in the fasting state consists of only a minor portion of lipids (<10% of the total plasma volume). The rest of the plasma is water. The increase in the concentration of lipoprotein particles leads to the increase of the plasma volume occupied by lipids. Particles that are not lipid soluble are displaced by the lipids to the water part of the plasma. Through this mechanism lipemia may lead to the false decrease of the concentration of the measured analyte, in all methods in which the concentration of respective analyte is measured in the total plasma volume (e.g., pseudo-hyponatremia if sodium is measured by flame photometry and indirect measurement using ion-selective electrodes, but not in direct potentiometry).

Interference caused by partitioning of the sample. Upon centrifugation of a lipemic sample, lipoproteins are not homogeneously distributed in the serum or plasma due to the lipid gradient. Water-soluble analytes are more concentrated in the lower layer of the plasma or serum, whereas lipids and lipid soluble analytes, such as drugs and some lipid-soluble hormones, are more concentrated in the top lipid-rich layer. This is especially important in automated chemistry analyzers with a fixed path length of the sample probe. Test results may differ for those analytes that are not evenly distributed between the lipid and water portions of the sample, depending on the part of the specimen from which the probe takes the sample for analysis.

Interference caused by physico-chemical mechanisms. Excess lipoproteins in blood may interfere in electrophoretic and chromatographic methods by causing abnormal peaks. Elevated levels of triglycerides and lipoprotein particles may disturb the electrophoretic pattern as well as falsely increase the relative percentage of the prealbumin, albumin, and α1- and α2-globulin regions. Moreover, lipemia may affect some immunochemistry assays by masking the binding sites on antigens and antibodies, and thus physically interfere with antigen–antibody binding.

Removal of Lipids from the Sample

Unlike hemolysis, the interference caused by lipemia can be fully eliminated or at least reduced, by removing the excess lipids from the sample. Methods for lipid removal include ultracentrifugation, high-speed centrifugation, and some lipid clearing agents.

Ultracentrifugation is the recommended method for the removal of the excess of lipids in the sample. Ultracentrifuges use the centrifugation force of almost 200,000 g and are very effective in clearing lipemic sera by separating lipids, especially chylomicrons (top layer), from the aqueous part (lower layer) of the sample. After centrifugation, the infranatant (lower part of the sample) can be analyzed. It should be kept in mind that by removing the upper lipid layer one also removes lipid soluble analytes such as drugs and hormones. Results reported from ultracentrifuged samples or samples from which lipids have been removed in any other way, should be appropriately annotated to ensure clinicians are aware that the sample has been manipulated to obtain the reported results.

In laboratories where ultracentrifuge is not available, high-speed centrifugation using the microcentrifuge with a maximum centrifugation speed of up to 20,000 g for 15 minutes at room temperature may serve as an acceptable alternative.

Lipid clearing agents are widely used in many laboratories due to their low cost, convenience, and ease of use. Those agents (cyclodextrin, polyethylene glycol, dextran sulfate, hexane, and others) may vary in their ability to extract lipids from a lipemic sample, and may lead to reduction of a significant amount of protein from the sample. It is therefore very important for laboratories to verify the performance of such reagents before their routine use.

Icterus

The normal concentration of bilirubin in human plasma (or serum) is up to 20 μmol/L. A change in the color of the serum (or plasma) becomes detectable when bilirubin concentration exceeds 34 μmol/L. Bilirubin concentrations above 100 μmol/L are clinically defined as **icterus**. Bilirubin interferes with numerous chemistry tests, such as enzymes (alanine aminotransferase, alkaline phosphatase, creatine kinase, lipase), electrolytes, metabolites (urea, creatinine, glucose), lipids (cholesterol, triglycerides), proteins (albumin, total proteins, immunoglobulin [Ig]-G), hormones (estradiol, beta-human chorionic gonadotropin [β-hCG], free T3), and even some drugs (gentamicin, phenobarbital, theophylline, tobramycin). As is the case for hemolysis and lipemia, interference caused by bilirubin differs between instruments and assays.

Unfortunately, there is not much that can be done by the laboratory to remove or minimize the effect of icteric interference. Possible options are diluting the sample (possible only for analytes present at high enough concentrations in the blood) and testing the requested analytes with a different method or on a different instrument for which icterus does not cause clinically significant interference.

Mechanisms of Interference Caused by Icterus

Icterus interferes through two mechanisms: spectrophotometric interference and by interfering with chemical reaction. It is important to recognize that both mechanisms may occur simultaneously in one sample.

Bilirubin causes spectrophotometric interference due to its ability to absorb light in the wide range of wavelengths between 400 and 540 nm.

Chemical interference of bilirubin is exerted through its interaction with various components in the blood. For example, bilirubin produces negative bias on assays that involve H_2O_2 as an intermediate reaction (e.g., cholesterol, glucose, uric acid, triglycerides).

Detection of Hemolysis, Lipemia, and Icterus

Hemolysis becomes visible at the concentration of 0.3 to 0.5 g/L of free hemoglobin and the intensity of the red color of the serum or plasma further increases with the increase in concentration of free serum hemoglobin (Fig. 3.1A). Lipemia causes sample turbidity, which approximately corresponds to the concentration of serum triglycerides (see Fig. 3.1B). Increased concentrations of serum bilirubin lead to yellow-to-orange coloration of the serum, and the change of the color correlates with the increasing concentration of the bilirubin in serum (see Fig. 3.1C).

Detection of the degree of hemolysis, icterus, and lipemia by visual inspection is highly unreliable and should be replaced with automated detection. Most high-throughput chemistry analyzers currently available on the market have the ability to detect serum indices by the use of semi-quantitative, spectrophotometric measurement, and grading the interfering substances into categories. The serum index is automatically reported for every sample and can be used to determine the degree of interference and its effect on the requested parameters. Such systems are highly reproducible and provide an objective and standardized way to screen for common interferences and manage specimen rejection via built-in rules in analyzer software.

Although the information about serum indices and their cut-offs for tested analytes are provided by the manufacturers of in vitro diagnostic systems and reagents, their claims are often not complete, accurate, and reproducible. It is therefore good practice for a laboratory to perform its own verification of serum indices. Alternatively, laboratories may rely on the evidence from the literature, if available and of adequate quality.

As for all other laboratory methods, analytical quality of serum indices should be continuously monitored by using appropriate internal quality control (IQC) and through participation in an external quality assessment program (EQA). IQC material for serum indices has been recently made available on the market by several manufacturers.

EXOGENOUS INTERFERENCE FACTORS

The effects of exogenous interfering substances are difficult to identify. They are introduced into the sample in different ways; for example, through therapy (drugs, natural preparations for self-medication, or supportive therapy), diagnostic

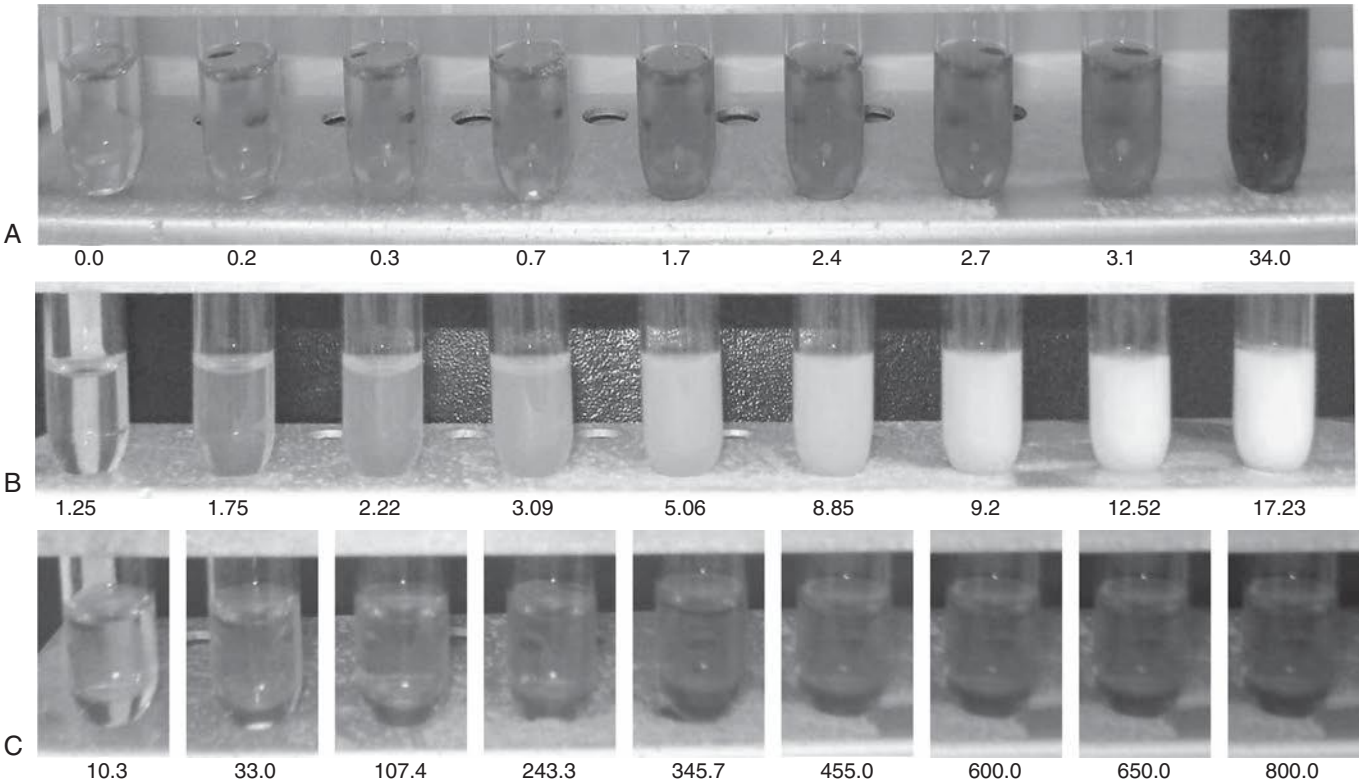


Fig. 3.1 (A) Hemolysis—the intensity of the red color of the serum and corresponding concentrations of free serum hemoglobin. (B) Lipemia—the degree of turbidity and corresponding concentrations of triglycerides. (C) Icterus—the intensity of the yellow color of the serum and corresponding concentrations of bilirubin. *Concentration of Intralipid; **concentration of triglycerides. (Standard color scales; Courtesy Clinical Institute of Chemistry, University Hospital Center "Sestre Milosrdnice," Zagreb, Croatia.)

procedures (contrast media), by accidental exposures and poisonings (with drugs, herbs, household chemicals, etc.), and through contamination of the sample during sample handling (from rubber tube stoppers, lubricants, anticoagulants, or surfactants) or sample analysis (antibiotics used for reagent and buffer stability).

POINTS TO REMEMBER

Hemolysis, Lipemia, Icterus

- Visual assessment of the degree of hemolysis, lipemia, and icterus is not reliable and leads to errors.
- Hemolysis is the most common preanalytical error and most common cause of sample rejection.
- Hemolysis may cause clinically relevant bias through spectrophotometric and chemical interference, sample dilution, and the release of cell components into the sample.
- Lipemia causes interference by spectrophotometric interference (light absorbance and light scattering), by the volume depletion effect, by partitioning of the sample, and by physico-chemical mechanisms (e.g., disturbance of the electrophoretic pattern).
- Laboratories should verify the performance of lipid removal reagents before their routine use, since they may not be appropriate for a wide range of analytes due to their low recovery.
- Different forms of bilirubin cause varying degrees of interference with different laboratory methods, and the same forms of bilirubin act differently with the same assays on different instruments.

SPECIFIC CONSIDERATIONS

Besides the general aspects of preanalytical quality (i.e., patient and sample identification, controllable and noncontrollable variables), some specific aspects related to urine, saliva, and blood gas testing are of particular relevance to laboratory medicine. The following sections deal with the preanalytical aspects of sample types other than serum or plasma, and other types of testing that deserve special preanalytical considerations.

Influences and Interferences in Urine Testing

Some of the most important preanalytical issues in the analysis of urine are patient preparation, choice of urine type, type of collection vessel, sample stability during transport and storage, sample homogeneity, and sample contamination. These aspects are covered below.

Types of Urine

Not every urine sample is fit for the purpose of every type of laboratory testing. The possible types of urine sample and their intended use are described in Table 3.3.

When collecting urine specimens, in order to prevent contamination with bacteria that normally reside on the skin, hygiene of the hands and genitalia is essential. Hands should be washed with soap, thoroughly rinsed with water, and dried. Not only proper washing with soap, but also thorough rinsing is extremely important, as baby soaps are known

TABLE 3.3 Urine Specimens and Their Diagnostic Use

Urine Type	Use
First morning urine	Urine sediment, test strip
Second morning urine	Proteins, urinalysis
Timed urine (6–24 h)	Hormones, drugs, electrolytes
First void urine	Chlamydia
Midstream urine	Urinalysis, microbiologic examination
Urine obtained through a catheter or sterile suprapubic aspiration	Exclusion and confirmation of urinary tract infection

to cause significant interference with tetrahydrocannabinol (THC) immunoassays. To prevent sample contamination by skin microorganisms, it is very important not to touch the inner surface of the cap and the container during collection. Urine should be collected while holding the skin folds (labia) apart (females) or retracting the foreskin (uncircumcised men) during voiding.

First morning urine (also called overnight or early morning specimen) is collected immediately after arising from an overnight sleep. Exceptionally, in patients suffering from insomnia or in night-shift workers, other 8-hour periods may also be used for the purpose of first morning urine collection. It is important that the patient's bladder is emptied immediately before going to sleep and that any amount of urine voided during the night is also collected and pooled together with the first voided morning specimen.

Timed urine specimens are collected at a specified time or in relation to some activity (e.g., before a meal, after a meal, before therapy, before exercise, and so on) during a 24-hour period. The exact time of the collection should always be reported with a test report.

First void urine comprises the first portion of the urine (usually first 15 to 50 mL) passed at any time of the day. It is collected after a patient has not urinated for at least 1 to 2 hours. The exact time depends on the sensitivity of the actual test method and is usually designated on the method insert sheet.

Midstream urine (also called clean catch specimen) is a specimen collected during the middle of a urine flow, after the urinary opening has been carefully cleaned. The first amount of urine should be passed into the toilet instead of the container. This will prevent contamination with skin, vaginal or urethral cells, and bacteria. The mid-portion of the urine is collected and once the container is filled, the rest of urine is voided into the toilet until the bladder is empty.

Suprapubic aspiration and catheterization are procedures usually applied for bacteriologic studies to obtain uncontaminated urine. A suprapubic specimen is collected by aspirating urine from the distended bladder through the abdominal wall, while catheterization enters the bladder through the urethra. Both collection methods use the sterile technique.

The ideal container for any urine specimen is a wide-mouthed bottle of appropriate size. Containers should be clean, leak proof, particle-free, and preferably made of a clear, disposable material that is inert with regard to urinary constituents. If urine is to be transported, the container used during transportation should have a secure closure to prevent leakage of the contents during transportation. If the urine is to be analyzed bacteriologically, containers have to be sterile.

For pediatric and newborn patients, urine specimen collection bags with hypoallergenic skin adhesive should be used. First, the pubic and perineal areas should be cleaned with soap and water. Then, the adhesive strip should be pressed all around the vulva or the bag fixed over the penis and the flaps pressed to the perineum.

Transport and Storage of Sample

Since some urine parameters have limited stability in unpreserved urine, temperature and time conditions during transport and storage of urine are very important to ensure adequate sample quality. Urine samples may be stored for up to 1 hour at room temperature and up to 4 hours if refrigerated without significant variation in the results of the physical, chemical, and morphologic analysis of particles. Chemical constituents in urine (for test strip analysis) are stable for 24 hours if urine has been kept at +4°C.

POINTS TO REMEMBER

Urine

- The contamination rate is much lower in urine specimens collected after proper hand and genital hygiene.
- Laboratories should provide instructions to patients about reasons for urine collection, how to prepare for urine sampling (e.g., effects of diet and fluid intake, diuresis, exercise, and other interferents), and how to properly collect urine.
- Stability of different analytes in urine is limited and decreases during prolonged storage and at higher temperatures.
- Preservatives like ethanol, polyethylene glycol, sodium fluoride, mercuric chloride, boric acid, and formaldehyde- and formate-based solutions may enhance the stability of urine particles.
- The addition of urine stabilizers inhibits the bacterial growth, metabolic processes, and degradation of urine analytes and particles.

Saliva

Saliva is produced by *major* salivary glands (parotid, submandibular, and sublingual glands) and by oral secretion of the mucus by hundreds of *minor* salivary glands. The primary function of saliva is to provide oral protection by lubrication, digestion, and an immune response. Saliva is an attractive alternative to blood because it is collected noninvasively. As a diagnostic sample, saliva offers many advantages as well as some challenges (Box 3.1).

Types of Saliva Samples

The easiest way to collect saliva is to collect whole oral fluid (whole saliva). Whole saliva is representative of the oral milieu. However, depending on the intended aim of the

BOX 3.1 Advantages and Challenges of Using Saliva as a Diagnostic Specimen

Advantages	Challenges
• Rapid and easy collection by minimally trained individuals • Sampling can be done by patients, at home, out of hospital • Multiple sample collection is possible • Procedure safe and painless for the patient • Convenient in children, psychiatric patients, and stress research • Availability • Low cost associated with sampling (skilled staff not required) • Convenient method for population screening programs • Low risk of infections associated with sampling	• Low analyte concentration • Some analytes may be affected by circadian cycle • Questionable recovery for some analytes • Risk of contamination during collection • Difficult sampling and low patient compliance (in small children)

sample collection and analyte to be tested, some other sample types are also possible. Various sampling techniques and devices are available. Whereas collection of whole oral fluid is an easy procedure, other sampling methods are more complicated; they require trained staff and are not commonly used.

Whole saliva may be collected as unstimulated (mostly produced by the submandibular glands) and stimulated (originating predominantly from the parotid glands) samples. Stimulation of saliva is obtained by oral movements (yawning, chewing paraffin or gum), or using a cotton roll soaked with citric acid. It has to be noted that stimulated and unstimulated saliva are of different origin (produced by different salivary glands) and composition (concentration of some analytes may vary), and thus are not equally suitable for all assays.

Several commercially available devices may be used for collection of saliva. **Unstimulated whole saliva** can be collected by:

- Passive drooling into the plastic vial
- Spitting into the collector vial

Passive drooling is considered by many to be the gold standard collection method for many analytes, as it enables collection of a representative portion of saliva present in the oral cavity. Moreover, passive drooling is preferred over spitting, since saliva collected by spitting is more likely to be contaminated with bacteria.

Stimulated saliva may be obtained by adsorbing saliva by cotton rolls, cotton swabs, or filter paper. Cotton is not an ideal collection material, not only due to its unpleasant texture, but also because it may induce some variations in salivary immunoassays (some analytes are difficult to elute from cotton). To address this problem, some synthetic materials have been made available, such as inert polymers, polystyrene, rayon, and polyester. It must be emphasized that the composition of saliva collected by the use of various adsorbent devices may differ from the whole saliva, as adsorbent devices mostly collect localized saliva.

Patient Preparation for Saliva Sampling

The laboratory is responsible for providing information to the patient about the purpose of saliva collection, as well as how to prepare for collection, and how to collect saliva. To avoid sample contamination, it is recommended to wear gloves during saliva collection. Below are some important measures aimed to minimize errors and ensure high-quality saliva samples:

- If not otherwise decided by the requesting physician, saliva should be collected in the morning, preferably in the fasting state (the exact time of collection is important, as some analytes may have diurnal variations)
- The patient should wash his/her face with water and soap and rinse it thoroughly to avoid contamination with facial creams and lotions
- The patient should not consume alcohol 12 hours before the collection
- The patient should not brush or floss his or her teeth at least 30 minutes before the collection
- The patient should not have any dental work done 2 days before the collection (dental bleeding may affect the test results)
- The patient should not ingest any food and drinks (except water) within 30 minutes before the collection
- The patient should not chew gum for at least 30 minutes before the collection
- Before collection, the patient should rinse his or her mouth with water to remove some food remnants. To avoid sample dilution with water, the sample should be collected 10 to 15 minutes after rinsing the mouth.

Any sample visibly contaminated with blood or food remnants should be rejected for analysis, and sample re-collection should be requested.

Preanalytical Aspects of Arterial Blood Gas Testing

Blood sampling for assessing the oxygenation status of the patient or acid–base balance should be done when a patient is in a stable, resting state, to ensure that test results reflect the actual condition of the patient. Furthermore, the exact time of the blood collection should always be recorded and reported with a test result. Any deviation from the steady state should be noted as a comment and accompany the test report in order to allow proper interpretation of the results and appropriate patient management.

Relevant patient condition determinants (at the time of blood collection) are:

- Patient status (resting, exercising, crying, anxious)
- Ventilatory setting (spontaneous breathing or assisted mechanical ventilation)
- Mode of oxygen delivery (fraction of inspired oxygen [FiO_2] through nasal cannula or Venturi mask)

- Respiratory rate (hyperventilation, hypoventilation)
- Body temperature

If a patient's condition is changing, sufficient time should be allowed for the patient to stabilize. If there has been any change in the patient ventilatory setting or the mode of oxygen delivery, the patient should also be left in the resting state to stabilize. For patients without pulmonary disease, a period of 3 to 5 minutes is usually enough to stabilize. However, in patients with pulmonary disease, this period is significantly longer. For most patients, 20 to 30 minutes is adequate time to reach a stable state following ventilatory changes.

Arterial blood collected under anaerobic conditions and anticoagulated with lyophilized balanced Li-heparin is the ideal sample type for an accurate evaluation of the gas exchange function of the lungs (pO_2 and pCO_2). If arterial blood is not available (e.g., neonates, small children, patients with burns), and during medical transport and prehospital critical care, a capillary sample is an acceptable alternative. Capillary blood sampling is not recommended in patients with circulatory shock, poorly perfused (cyanotic), infected, inflamed, swollen, or edematous tissues. For more details on blood gas analysis and sample collection and processing, see Chapters 6 and 24.

Hemostasis Testing

Some specific considerations related to hemostasis testing are associated with the type of anticoagulant, sampling technique (fasting state, length of the venous stasis, order of draw, sampling from a catheter), sample handling (centrifugation), transport, and storage prior to analysis.

Samples for hemostasis testing should be anticoagulated with 3.2% sodium citrate, although 3.8% may also be acceptable. It is important that the same concentration of sodium citrate is used within the laboratory, since clotting times may be longer in 3.8% than in 3.2% sodium citrate.

The mixing of samples is extremely important for adequate sample coagulation. Samples must be promptly mixed to avoid in vitro clot formation. Tubes should be mixed by gentle inversion (at 180 degrees) several times. For proper mixing, instructions from the tube manufacturer should be followed. Vigorous mixing and shaking the tubes is discouraged, as it may lead to sample hemolysis, plus the activation of platelets and coagulation factors, resulting in false shortening of clotting times and even the possible false increasing of clotting factor activity. The transport of samples by pneumatic tubes is still under discussion. While some claim that pneumatic tube transport is acceptable for routine hematology and coagulation testing, others argue that it is not appropriate for platelet aggregation studies. It is therefore recommended that each institution verifies the acceptability of their tube transport systems by comparing paired samples for differences.

Samples with visible clots should not be accepted for hemostasis testing. To prevent clot formation during blood sampling, handling, and transport, the following situations should be avoided:

- Blood flow (during blood sampling) too slow
- Collecting the sample into a syringe and then transferring it into a citrated tube
- Tubes underfilled
- Prolonged use of tourniquet (longer than 1 minute)
- Considerable manipulation of the vein by the needle
- Incomplete mixing

Clot formation induces the activation of platelets and clotting factors and the release of granules from the platelets, and it may cause false results in coagulation assays. It is very important to point out that even small clots that are invisible to the human eye may significantly impact coagulation assays.

The required blood to anticoagulant ratio is 9:1 and it is therefore very important that tubes for coagulation assays are filled to the mark noted on the tube. Acceptable deviation is a maximum 10% of the total volume; overfilling or underfilling the designated volume is strictly discouraged as this can introduce bias in test results. Laboratories should have preanalytical procedures in place for the rejection of over- or underfilled tubes. Other anticoagulants (e.g., ethylenediaminetetraacetic acid [EDTA] or heparin) are not acceptable for hemostasis testing, since they lead to erroneous results and cause clinically significant errors.

A standardized order of draw has been recommended, with the coagulation tube as the first tube. Although a "discard" tube before the coagulation tube has long been recommended, more recent evidence suggests that it may not be necessary. If intravenous catheter systems are used for blood sampling, 5 to 6 times the dead space (priming) volume of the catheter should be discarded prior to blood sampling.

Following collection, citrated samples should ideally be transported to the laboratory immediately and at room temperature, but no later than within 1 hour of blood draw. All coagulation assays should be performed in fresh samples. Blood samples for coagulation testing must be kept at room temperature (20 to 25°C) until analysis. Storage at a lower temperature, or on ice, and sample freezing are discouraged as some coagulation factors are activated by cold.

Whole blood coagulation assays should be performed within 4 hours after blood sampling. For platelet function assays, samples should rest (at room temperature) for 30 minutes before analysis. Platelet function assays require the careful preparation of platelet-rich plasma by centrifugation of the blood at 200 to 250 g for 10 minutes without application of a centrifuge brake. To generate platelet-poor plasma, centrifugation is normally performed at 1500 g at room temperature for 10 to 15 minutes. Higher speed with a shorter centrifugation time is not recommended, as this may induce hemolysis and activation of platelets.

Hematology

EDTA is the anticoagulant of choice for hematology testing given its efficacy of preserving cellular morphology. Due to its higher solubility, lower osmotic effect, and best overall performance, the International Council for Standardization in Haematology (ICSH) recommends dipotassium EDTA salt as the anticoagulant of choice for hematology testing.

Blood tubes should be filled to ±10% of the stated draw volume. In underfilled EDTA tubes, cell count and hematocrit might be falsely decreased due to the excess EDTA. In

overfilled tubes, clot formation and platelet clumping is likely to occur, due to difficulty of appropriate mixing.

In some individuals EDTA may cause pseudothrombocytopenia (i.e., platelet clumping or platelet satellitism [platelet adhering to neutrophils]) and subsequently inaccurate platelet results. Since most cell counters are not able to identify this preanalytical problem, platelets are thus counted as white blood cells, resulting in spurious leukocytosis and false thrombocytopenia. EDTA-induced pseudothrombocytopenia has so far been observed in healthy and diseased individuals, and is not related to gender and age. The hypothesized mechanism in pseudothrombocytopenia involves IgM autoantibodies, directed against platelet IIb/IIIa fibrinogen receptors. This is further supported by the fact that platelets from patients with Glanzmann thrombasthenia (in which platelets have either defective or low levels of glycoprotein IIb/IIIa) do not react with autoantibodies from pseudothrombocytopenic patients.

EDTA anticoagulated blood should be stored at room temperature and analyzed within 6 hours of collection. Stability of hematologic parameters may differ depending on the parameter that is being measured, the instrument type, the transport, and the storage conditions. Therefore, on some occasions, a shorter time is necessary to ensure accurate and reliable results, whereas some parameters show excellent stability even over much longer time intervals. Some parameters are very stable (hemoglobin and red blood cells), while some others are not (reticulocytes, mean cell volume, and hematocrit). The stability of hematologic parameters is improved if samples are kept at 4°C. The following are the data provided by the ICSH on the stability of some hematology parameters:

- Hemoglobin concentration and red blood cell count are stable up to 72 hours, if blood is kept at 4°C
- Platelet and reticulocyte count are stable for 24 to 72 hours, if blood is stored at 4°C
- White blood cell count with automated differential count is stable for at least 24 hours, if blood is kept at 4°C, and up to 6 hours if stored at room temperature
- Peripheral blood smear should be made from blood stored no longer than 1 hour at room temperature (18°C to 25°C)

Since stability may vary depending on the instrument, it is the responsibility of the individual laboratory to verify the stability of hematologic parameters on the instruments used in their laboratory.

Antibodies may affect the cell count of erythrocytes, leukocytes, and platelets. The following antibodies are known to interfere with hematologic analytes:

- Cold agglutinins (erythrocyte specific)
- Cryoglobulins
- EDTA-dependent antibodies with thrombocyte and leukocyte specificity

Cold Agglutinins

Cold agglutinins are monoclonal or polyclonal antibodies specific for erythrocyte surface carbohydrate antigens, which bind to the erythrocyte surface at 0°C to 4°C. Binding of these antibodies causes agglutination of erythrocytes, induces complement activation and hemolysis, and impairs peripheral circulation. Some rare cases of cold agglutinins toward platelets have also been described, causing pseudothrombocytopenia independent of EDTA.

Cold agglutinins, if undetected, may cause diagnostic confusion and lead to subsequent extensive diagnostic workup as well as wrong and unnecessary therapy, risking patient safety and increasing healthcare costs. It is therefore very important to recognize cold agglutinins promptly.

Cold agglutinins should be suspected if the following anomalies are observed:

- Red blood cell count too low (in the presence of normal hemoglobin concentration)
- Significantly increased mean corpuscular volume (MCV) values
- Significantly low values of calculated hematocrit in samples with too high mean corpuscular hemoglobin (MCH) and MCH concentration (MCHC) values without any obvious explanation
- Falsely elevated white blood cell and platelet count

White blood cell and platelet counts may be falsely elevated because agglutinates, depending on their size, are either counted in the leukocyte or platelet channel. The blood smear may also show agglutination of erythrocytes. For adequate analysis of samples in which cold agglutinins are suspected, it is essential to warm up the EDTA blood sample at 37°C and analyze it immediately afterwards. This anomaly will appear again if a sample is kept at 4°C and analyzed cold.

Cryoglobulins

Cryoglobulins are immunoglobulins with temperature-dependent solubility that precipitate at temperatures below 37°C. Cryoglobulins are often associated with infections, autoimmune disorders, and malignancies, and cause organ damage through immune-mediated mechanisms and vascular damage due to increased viscosity of the blood. The precipitation of cryoglobulins depends on the immunoglobulin class to which they belong. Also, precipitation is absent at pH less than 5.0 or more than 8.0.

In samples kept at room temperature, cryoglobulins tend to form globular or cylindrical precipitates, which are then counted by automated hematologic analyzers as cells, thus affecting hematologic laboratory tests and leading to false leukocytosis (pseudoleukocytosis), or false thrombocytosis (pseudothrombocytosis). The degree of pseudoleukocytosis and pseudothrombocytosis depends on the time of exposure, temperature, cryoglobulin concentration, and the interaction of cryoglobulins with other plasma proteins. The following indices may point to the presence of cryoglobulins:

- Very different cell counts in different investigations
- Blue sediments in differential count samples
- In a sample warmed up to 37°C, significantly lower cell counts are measured

As is the case with cold agglutinins, for adequate analysis of samples in which cryoglobulins are suspected, blood sample should be kept warm at 37°C and analyzed immediately afterward.

REVIEW QUESTIONS

1. Which of the following statements best describes the way interference factors may be reduced or eliminated?
   a. By standardizing the preanalytical conditions
   b. By providing proper instructions to the patients on how to prepare for blood sampling
   c. By selecting a more specific method
   d. By selecting the appropriate sampling procedure
   e. By maintaining the analytical variability (method CV_A) to a minimum
2. Spectrophotometric interference of hemolysis occurs due to the ability of hemoglobin to absorb light at which wavelengths?
   a. 400, 500, and 600 nm
   b. 550 and 570 nm
   c. 540 and 600 nm
   d. 415 and 540 nm
   e. 415, 540, and 570 nm
3. Recommended sample for the accurate measurement of potassium is which of the following?
   a. Plasma
   b. Serum
   c. Whole blood
   d. Capillary blood
   e. Arterial blood
4. Which of the following statements is true for lipid testing and testing for lipid-soluble drugs and hormones?
   a. Testing should always be done in a delipidated sample.
   b. Testing should always be done on the native sample before delipidation.
   c. Delipidation does not affect the concentration of lipids and lipid-soluble drugs.
   d. The most suitable delipidation method is ultracentrifugation.
   e. The most suitable delipidation method is lipid removal using the lipid clearing agents.
5. Major sources of exogenous interferences are described below, EXCEPT?
   a. Prescribed medications
   b. Supportive medical therapy like parenteral emulsions, contrast media agents, or infusion solutions
   c. Dietary supplements
   d. Accidental exposure and poisoning
   e. In vivo hemolysis
6. According to the International Council for Standardization in Haematology, the anticoagulant of choice for hematology testing is:
   a. dipotassium EDTA.
   b. tripotassium EDTA.
   c. disodium EDTA.
   d. 3.8% sodium citrate.
   e. 3.2% sodium citrate.

SUGGESTED READINGS

Baird, G. (2013). Preanalytical considerations in blood gas analysis. *Biochemia Medica*, *23*(1), 19–27.

Banfi, G., & Germagnoli, L. (2008). Preanalytical phase in haematology. *Journal of Medical Biochemistry*, *27*, 348–353.

Bonar, R., Favaloro, E. J., & Adcock, D. M. (2010). Quality in coagulation and haemostasis testing. *Biochemia Medica*, *20*, 184–199.

Chiappin, S., Antonelli, G., Gatti, R., & De Palo, E. F. (2007). Saliva specimen: a new laboratory tool for diagnostic and basic investigation. *Clinica Chimica Acta*, *383*(1–2), 30–40.

Delanghe, J., & Speeckaert, M. (2014). Preanalytical requirements of urinalysis. *Biochemia Medica*, *24*(1), 89–104.

Dimeski, G. (2008). Interference testing. *Clinical Biochemist Review*, *29*(Suppl. 1), S43–S48.

Dimeski, G., & Jones, B. W. (2011). Lipaemic samples: effective process for lipid reduction using high speed centrifugation compared with ultracentrifugation. *Biochemia Medica*, *21*, 86–94.

Favaloro, E. J., Adcock Funk, D. M., & Lippi, G. (2012). Preanalytical variables in coagulation testing associated with diagnostic errors in hemostasis. *LABMEDICINE*, *43*(2), 1–10.

Harrison, P., Mackie, I., Mumford, A., Briggs, C., Liesner, R., Winter, M., et al. (October 2011). Guidelines for the laboratory investigation of heritable disorders of platelet function. *British Journal of Haematology*, *155*(1), 30–44.

International Council for Standardization in Haematology (ICSH). (1993). Recommendations of the ICSH for ethylene diamine tetraacetic acid anticoagulation of blood for blood cell counting and sizing: expert panel on cytometry. *American Journal of Clinical Pathology*, *100*, 371–372.

Lippi, G., Daves, M., & Mattiuzzi, C. (2014). Interference of medical contrast media on laboratory testing. *Biochemia Medica*, *24*, 80–88.

Lippi, G., Franchini, M., Montagnana, M., Salvagno, G. L., Poli, G., & Guidi, G. C. (2006). Quality and reliability of routine coagulation testing: can we trust that sample? *Blood Coagul Fibrinolysis*, *17*, 513–519.

Narayanan, S. (2000). The preanalytic phase – An important component of laboratory medicine. *American Journal of Clinical Pathology*, *113*, 429–452.

Nikolac, N. (2014). Lipemia: causes, interference mechanisms, detection and management. *Biochemia Medica*, *24*(1), 57–67.

Nunes, L. A., Mussavira, S., & Bindhu, O. S. (2015). Clinical and diagnostic utility of saliva as a non-invasive diagnostic fluid: a systematic review. *Biochemia Medica (Zagreb)*, *25*(2), 177–192.

Saracevic, A., Nikolac, N., & Simundic, A. M. (2014). The evaluation and comparison of consecutive high speed centrifugation and LipoClear® reagent for lipemia removal. *Clinical Biochemistry*, *47*(4–5), 309–314.

Simundic, A. M. (2015). Blood gases, ions and electrolytes; Who is doing Phlebotomy in Europe? In W. G. Guder, & S. Narayanan (Eds.), *Pre-examination procedures in laboratory diagnostics. Preanalytical aspects and their impact on the quality of medical laboratory results*. De Gruyter.

Simundic, A. M., Cornes, M., Grankvist, K., Lippi, G., & Nybo, M. (2014). Standardization of collection requirements for fasting samples: for the Working Group on Preanalytical Phase (WG-PA) of the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM). *Clinica Chimica Acta*, *432*, 33–37.

Simundic, A. M., Nikolac, N., & Guder, W. G. (2017). Preanalytical variation and preexamination processes. In N. Rifai, R. Horvath, & C. Wittwer (Eds.), *Tietz textbook of clinical chemistry and molecular diagnostics* (6th ed.). Elsevier.

Zini, G. (2014). International Council for Standardization in Haematology (ICSH) Stability of complete blood count parameters with storage: toward defined specifications for different diagnostic applications. *International Journal of Laboratory Hematology*, *36*(2), 111–113.

4

Biological Variation

Sverre Sandberg, Thomas Røraas, Aasne K. Aarsand, Callum G. Fraser

OBJECTIVES

1. Explain the nature of biological variation and how estimates of within- and between-subject biological variation can be generated.
2. Explain how estimates of biological variation can be applied in laboratory and clinical practice.
3. Design biological variation studies for measurands in clinical chemistry.
4. State the formula for reference change value calculations and illustrate its use with examples.
5. Describe the various models for setting analytical performance specifications, and critically assess the advantages and disadvantages of each model.

KEY WORDS AND DEFINITIONS

Analytical performance specifications The analytical performance needed to fulfill certain requirements—often given as numerical values.

Analytical variation (CV_A) Variation of the concentration of a measurand in a sample measured several times with the same measurement procedure (imprecision)

Between-subject variation (CV_G) The variation in the concentration of a measurand among the homeostatic set points of different individuals.

Biological variation Variation in the concentration of a measurand in the body not caused by disease (or analytical variation).

Critical difference The change in a result that, with a certain probability, makes it significantly different from a previous result. The preferred synonym is reference change value (RCV; see later).

Index of individuality (II) The distributions of measured values from samples from healthy individuals in relation to the population-based reference interval.

$$II = (CV_A^2 + CV_I^2)^{1/2}/CV_G$$

Reference change value The change in a result that, with a certain probability, makes it significantly different from a previous result.

Within-subject variation (CVI) The variation in the concentration of a measurand within each individual around their own homeostatic set point.

There are many sources of variation in numerical results generated in laboratory medicine. Although some measurands have biological variations over the span of life, and others have predictable cyclical variation, most measurands have random variation around homeostatic set points, which differ between individuals. Knowledge of the generation and application of data on within-subject and between-subject biological variation is essential for the correct interpretation of results.

NATURE OF BIOLOGICAL VARIATION

There are many sources of variation that contribute to the uncertainty of any result generated in laboratory medicine. **Biological variation** is one of the most important and should be taken into account when results are interpreted.

There are various types of biological variation. The concentration or activity of some measurands changes over the span of life—some slowly and some more quickly, particularly at times of physiological change, such as the neonatal period, childhood, puberty, the menopause, and old age. The concentration or activity of measurands can also differ between men and women. This variation is taken care of by the creation of age- and/or sex-stratified (partitioned) reference intervals when these are needed. In addition, many measurands have predictable cyclical rhythms in their concentrations. These can be daily, monthly, or seasonal in nature. The major ramifications for interpretation are that reference intervals cannot be generated for every point during these cycles. Therefore specimen collection must be performed at appropriate times.

TABLE 4.1 Serum Sodium Activity (mmol/L) in Four Specimens Collected at Daily Intervals from Each of a Cohort of Four Individuals

Sodium	Day 1	Day 2	Day 3	Day 4
Individual 1	137	139	136	138
Individual 2	144	146	145	144
Individual 3	141	143	142	140
Individual 4	139	138	141	140

The most important type of biological variation is random biological variation. As an example, four specimens were taken from four individuals at daily intervals, and serum sodium activity (reference interval: 135 to 147 mmol/L) was examined (Table 4.1). It is evident that the results for each individual vary from day to day; this is due to three sources of variation, namely, preanalytical, analytical, and within-subject biological variations. In addition, each individual has a different average serum sodium activity, which is termed the homeostatic set point. The variation among the homeostatic set points of individuals is the **between-subject variation**, whereas the variation within each individual around their own homeostatic set point is the **within-subject variation**. The individuality of a measurand and the use of conventional population-based reference intervals are determined by comparison of the within-subject and between-subject biological variations. Within-subject biological variation and analytical imprecision can be used to create **reference change values** (RCVs), sometimes called **critical differences**, to assess the statistical significance of changes in serial results from an individual or the probability that any change seen is significant. Analytical performance specifications (APS) for imprecision, bias, allowable total error, measurement uncertainty, and other characteristics also can be created using within-subject and between-subject variation. The generation and subsequent application of numerical data on the components of biological variation are crucial facets of laboratory medicine and are described in detail in this chapter.

The range of terms and the variety of symbols used to define the components of biological variation have grown with the increasing body of literature. The internationally agreed terms and symbols used throughout this chapter are:

- CV_I: within-subject biological variation (variation within a single individual estimated as a pooled variation from a [homogeneous] group of individuals)
- CV_G: between-subject biological variation (variation between the central tendencies of a group of individuals)
- CV_A: **analytical variation** (analytical imprecision)

GENERATION OF DATA ON COMPONENTS OF BIOLOGICAL VARIATION

Production of data on biological variation is similar to derivation of population-based reference intervals except that, instead of one specimen being taken from many reference individuals, the specimens are taken from a smaller cohort. In general terms, the traditional approach recommends that numerical estimates of CV_A and both CV_I and CV_G components of biological variation should be generated using the following experimental approach:

- Select a group of reference individuals (usually apparently healthy volunteers)
- Take a set of specimens from each of the individuals at regular time intervals while minimizing all sources of preanalytical variation in preparation of the subjects for specimen collection
- Transport specimens in a standardized way and store aliquots under controlled conditions until ready for analysis
- Undertake the analyses in duplicate while minimizing analytical sources of variation
- Remove statistical outliers and confirm that all results from the individuals are homogeneous
- Dissect out the CV_A, CV_I, and CV_G components using nested analysis of variance (ANOVA) (or alternatively general linear models or generalized linear models)

Design of Studies

The preceding general design has been widely used and is very suitable for those measurands that have low CV_I and tight homeostatic control, but it is somewhat simplistic for many reasons. The measurand may be unstable, and analyses would need to be performed soon after the collection of specimens (e.g., for some hematologic measurands, such as mean cell volume, and erythrocyte and leukocyte counts per volume). In this case, to obtain the necessary statistically unconfounded estimate(s) of CV_I, quality control materials have to be analyzed during each run to ascertain that variation, due to random or systematic error in the procedure between each analysis, is not introduced. Using this strategy, the quality control material should preferentially be commutable; meaning that the variations of the analyses of the individual samples and the quality control materials are comparable. Furthermore, it must be ensured that the concentrations or activities of the quality control materials are similar to those of the samples from the subjects studied because CV_A often varies with concentration or activity.

Estimates in laboratory medicine are usually accompanied by confidence intervals (CIs); unfortunately, this has rarely been delivered in most published reports to date that provide estimates of the components of biological variation. CIs are essential for appraising the results and comparing results between studies. It has been found that the reliability of an estimate for CV_I is greatly influenced by the study design and by the ratio between CV_A and CV_I. For a limited number of measurements, it is preferable to have a relatively high number of specimens from each individual. On the other hand, if the influence of factors, such as age and sex, on within-subject biological variation is under study, it is important that an adequate number of individuals are included in each sub-group to detect clinically important differences. If the CV_A is high compared to the CV_I, either the total number of specimens or the number of replicate analyses should be increased.

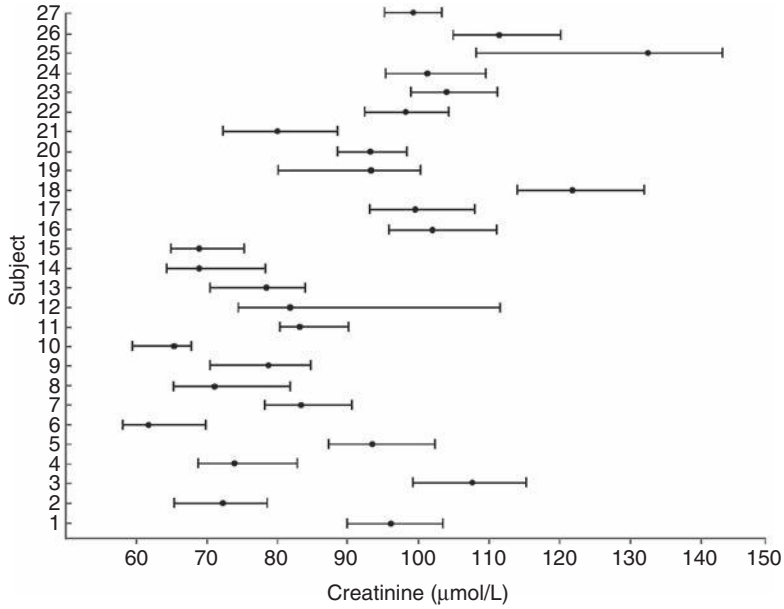


Fig. 4.1 Means and extreme values for serum creatinine in 27 older adults. Note: 100 μmol/L = 1.13 mg/dL. (From Fraser, C. G. [2004]. Inherent biological variation and reference values. *Clinical Chemistry and Laboratory Medicine, 42,* 758–764.)

Methods for Analysis

Steady State, Outliers, and Homogeneity

The currently accepted best method for sample collection and data analysis is detailed by Fraser and Harris, whereby duplicate analyses are performed on samples from a cohort of individuals. However, before any components of variation are estimated, it is important (1) to verify that the individuals are in a steady state; (2) to exclude outliers in the data set; and (3) to assess whether the individuals have a homogeneous within-subject CV_I.

Steady state. The calculation of biological variation data assumes that the individuals assessed are in a "steady state," that is, the homeostatic set points do not change during the time span of the study. If this is not the case, data should be transformed to a "steady state," for example, by correcting for trends using regression analysis.

Outliers. The assessment of outliers is important because such aberrant values will lead to erroneous estimates of the components of biological variation. It is important that this assessment is done using the same measure of variability that is estimated; that is, if coefficients of variation (CVs) are estimated, which is usually the case, all the calculations should be performed using CVs. This can be achieved by normalizing data, for example, through $\log_e$-transformation. After transformation, outlier assessments are performed at three levels: (1) between duplicates or replicates, (2) between samples within an individual, and (3) between individuals. For levels 1 and 2, typically Cochrane's test is used, but applied to the CV^2 and not to standard deviation $(SD)^2$. Failure to remove outliers in the replicates can result in a falsely high CV_A and an erroneous CV_I while failure to remove outliers from results from each of the individuals can result in a falsely increased CV_I. Finally, outliers among the mean values of the individuals (level 3) are assessed. A simple strategy to perform this process is to use Reed's criterion—that the difference between any mean value and the next value in the series should be less than one-third of the overall range of all values. Another useful approach for the assessment of outliers between individuals is a simple graphical approach in which the mean values and the range of all these values are plotted for each individual (on the *y*-axis) against concentration or activity (on the *x*-axis). An example is provided in Fig. 4.1 and discussed later in this chapter. Failure to exclude outliers of the mean values of the different individuals will result in a falsely large CV_G, and because the overall mean value will be different; this may also affect CV_A and CV_I depending on the transformation chosen.

Normal distribution. If the data are not normally distributed, CIs cannot be calculated using formula-based methods. It is important to specify that the normality relates to the model effects; that is, that both the analytical variation around the true sample value and the individuals' variation around their homeostatic set point are normally distributed. It does not relate to the total pooled data. Pooling the standardized residuals for each level, namely, residuals from replicates (difference between replicates and mean of replicates from each sample), residuals from samples (difference between mean of samples and mean for the individual), and residuals from subjects (differences between individual means and total mean), can be used to assess normality.

Example: Assume observations of 2.25, 2.50, and 2.75 are from individual A and 3.50, 4.00, and 4.50 from individual B. Standardized residuals are generated by first dividing by the mean for each individual. Individual A has average 2.50, so standardized observations are 0.90, 1.00, and 1.10, and corresponding standardized residuals are −0.10, 0.00,

and 0.10. Individual B has average 4.00, standardized observations 0.875, 1.00, and 1.125, and corresponding residuals −0.125, 0.00, and 0.125. The pooled standardized residuals are (−0.10, 0.00, 0.10, −0.125, 0.00, 0.125). The standardized residuals can then be examined using Kolmogorov-Smirnov or Anderson-Darling or other techniques for the assessment of normality.

Homogeneity. Applications of within-subject biological variation data, particularly for estimation of RCVs, depend on the within-subject variation data being homogeneously distributed. If the specimens have not been obtained from a homogeneous population, the results will not be representative of the population, only an average individual, and stratified analysis by subgroups should be applied. Thus, although estimates of CV_I and RCVs can be calculated, these are not generalizable to the overall population. It is therefore important to know that the variances in specimens drawn from a population are "homogeneous" by definition, and consequently, that the ranked cumulative distributions of these variances are distributed around the true variance of the population according to χ^2/df (χ^2 distribution for degrees of freedom [df] according to the individual sample sizes). In contrast, when a series of different variances have a dispersion around the pooled variance according to a χ^2/df distribution, they are considered to be heterogeneous. This can be illustrated by plotting the cumulated ranked fractions of within-subject variations as a function of the within-subject variation estimates on a Rankit scale (Fig. 4.2). If homogeneous, this curve will fit to the theoretical of the square root of the pooled variance multiplied by χ^2/df. Variance homogeneity can also be tested further by Bartlett's test. Although Cochrane's test of the variances of the mean values of the specimens is primarily an outlier test, it also can be used as a test for homogeneity.

After testing for homogeneity, it is important to indicate how many individuals and results had to be removed to obtain homogeneity of the data used to estimate the CV_I. This again provides an indication of the representative nature of the data and underscores its suitability for wide application. It is also recommended to check if excluded individuals share common traits that can explain their heterogeneity, which can be valuable information to the applicability of the estimated CV_I.

Data Analysis

There are several estimators available: ANOVA, maximum likelihood (ML), restricted maximum likelihood (REML), minimum variance quadratic unbiased estimator (MIVQUE), and weighted analysis of means (WAM). For balanced designs, the choice of estimator is not important; however, for unbalanced designs, the REML, MIVQUE, or WAM are preferred for the variance components. These estimators are available in most statistical software packages using general linear models or generalized linear models.

In previous publications, many have not used formal ANOVA but instead have simply used subtracted variances. The thesis is that, because preanalytical sources of variation have been minimized and can be considered negligible, the total CV (CV_T) of a set of results from each cohort of individuals includes CV_A, CV_I, and CV_G. Then, because:

$$CV_T = \left[(CV_A)^2 + (CV_I)^2 + (CV_G)^2\right]^{1/2}$$

the components can be calculated by simple subtraction. However, this approach will require calculations including degrees of freedom and, therefore, a formal ANOVA is a simpler approach for correct calculations and can take into account unbalanced designs. Additionally, in many of these studies, CV_A has been estimated based on quality control materials, the outliers have not been assessed, and a normal distribution and homogeneity of data are assumed; in consequence, estimates of the components of biological variation are likely to be less precise.

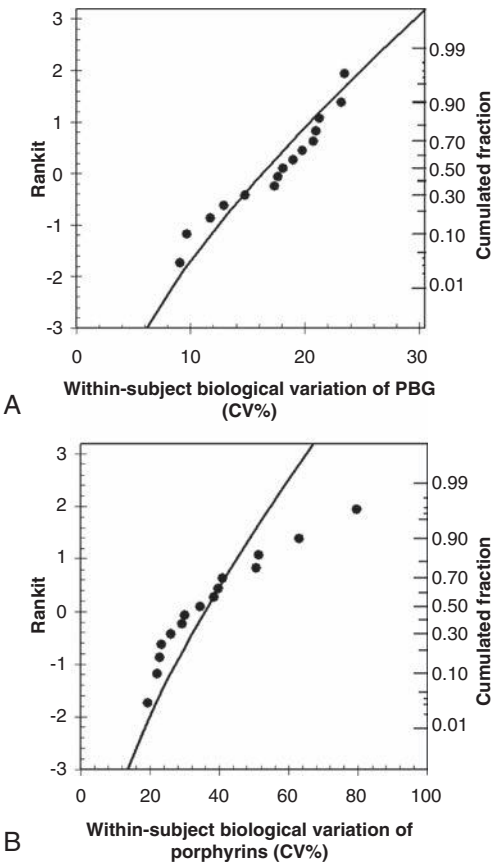


Fig. 4.2 Variance homogeneity plots. Rankit plots show the accumulated fractions as function of within-subject biological coefficient of variations *(CVs)*. The filled circles represent individual CVs for healthy individuals; the *line* indicates the expected distribution of measured CV for "true" CV values of 17.3% for porphobilinogen *(PBG)* (A) and 38.1% for porphyrins (B). The Cochrane test gives values of 0.11 for PBG and 0.25 for porphyrins, and a value indicating heterogeneity is >0.17. *CVs*, Coefficients of variation. (From Aarsand, A. K., Petersen, P. H., & Sandberg, S. [2006]. Estimation and application of biological variation of urinary delta-aminolevulinic acid and porphobilinogen in healthy individuals and in patients with acute intermittent porphyria. *Clinical Chemistry, 52,* 650–656.)

DATABASES

Many data have been generated over the last 45 years on the components of biological variation in a broad range of measurands. Before considering the uses of the existing published data, it is necessary to consider their reliability. In many publications, duplicate measurements are lacking and outlier analysis and homogeneity testing have not been performed. For a few measurands, reviews have been performed on the robustness of estimates; examples are glycated hemoglobin (HbA_{1c}) and serum enzymes (Fig. 4.3). These reviews concluded that there is a great variation of CV_I estimates, probably due to factors including examination methodology, population selection, specimen collection frequencies, protocol application, and statistical analyses.

Current Databases and Their Merits and Disadvantages

As stated previously, derivation of the components of biological variation is not without difficulties and requires considerable resources; hence the need for a biological variation database of published information. Generation of such a comprehensive database was initiated in 1997 by the Analytical Quality Commission of the Spanish Society of Clinical Chemistry (SEQC), and it was first published in 1999. The last update of this database was in 2014. The database, which has been much cited and widely used, has many advantages. CV_I has been determined for 358 measurands in 247 articles, and this large resource has been developed and refined over nearly 20 years. Data are available on measurands in various matrices, namely, serum (n = 185), plasma (n = 74), whole blood (n = 55), and urine (n = 47).

However, due to concerns over the quality of evidence available on biological variation, the database should be used with caution. In addition, this database includes lack of data for many measurands of interest in laboratory medicine, and there has been a paucity of new publications available each year for inclusion in the database, with only 25% published since 2000. Limited biological variation data are available for some measurands; for example, 202 were found in a single publication, 129 had data from 2 to 9 publications, and 27 had data from 10 or more publications. Further, CIs for the estimates, allowing for the comparison of these results, are lacking, and thus, the robustness of the data is difficult to assess.

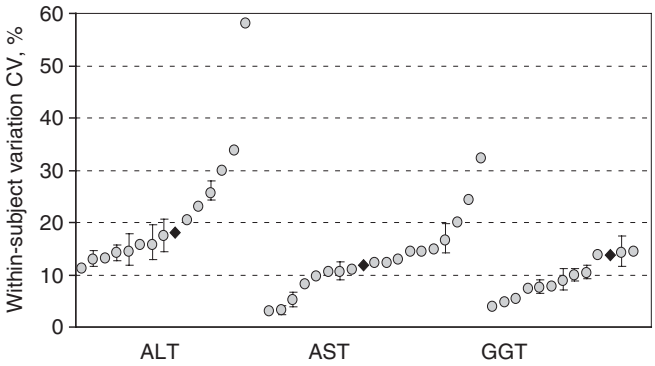


Fig. 4.3 Estimates of within-subject biological variation from different studies for three enzyme activities. *ALT*, Alanine aminotransferase; *AST*, aspartate aminotransferase; *CV*, coefficient of variation. *GGT*, γ-glutamyl transferase. (From Carobene, A., Braga, F., Roraas, T., Sandberg, S., & Bartlett, W. A. [2013]. A systematic review of data on biological variation for alanine aminotransferase, aspartate aminotransferase and γ-glutamyl transferase. *Clinical Chemistry and Laboratory Medicine*, *51*, 1997–2007.)

Due to growing concerns about the apparent heterogeneity and applicability of the estimates, an Expert Working Group on Biological Variation of the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) produced a checklist to enable standardized production and reporting of biological variation data. The checklist identifies key elements that are essential for the safe, accurate, and effective transfer of biological variation data sets across health care systems. Following this work, a new EFLM Task and Finish Group has produced another tool, the Biological Variation Data Critical Appraisal Checklist (BIVAC), which is used to assess the methodological quality of existing studies. The result from this work will be used to populate a new database on the EFLM website with those estimates that have sufficient quality. An example of data generated on nine enzymes in serum using these up-to-date approaches has been recently published. Moreover, compliance with the checklist for new studies will enable authors, reviewers, and journal editors to ensure that studies are fit for purpose, appropriately powered, share common terminology, and deliver robust estimates of CV_I and CV_G, accompanied by the key metadata to enable general application.

Biological Variation in Health and Disease

Because the quality of published papers varies, it is strongly recommended that all users investigate the sources and quality of the data selected before application. It is planned that high-quality data will soon be available once the EFLM database is up and running. Some argue that many of the data on the components of biological variation are inappropriate for wider use in laboratory medicine because they have been derived from studies on healthy individuals, and not on patients with disease who are the primary targets of most laboratory examinations.

The group from the Analytical Quality Commission of the SEQC has published a database on the components of biological variation in disease. The 2014 database has information on 97 measurands in subjects with 41 disease states. This work indicates that estimates of CV_I are generally independent of the state of health, except when the measurand is the one that is pathologically changed, such as tumor markers in patients with cancers.

Generation of Data from Patient Populations

Because estimates of CV_I, a measure of the homeostatic mechanisms in individuals, seem constant, "big data" from hospital laboratories might also be used to determine this component of biological variation. For example, estimation of CV_I of HbA_{1c} was done in specimens taken routinely from children

with cystic fibrosis who had no evidence of diabetes mellitus or impaired glucose tolerance. Similarly, it would be difficult to justify collecting multiple samples from apparently healthy individuals for examinations, such as arterial pH, gas partial pressures, and electrolytes; derivation of CV_I from results of examinations in diseased individuals performed regularly in the intensive care unit has provided reliable estimates. These interesting examples showed that, although it is preferable to undertake the well-documented experimental protocols and statistical analyses, novel approaches might be required for examinations where the generation and application of traditional estimates of CV_I are difficult.

INTERPRETATION AND USE OF DATA

Individuality

The results of most examinations in laboratory medicine are compared with conventional population-based reference intervals or sometimes fixed limits for clinical decision-making. Reference intervals represent the values found in 95% of the reference population. The ramification of biological variation on the use of reference intervals is determined by the individuality of the measurand.

Fig. 4.1 shows a graph, which should always be prepared when generating data on the components of biological variation to visually assess the presence of outlying observations. This shows the means and extreme values in a cohort of 27 older adults for serum creatinine concentration. Subjects 1 to 13 were women and subjects 14 to 27 were men. The conventional reference intervals for creatinine generated in the laboratory for individuals older than 55 years of age were 60 to 98 μmol/L (0.68 to 1.10 mg/dL) for women and 66 to 128 μmol/L (0.75 to 1.45 mg/dL) for men.[a] The following conclusions can be drawn when assessing the data in Fig. 4.1:

- No individual has observed values that span the entire reference interval.
- Most individuals have all observed values within the reference interval.
- The means of the observed values of most individuals lie within the reference interval, but they are different from each other.
- A few individuals have observed values that span the lower or the upper reference limit, and these individuals have values that change from usual to unusual over time.

It is clear that CV_I (the variation around the homeostatic set points) of creatinine is smaller than the CV_G (the difference among homeostatic set points). In numerical terms, CV_I is 4.3% and CV_G is 18.3%. Thus CV_I is less than CV_G, meaning that the measurand has marked individuality. This characteristic can be expressed mathematically as an **index of individuality** (II) and is best calculated as: $II = \left(CV_A^2 + CV_I^2\right)^{1/2}/CV_G$. It is common now for II to be simply calculated as CV_I/CV_G. This is satisfactory if CV_A is less than ½ CV_I, as is often the case with modern analytical technology and methodology, because CV_A will then contribute little to the numerator $\left(CV_A^2 + CV_I^2\right)^{1/2}$. Most common measurands have marked individuality with a low II.

This example provides a biological explanation for the fact that serum creatinine concentration, when compared with conventional reference intervals, even if partitioned by age or sex, does not have high sensitivity for the detection of mild renal impairment, and provides a sound rationale for the undoubted benefits of the estimated glomerular filtration rate. The estimated glomerular filtration rate uses formulae that take age, sex, and ethnicity into account, as well as the serum creatinine concentration. This example also highlights that most measurands examined in laboratory medicine are not very useful in population screening or in case finding.

Consequences for Population-Based Reference Intervals

The consequences of individuality were first postulated by Harris who showed that, when II is high (the criterion usually applied is II >1.4), the distribution of values from any single individual will cover much of the entire dispersion of the reference interval derived from values found in reference individuals. In contrast, if II is low (especially when II is <0.6), the dispersion of values for any individual will span only a small part of the conventional population-based reference interval.

The ramifications of this individuality on the interpretation of the results of examinations are profound. When II is low, most individuals can have values that are unusual for them, but these will often lie within the reference interval. Consequently, these results would not be flagged by laboratories as deserving further attention because, although they are unusual for that individual, they are within the reference interval. Moreover, the users of the laboratory results would be highly unlikely to pay attention to such unusual values. Thus taking one specimen from an individual and comparing, for example, its creatinine result to the population-based reference interval will not be an effective way of picking up the small changes often seen in early pathological processes. However, when only one sample is examined in an individual, the II will have no influence on the percentage of false positives and true positives detected, irrespective of whether the upper reference limit or a clinical decision limit is used. However, if a "confirmatory" measurement is performed, the II is important. For quantities with a very low II, such as creatinine, the new result will be close to the first, and will only provide limited new information. For quantities with a high II, a repeat measurement will decrease the number of true positives and false positives.

If II is defined as CV_I/CV_G, this ratio must be made larger to make conventional population-based reference intervals of higher clinical usefulness, especially in diagnosis, case finding, and screening, when no previous results on an individual are available. This is sometimes easily achieved because CV_G can be made smaller by stratifying (or partitioning) the data. An example is shown in Table 4.2.

[a] Laboratories should verify that these ranges are appropriate for use in their own settings.

TABLE 4.2 Within-Subject (CV_I) and Between-Subject (CV_G) Biological Variation of Urine Creatinine and Indices of Individuality (II)

Group	CV_I (%)	CV_G (%)	II
Men (n = 7)	11.0	6.0	1.83
Women (n = 8)	15.7	11.0	1.42
Total	13.0	28.2	0.46

Table 4.2 shows that the II for the total cohort is 0.46; therefore the reference interval is of low usefulness, especially for monitoring individuals. However, for men and women taken separately, the II are 1.83 and 1.42, respectively, and reference intervals will be very useful. Stratification by sex has vastly increased the usefulness of the conventional population-based reference interval. Knowledge of individuality gives a sound scientific basis for stratification; II must be high for conventional, population-based reference intervals to be useful. Because CV_I is generally considered constant and rarely stratified, it must be CV_G that is made smaller, if possible.

REFERENCE CHANGE VALUES

Most laboratory results are used for monitoring acute diseases in the short term in secondary or tertiary care institutions or chronic conditions in primary care. *Monitoring,* by definition, means assessment of results over time. Because most measurands have marked individuality and low II, conventional population-based reference intervals provide limited assistance when serial results of an individual need to be interpreted.

Harris first introduced the concept of evaluating consecutive measurements in 1975 and 1976, and the concept was later simplified and introduced by Harris and Yasaka as the RCV, sometimes also referred to as *critical difference.* The former term is preferred to indicate an analogy with population-based reference intervals.

The result of one examination will have: dispersion $= Z \times \left(SD_A^2 + SD_I^2\right)^{1/2}$, where Z is the Z-score equal to the number of standard deviations appropriate for the probability desired. The result of a second examination will have the same dispersion, so the total dispersion of two results will be $2^{1/2} \times Z \times \left(SD_A^2 + SD_I^2\right)^{1/2}$. Thus, for two results on the same individual to be different, this inherent difference due to SD_A and SD_I must be exceeded, and this is the RCV.

The original RCV, as shown previously, is only applicable when SDs are used (i.e., changes expressed in units). When used for changes expressed in percentages (i.e., CV_A and CV_I combined using CVs), a transformation is required. This is due to the fact that the sum of normally distributed variables is normally distributed, but the ratio between normally distributed variables is not.

A sufficient approximation for RCV based on an already established CV_I and the individual laboratory's CV_A can be estimated using the formula:

$$SD_A^2 = \log_e\left(\left(\frac{CV_A\,\%}{100\,\%}\right)^2 + 1\right)$$

$$SD_I^2 = \log_e\left(\left(\frac{CV_I\,\%}{100\,\%}\right)^2 + 1\right)$$

$$RCV_{\%} = \pm 100\,\% \times \left(\exp\left(\pm Z_\alpha \times \sqrt{2} \times \sqrt{SD_A^2 + SD_I^2}\right) - 1\right)$$

Example: With a within-subject biological variation found to be CV_I = 5.1% and between-run CV_A from the laboratory = 1.4%, the $\log_e$-transformed measure of variability is $SD_I^2 = \log_e(0.051_A^2 + 1) = 0.00259762$ and $SD_A^2 = \log_e(0.014_A^2 + 1) = 0.00019598$. The correct RCV in % can then be calculated as RCV = $\pm 100 \times \left(\exp\left(\pm 1.65 \times \sqrt{2} \times \sqrt{0.00019598 + 0.00259762}\right) - 1\right)$ = (−11.6% + 13.1%). The RCVs are skewed, with bigger tolerance for a rise than a fall in values.

It is usually assumed that preanalytical sources of variation are negligible, and in clinical and laboratory practice this means having well-documented standard operating procedures for patient preparation and specimen collection, transport, and handling before analysis, as well as appropriate training of staff performing these tasks. Moreover, it is important to realize that changes in the bias of the analysis between the collections of the serial specimens can also add to the RCV; if these can be quantitated, as a difference due to bias in percentage terms, ΔB, the formula becomes $RCV = \Delta B + 2^{1/2} \times Z \times \left(SD_A^2 + SD_I^2\right)^{1/2}$. However, in a single laboratory the main source of ΔB over time is due to recalibration, and this "random" bias is usually an integral component of the longer term CV_A as estimated from replicate analyses of internal quality control materials. Thus it can be assumed that a "systematic" bias of the analysis does not change during the period between the two analyses, and the simpler RCV formula applies.

It is often assumed that a Z-score of 1.96 for $P < .05$ (and sometimes also 2.58 for $P < .01$) is appropriate. It is almost ubiquitously stated in studies on biological variation that the RCV is calculated as $2.77 \times \left(SD_A^2 + SD_I^2\right)^{1/2}$. This is incorrect for various reasons. These Z-scores are termed bidirectional (or two-tailed or two-sided), and this infers that the difference between the two serial results can be either an increase or a decrease. However, in most situations, clinical decisions rely on the assessment of a significant decrease (e.g., of HbA_{1c} after treatment for diabetes mellitus or in glucose after adjustment of insulin dosage) or a significant increase in the measurand (e.g., an increase in serum creatinine to assess acute kidney injury, or serum troponin after acute chest pain). Thus unidirectional (one-tailed or one-sided) Z-scores must be used in most clinical situations to facilitate correct interpretation; these are 1.65 for $P < .05$ and

2.33 for $P < .01$. Precise definition of the real clinical setting and accurate description of what is considered a "change" in terms of "increase" or "decrease" and their many synonyms are required for correct calculation and use of RCVs in clinical decision making.

It is also important to recognize that clinical decision making is not always done at $P < .05$, which is the most commonly used probability in analysis of research data. The semantics used are crucial to understanding the probability that is clinically appropriate. An example was given in a recent study on RCVs for dehydration markers, which, in addition to graphs that showed probability against change for plasma osmolality, urine specific gravity, and body mass, integrated a series of semantic interpretative anchors; namely, change was "likely" at $P > .80$, "more likely" at $P > .90$, "very likely" at $P > .95$, and "virtually certain" at $P > .99$. RCVs should be used in a spectrum of postanalytical processes, including provision of graphs and tables of change versus probability, Δ-checking, and flagging of significant changes at different levels of probability on laboratory reports.

In addition, because RCV is dependent on both SD_A and SD_I, RCVs must be calculated for each individual laboratory since they depend on the SD_A achieved locally. Furthermore, it is essential that the SD_I used is reliable and obtained from a population with a homogeneous SD_I that is similar to the population for whom the RCV is being applied. In addition, the time interval used for obtaining the SD_I must be comparable to the one used in practice. Thus laboratories can calculate relevant RCVs by using SD_A derived from their own internal quality control programs, using data close to clinical decision-making concentrations or activities, along with robust SD_I estimates from publications cited in the most up-to-date database.

Traditional RCVs are believed by some to be rather simplistic, especially because they only address how likely it is that a certain change can be explained by SD_A and SD_I, but not the probability that a change in the disease state has occurred. It has been suggested that a tool for better understanding and interpretation of measured differences in monitoring is needed; that the concepts of sensitivity, specificity, likelihood ratios, and odds used for diagnostic test accuracy evaluations should be applied to monitoring by substituting measured concentrations with measured differences. It was suggested that this idea expanded the earlier concept of RCV by making it possible to have an estimate of the posttest odds for a certain difference to occur. Consequently, the likelihood ratio for change increases with a larger measured difference, and when used together with the pretest odds or pretest probability, the posttest odds and posttest probability, which are related to the clinical situation, can be calculated (see Chapter 2 for details).

It has been proposed that the probability of significance of any difference seen in clinical practice between two results can be calculated using a simple rearrangement of the RCV equation making the Z-score (and thus the probability) the unknown, namely:

$$Z = \text{difference} \Big/ \left[2^{1/2} \times \left(SD_A^2 + SD_I^2\right)^{1/2}\right]$$

This does not seem to have been applied much in practice, but there is clearly scope for adoption of this technique to enhance interpretation of serial results. Exactly as for RCVs, this application would have important consequences for SD_A; the smaller the SD_A, the smaller the RCV will be for any probability, and the significance of any difference seen will be of higher probability.

Reference Change Values for More Than Two Serial Analyses

Frequently, in practice, more than two results of analysis are available for the individuals over time. Using the traditional RCVs described previously, it is only possible to calculate the significance of changes between each of the two consecutive analyses. Thus an RCV method including all available serial results might be useful for interpretation of significant differences over time.

Mathematical methods have been developed to assess serial results from an individual, which are sometimes called methods for time series analysis. One is called the *homeostatic model*. This model assumes that the measurand varies randomly around a homeostatic set point. After collection of results from a small number of analyses, the mean and SD are calculated. The next result should fall within the range calculated from the mean and the SD. Then, new data from the same individual are used in an ongoing manner to refine the estimates of the mean and SD for that individual to interpret whether the next new result has changed significantly from previous data. In contrast, the "random walk" model assumes that the measurand behaves randomly, and there is no homeostatic set point. The dispersion of each result is calculated. However, instead of calculating a mean for the individual and recalculating this mean as further data are gathered, it is assumed that the most recent result is the starting point from which to assess whether change has occurred. These models have not been widely applied; they generate many false positive signals that change has occurred, when there is no obvious usual or pathological change. They also seem too complex and time consuming to be used in everyday practice.

Recently, studies on both unidirectional and bidirectional changes in serial results that used computer simulations have been performed. Factors used to multiply the first result from an individual were calculated to create the limits for constant cumulated significant differences. The factors were shown to become a simple function of the number of results and the total CV. The first result is multiplied by the appropriate factor for an increase or a decrease, which gives the limits for a significant difference. It remains to be seen if such apparently simple techniques become used in practice.

Analytical Performance Specifications

Analytical performance specifications (APSs) are the numerical standards of analytical performance that are required to facilitate optimum patient care. There are many strategies available to set these specifications, which have been described over time as this facet of laboratory medicine has evolved. A concise historical perspective has been published recently.

Following pioneering studies done in the United States on the definition of the components of biological variation, a College of American Pathologists conference held in 1976 supported the concept that APS should be best based on biology. The consensus statement, restated in current terms and symbols, was: "For group screening, in which an individual is to be selected from a population, a specification for imprecision (CV_A) is defined as: $CV_A = 0.5 \times \left(CV_I^2 + CV_G^2\right)^{1/2}$" and "For individual single and multipoint testing, in which an individual is evaluated on the basis of discrimination values: $CV_A = 0.5 \times CV_I$."

As the topic of APS matured and a profusion of recommendations appeared, many leading professionals in laboratory medicine realized that there was a need for a global consensus on the setting of such specifications. The Stockholm Conference held in 1999 on "Strategies to set global analytical quality specifications in laboratory medicine" achieved this and advocated the ubiquitous application of a hierarchical structure of approaches, based on a model proposed by Fraser and Petersen. The hierarchy had five levels, stated as: (1) evaluation of the effect of analytical performance on clinical outcomes in specific clinical settings; (2) evaluation of the effect of analytical performance on clinical decisions in general, using (a) data based on components of biological variation or (b) analysis of clinicians' opinions; (3) published professional recommendations from (a) national and international expert bodies or (b) expert local groups or individuals; (4) performance goals set by (a) regulatory bodies or (b) organizers of external quality assessment (EQA) schemes; and (5) goals based on the current state of the art as (a) demonstrated by data from EQA or proficiency testing scheme, or (b) found in current publications on methodology. Since the Stockholm conference, there have been further proposals for setting APS, but all of these have included biological variation data in their derivation.

In 2015 the EFLM organized the First Strategic Conference on Defining Analytical Goals 15 Years after the Stockholm conference. The consensus statement and formal papers emanating from the speakers are included in a Special Issue of *Clinical Chemistry and Laboratory Medicine*, and they represent an invaluable up-to-date resource on all aspects of setting APS and on generation and application of data on biological variation. The consensus was that the Stockholm hierarchical approach was supported but could be simplified. In this revision, the hierarchy was streamlined and represented by three different models to set APS.

Model 1 is based on the effect of analytical performance on clinical outcomes using direct (1a) or indirect (1b) outcome studies. The latter investigates the impact of analytical performance of the investigation on clinical classifications or decisions and thereby on the probability of patient outcomes.

Model 2 is based on components of biological variation of the measurand.

Model 3 is based on the analytical state of the art.

The three models use different principles, and some models will be better suited for certain measurands than for others. Application of model 1 is difficult and will probably be limited to a few measurands. An EFLM Task and Finish Group established after the conference has proposed and visualized through a flow chart a model for how to allocate different measurands to different models. In general, it states that model 2 can be used when the measurand has a stable concentration or activity in serum/plasma and the measurand is not suited for clinical outcome studies or such data do not exist. Generally applicable APS will still be best based on biological variation data for most measurands. In view of the caveats regarding the reliability of certain data on the components of biological variation and the central role measures of biological variation play, there will be significant advances when the EFLM Task and Finish Group on the Biological Variation Database fulfills its remit of creating a database with high-quality estimates, and its recommendations on standards of generation and reporting of data are carried through into practice.

Specifications for Analytical Imprecision

This approach became more popular as the quantity of data on the components of biological variation increased. In most cases, the APS for CV_A was simply taken as $CV_A = 0.5\ CV_I$, with the rationale that the same technology and methodology were used to examine specimens in both of the clinical settings of group screening and individual single and multipoint testing (and other situations). In consequence, the APS derived for the most stringent situation would allow both specifications to be met for all purposes.

Specifications for Analytical Bias

Examination bias was not considered in most early works on setting APS, possibly because the thesis was that laboratories all had their own reference intervals to which the results of their analyses were compared. However, as the interest in harmonization of laboratory data across time and geography increased, and the benefits for harmonized reference intervals were appreciated, criteria for bias needed to be developed. Therefore it was proposed that the APS for bias (B), to allow use of harmonized reference intervals, should be $B < 0.25\left(CV_I^2 + CV_G^2\right)^{1/2}$.

Specifications for Allowable Total Error

As the concepts of total laboratory quality management evolved, and the idea that random and systematic sources of variation (imprecision and bias) were both important, it became clear that total analytical error was the most relevant clinically. It was proposed that the linear model of combining imprecision and bias could be used to set APS for what was termed total error allowable (TEa) as a simple linear addition; for $P < .05$: $TEa < 1.65 \times 0.5\ CV_A + 0.25\left(CV_I^2 + CV_G^2\right)^{1/2}$. It should be noted that this model for combining imprecision and bias is only one of the many models available. An EFLM Task and Finish Group has summarized the pros and cons of the concepts of allowable total error and measurement uncertainty.

The total error concept is widely used in current practice in laboratory medicine, albeit sometimes with a different multiplier for the imprecision term, to deliver different levels of probability. Application of this has become very widespread because the formula is simple, many data on biological variation exist, and they are directly related to the general use of results of examinations in laboratory medicine. However, concerns have been expressed. First, because the biological variation data currently used for some measurands vary, and second, because the APS derived using biological variation are unattainable for some measurands with current technology and methodology, including serum sodium, chloride, and calcium. However, some APS were easy to obtain for others, including serum urea, triglycerides, and many enzyme activities. A three-tier model to cater for this was proposed, giving minimum, desirable, and maximum APS based on biological variation using 0.25 CV_A, 0.50 CV_A, and 0.75 CV_A for imprecision and $0.125\times\left(CV_I^2+CV_G^2\right)^{1/2}$, $0.25\times\left(CV_I^2+CV_G^2\right)^{1/2}$, and $0.325\times\left(CV_I^2+CV_G^2\right)^{1/2}$ for bias, respectively. Furthermore, because analytical technology has evolved rapidly over recent time, and examination performance has improved, it has been suggested that a fourth "ideal" level should be introduced with multipliers of 0.1 for both imprecision and bias. Combinations of the three (or four) levels can be done to obtain relevant performance specifications for TEa.

Specifications Derived from Opinions of Clinicians

Because it is the users of the results of laboratory tests that apply the data provided, it is logical to think that they should set analytical quality requirements that facilitate their decision making. One way of attempting to do this is with clinical vignettes in which the clinician provides information about the change in serial results, which is required to stimulate a clinical decision. The principle behind the studies should be that a difference in serial results is due to preanalytical sources of variation (CV_P), CV_A, CV_I, and changes in bias (ΔB), using the transformation mentioned above from CV to SD:

$$\text{difference}=2^{1/2}\times Z\times\left[SD_P^2+SD_I^2+SD_A^2\right]^{1/2}+\Delta B$$

now, let SD_P and ΔB be zero and rearrange the equation:

$$SD_A=\left[\text{difference}\Big/\left(2^{1/2}\times Z+SD_I^2\right)\right]^{1/2}$$

Studies have been done using this model on a variety of measurands, and have included some studies that involved patients undertaking self-testing. An early example is provided by the studies on blood hemoglobin; other studies performed since then have investigated HbA_{1c}, glucose, prothrombin time, erythrocyte sedimentation rate, and urinary albumin. It was found that the derived APS depends on the clinical setting and local management pathway, the semantics of the questions posed, how close the result was to a decision limit or reference limit, and probably on perceptions related to current analytical performance. Not unexpectedly, there was large between-clinician variation in response. This approach is probably possible only for the assessment of a single measurand in a well-specified clinical situation. In addition, it has been suggested that this strategy is possible only for measurands that have a major role in monitoring and/or diagnosis and where more or less uniform recommendations guide clinical standards of practice.

Specifications for Measurement Uncertainty

Because there is now a requirement for medical laboratories to document the measurement uncertainty to comply with ISO 15189, there should be consideration of the setting of APS for this characteristic. Because the fundamental principles are that bias should be eliminated (if possible) and all sources of variation should be added linearly as variances, then the only possible performance specifications for measurement uncertainty are that $CV_A < f\times CV_I$, where $f = 0.10, 0.25, 0.50, 0.75$ for ideal, optimum, desirable, and minimum levels of performance, respectively, but, again, this is dependent on reliable data being available for CV_I.

Specifications for Other Applications

Data on biological variation have been used to generate performance specifications in other than general clinical settings, including an evaluation of systems, reference methods, and for the allowable difference in bias % (ΔB%) between two measurement systems or devices used to generate results on the same individuals, which is $\Delta B\% < 1/3CV_I$.

OTHER APPLICATIONS

Calculation of Reliability Coefficient

The individuality of tests and the consequences of the marked individuality (low II) of most measurands in laboratory diagnosis, case finding, screening, and monitoring have been discussed. Epidemiologists and others use similar information to the II, but in a slightly different way. The reliability coefficient is used, which is calculated as $CV_G^2\Big/\left(CV_A^2+CV_I^2+CV_G^2\right)^{1/2}$, which is the between-subject variance divided by the total variance. The reliability coefficient, usually called R, is numerically equal to the correlation coefficient of repeated measurements. It can be between 0 and 1. If R approached 0, II would be high, and if R approached 1, then II would be low. Most measurands have high values for R.

Number of Samples Needed

In usual clinical practice, only one sample is taken. Analytical result variation can be reduced by multiple sampling (or multiple analyses), and the variation is made smaller by the square root of the number of replicates. To estimate the number of specimens needed to determine the homeostatic set point within a certain percentage error with a stated probability, a

simple rearrangement of the usual standard error of the mean formula is used, namely:

$$n = \left(\frac{Z \times \sqrt{CV_A^2 + CV_I^2}}{D} \right)^2$$

where Z is the Z-score appropriate for the probability, and D is the desired percentage closeness to the homeostatic set point. It is important to note that taking multiple specimens and undertaking replicate analyses does affect the overall variability of the individual analytical result; the dispersion (expressed as 1 CV) can be calculated as:

$$\text{dispersion} = Z \times \left[\left(CV_A^2 / nA \right) + \left(CV_I^2 / nS \right) \right]^{1/2}$$

where Z is the number of standard deviations appropriate to the probability selected, nA is the number of replicate examinations, and nS is the number of specimens. The relative magnitudes of CV_A and CV_I are important in deciding if a lower dispersion is required, whether it is better to undertake replicate analyses on one specimen or singleton analysis on multiple specimens. Calculators for this can be found, along with a more detailed discussion of this topic, with real examples, at https://www.westgard.com/.

Reporting Results, Selecting the Best Specimen to Collect, and Choosing the Best Investigation

It is sometimes possible to report the results of investigations in different ways. For example, measurands in urine, such as creatinine, can be reported as concentration or output per day, and many others (e.g., electrolytes or proteins) are reported as a creatinine concentration ratio. Moreover, for some measurands, it is possible to collect different samples for the same clinical purpose (e.g., early morning or random or timed urine specimens for assessment of low concentration albumin and protein). In certain clinical situations, alternative investigations that might be considered to have a somewhat similar purpose are available, such as serum creatinine and cystatin-C, or blood HbA_{1c} and serum fructosamine. Knowledge of the components of biological variation can assist in making decisions about reporting results, selecting the best specimen to collect, and choosing the best test.

To undertake such comparisons, the influences of biological variation should be considered. The ideal measurand would have low CV_I so that a single analysis will give a good measure of the true value for that individual. Moreover, this would allow easy monitoring over time and detection of significant differences, because the RCV would be low, provided that the CV_A was also low. In addition, the ideal measurand would have no heterogeneity of CV_I among individuals and across studies, and would not be dependent on age and sex and other possible confounding factors so that the simple general formulae given in this chapter would hold for all.

Method Development and Evaluation

Introduction of new investigations is an ongoing task for most medical laboratories. Some years ago, Zweig and Robertson suggested that the introduction of a new procedure should be similar to the structured evolution of a new drug through phased trials. They suggested that the phases should be the following: analytical investigation and assessment of reliability and practicability characteristics; overlap investigation involving the generation of reference intervals and the assessment of values in disease; clinical investigation and evaluation of sensitivity, specificity, and predictive value; outcome investigation of whether individuals gain an advantage; and, finally, investigation of usefulness, which is a cost–benefit analysis with respect to individuals and the population. In contrast to the linear models, another approach has been proposed in which the essential components of analytical and clinical performance, clinical and cost effectiveness, and the broader impact of testing are assembled in a dynamic cycle. This approach emphasizes the interaction of the different components, and that clinical effectiveness data should be fed back to refine analytical and clinical performances to achieve improved outcomes.

One aspect that is not covered in any of these strategies, however, is the need to generate and apply data on biological variation early in the evolution of any investigation. Data on biological variation allow the setting of APS, calculation of the significance of changes in an individual, and assessment of the use of the generated population-based reference intervals. Generation and application of data on biological variation are essential prerequisites for introducing new laboratory investigations. Moreover, the data are also necessary for objective analysis of the often somewhat subjective guidelines from professional bodies that give recommendations on interpretation of numerical results and on APS; unfortunately, these recommendations are often flawed.

OVERALL CONCLUSIONS

Numerical estimates of within-subject and between-subject biological variation are best generated by analysis of a series of specimens taken from a cohort of individuals, analyzed in duplicate, and followed by statistical analysis of the sources of variation. The proper design and performance of such studies is complex. However, biological variation data have numerous applications, including:

- The determination of the individuality of a measurand and the usefulness of conventional population-based reference intervals
- The statistical significance of changes in serial results from an individual
- APS for imprecision, bias, allowable total error, measurement uncertainty, and other characteristics

Currently, there are concerns regarding the robustness of certain of the published data on biological variation, but important initiatives are on their way to provide easily accessible robust and globally adoptable estimates. Recent recommendations should be followed in the generation, reporting, and application of biological variation data.

POINTS TO REMEMBER

Biological variation may be cyclical and daily, monthly, or seasonal, but most measurands have random variation around homeostatic set points.

The generation of high-quality data on within-subject and between-subject components of biological variation requires adherence to strict experimental and statistical protocols.

Biological variation estimates are available for many measurands, but their reliability should be questioned, and users should ensure the relevance of published data to their applications.

Data on biological variation are used, inter alia, to determine:

- The individuality of a measurand and the usefulness of conventional population-based reference intervals
- The statistical significance of changes in serial results in an individual and the probability that any change documented is significant
- APS for imprecision, bias, TEa, measurement uncertainty, and other characteristics

REVIEW QUESTIONS

1. Which of the following usually best describes the variation in the concentration or activity of most measurands in laboratory medicine?
   a. Circadian rhythms
   b. Monthly cycles
   c. Systematic trends
   d. Random variations
   e. Seasonal fluctuations
2. Which of the following strategies is recommended to best set general analytical performance specifications?
   a. Effect of analytical performance on clinical outcomes
   b. Fractions of biological variation component estimates
   c. Professional body recommendations
   d. The limits of acceptance used in external quality assessment or proficiency testing
   e. The state of the art
3. Which of the following is irrelevant to the creation of Reference Change Values?
   a. Within-subject biological variation
   b. Analytical bias
   c. Analytical imprecision
   d. Preanalytical sources of variation
   e. Between-subject biological variation
4. What analytical imprecision should be used when the reference change value is to be created for a measurand that will be used by your own laboratory?
   a. The analytical imprecision obtained from duplicate analyses obtained during the study of the biological variation
   b. The best analytical imprecision you can obtain in your laboratory
   c. The analytical imprecision obtained in your laboratory during same time period as the time interval between the samples examined
   d. The analytical imprecision as assessed from an external quality assessment scheme
   e. The analytical imprecision documented in the manufacturer's literature
5. Which of the following represents the individuality of most measurands in laboratory medicine?
   a. The index of individuality is high
   b. The index of individuality is low
   c. The measurand has low individuality
   d. The within-subject variation is larger than the between-subject variation
   e. The analytical imprecision is lower than the within-subject variation

SUGGESTED READINGS

Aarsand, A. K., Røraas, T., Bartlett, W. A., Coşkun, A., Carobene, A., Fernandez-Calle, P., et al. (2018). Harmonization initiatives in the generation, reporting and application of biological variation data. *Clinical Chemistry and Laboratory Medicine, pii:/j/cclm.ahead-of-print/cclm-2018-0058/cclm-2018-0058.xml.* https://doi.org/10.1515/cclm-2018-0058. [Epub ahead of print].

Aarsand, A. K., Røraas, T., Fernandez-Calle, P., Ricos, C., Díaz-Garzón, J., Jonker, N., et al. (2018). The biological variation data critical appraisal checklist: A standard for evaluating studies on biological variation. *Clinical Chemistry*, *64*, 501–514.

Bartlett, W. A., Braga, F., Carobene, A., et al. (2015). A checklist for critical appraisal of studies of biological variation. *Clinical Chemistry and Laboratory Medicine*, *53*, 879–885.

Carobene, A., Røraas, T., Sølvik, U. Ø., Sylte, M. S., Sandberg, S., Guerra, E., et al. (2017). European Biological Variation Study of the EFLM Working Group on Biological Variation. Biological variation estimates obtained from 91 healthy study participants for 9 enzymes in serum. *Clinical Chemistry*, *63*, 1141–1150.

Carobene, A., Strollo, M., Jonker, N., Barla, G., Bartlett, W. A., Sandberg, S., et al. (2016). Sample collections from healthy volunteers for biological variation estimates' update: a new project undertaken by the Working Group on Biological Variation established by the European Federation of Clinical Chemistry and Laboratory Medicine. *Clinical Chemistry and Laboratory Medicine*, *54*, 1599–608.

Fraser, C. G. (2001). *Biological variation: from principles to practice*. Washington, DC: AACC Press.

Fraser, C. G., & Harris, E. K. (1989). Generation and application of data on biological variation in clinical chemistry. *Critical Reviews in Clinical Laboratory Sciences*, *27*, 409–437.

Fraser, C. G. (2011). Reference change values. *Clinical Chemistry and Laboratory Medicine*, *50*, 807–812.

Fraser, C. G. (2004). Test result variation and the quality of evidence-based clinical guidelines. *Clinica Chimica Acta*, *346*, 19–24.

Fraser, C. G. (1994). Data on biological variation; essential prerequisites for introducing new procedures. *Clinical Chemistry*, *40*, 1671–1673.

Fraser, C. G. (2004). Inherent biological variation and reference values. *Clinical Chemistry and Laboratory Medicine*, *42*, 758–764.

Fraser, C. G., Hyltoft Petersen, P., Libeer, J. C., et al. (1997). Proposals for setting generally applicable quality goals solely based on biology. *Annals of Clinical Biochemistry*, *34*, 8–12.

Fraser, C. G., & Petersen, P. H. (1999). Analytical performance characteristics should be judged against objective quality specifications. *Clinical Chemistry*, *45*, 321–323.

Hyltoft Petersen, P., Sandberg, S., Iglesias, N., et al. (2007). Likelihood-ratio and odds applied to monitoring of patients as a supplement to reference change value (RCV). *Clinical Chemistry and Laboratory Medicine*, *46*, 157–164.

Hyltoft Petersen, P., Fraser, C. G., Kallner, A., et al. (Eds.). (1999). Strategies to set global analytical quality specifications in laboratory medicine. *Scandinavian Journal of Clinical and Laboratory Investigation, 59, 475–585*.

Klee, G. G. (2010). Establishment of outcome-related analytic performance goals. *Clinical Chemistry*, *56*, 714–722.

Minchinela, J., Ricós, C., Perich, C., et al. (2014). Biological variation database and quality specifications for imprecision, bias and total error (desirable and minimum). The 2014 update http://www.westgard.com/biodatabase-2014-update.htm.

Perich, C., Minchinela, J., Ricós, C., et al. (2015). Biological variation database: structure and criteria used for generation and update. *Clinical Chemistry and Laboratory Medicine*, *53*, 299–305.

Petersen, P. H., Sandberg, S., Fraser, C. G., et al. (2001). Influence of index of individuality on false positives in repeated sampling from healthy individuals. *Clinical Chemistry and Laboratory Medicine*, *39*, 160–165.

Ricos, C., Iglesias, N., Garcia-Lario, J. V., et al. (2007). Within-subject biological variation in disease: collated data and clinical consequences. *Annals of Clinical Biochemistry*, *44*, 343–352.

Røraas, T., Petersen, P. H., & Sandberg, S. (2012). Confidence intervals and power calculations for within-person biological variation: effect of analytical imprecision, number of replicates, number of samples, and number of individuals. *Clinical Chemistry*, *58*, 1306–1313.

Sandberg, S., Fraser, C. G., Horvath, A. R., et al. (2015). Defining analytical performance specifications: consensus statement from the 1st Strategic Conference of the European Federation of Clinical Chemistry and Laboratory Medicine. *Clinical Chemistry and Laboratory Medicine*, 833–835.

5

Establishment and Use of Reference Values

*Gary L. Horowitz, Graham R.D. Jones**

OBJECTIVES

1. Define the following:
   - Clinical sensitivity
   - Clinical specificity
   - Exclusion criteria
   - Interpercentile interval
   - Outlier
   - Partitioning; partitioning criteria
   - Population-based reference value
   - Predictive value
   - Prevalence
   - Random sample
   - Range
   - Reference individual
   - Reference interval
   - Reference limits
   - Reference population
   - Reference value
   - Selection criteria
   - Subject-based reference value
   - Transferability or transference
2. List three conditions that are essential when a valid comparison of individual laboratory results with reference values is performed; state the need for establishing reference intervals.
3. Give three examples of exclusion criteria used in the production of health-associated reference values; give three examples of partitioning criteria used to subgroup a reference group.
4. State why standardization of specimen collection is important when reference values are established.
5. Compare the terms *reference value* and *reference interval;* list three categories of reference intervals.
6. Briefly state the parametric and nonparametric statistical methods of determining an interpercentile interval; state the important assumption that must be made when parametric statistics are used.
7. State the limitation of using population-based reference intervals instead of subject-based reference intervals, and state a solution to this limitation.
8. Discuss the issue of transferability of reference values with regard to prerequisites and solutions to the issue.
9. State the formulas used in calculating clinical sensitivity, clinical specificity, and predictive value of a laboratory test; given appropriate values, calculate clinical sensitivity, clinical specificity, and predictive value for a laboratory test.
10. State how the predictive value of a laboratory test is affected by prevalence.

KEY WORDS AND DEFINITIONS

Clinical sensitivity The proportion of subjects with disease who have positive test results.

Clinical specificity The proportion of subjects without disease who have negative test results.

Nonparametric analysis A statistical approach to reference value analysis that requires no assumptions about the nature of the distribution; thus it can be applied to distributions that are Gaussian or non-Gaussian.

Parametric analysis A statistical approach to reference value analysis that requires specific distributional assumptions. For example, it usually requires that the distribution of values be Gaussian (or that the values be mathematically manipulated so that they become Gaussian).

Partitioning The use of specific criteria in the subclassification of reference groups to reduce the biologic variation in each group; the most commonly used criteria are age and sex.

Predictive value The predictive value of a positive laboratory test is the number of true-positive results divided by the total number of positive results (true-positives plus false-positives); the negative predictive value is the number of true-negative results divided by the total number of negative results (true-negatives plus false-negatives).

* The authors gratefully acknowledge the original contribution by Helge Eric Solberg on which major portions of this chapter are based.

Prevalence The proportion of subjects in a specified population who have a specified disease or condition.

Reference individual An individual selected as the basis for comparison with individuals under clinical investigation through the use of defined criteria.

Reference interval (population based) A set of values usually defined by an upper reference limit and a lower reference limit, representing a specified proportion of the reference population; this is frequently the central 95% of values from the reference population.

Reference interval (subject based) A set of values usually defined by an upper reference limit and a lower reference limit, representing a specified proportion of the values from a reference individual; this is frequently the central 95% of values from the reference individual.

Reference population An undefined number of individuals that represent the demographic for which the reference intervals will be used. Reference individuals are chosen, preferably at random, from this larger population to provide reference samples for the establishment of a reference interval.

Reference value A value obtained by observation or measurement of a particular type of quantity on a reference individual; results of a certain type of quantity obtained from a single individual or group of individuals corresponding to a stated description.

Selection criteria A set of criteria that define the desired characteristics of a reference individual. The specific criteria chosen will depend on the purpose of the reference interval (RI) and the specific population the RI is intended to represent.

Transferability or Transference The adoption of reference intervals by a laboratory that were previously established elsewhere. Procedures for validation of reference intervals must be completed by the adopting laboratory prior to the use of the transferred RI to ensure that they are appropriate to the laboratory's patient population and laboratory methods.

In practice, data collected during (1) medical interviews, (2) clinical examinations, and (3) supplementary investigations are interpreted by comparison with reference data. If the condition of the patient resembles that typical of a particular disease, the clinician or health care provider may base the diagnosis on the observation (positive diagnosis). This diagnosis is made more likely if observed symptoms and signs do not fit the patterns that characterize a set of alternative diseases (diagnosis by exclusion).

Interpretation of medical laboratory data is an example of decision making by comparison. We therefore need *reference values* for all tests performed in the clinical laboratory, not only from healthy individuals but also from patients with relevant diseases. Ideally, observed values should be related to several collections of reference values, such as values from (1) healthy people, (2) undifferentiated hospital population, (3) people with typical diseases, or (4) ambulatory individuals, and to previous values from the same subjects. A patient's laboratory result is simply not medically useful if appropriate data for comparison are lacking. Establishment and use of such reference values are the topics of this chapter.

ESTABLISHMENT OF REFERENCE VALUES

Certain conditions are mandatory for making the comparison of a patient's laboratory results with reference values possible and valid:

1. All groups of reference individuals should be clearly defined.
2. The patient examined should resemble sufficiently the reference individuals (in all groups selected for comparison) in all respects other than those under investigation.
3. The conditions under which the samples were obtained and processed for analysis should be known.
4. All quantities compared should be of the same type.
5. All laboratory results should be produced with the use of adequately standardized methods under sufficient analytical quality control (see Chapter 7).
6. The clinical sensitivity, clinical specificity, and prevalence in the populations tested should be known so that laboratory tests can be interpreted intelligently (see Chapter 2).

Background

The term *normal values* has been used frequently in the past. Confusion arose because the word *normal* has several very different connotations. Consequently, this term is now considered obsolete and **should not be used.** Instead, the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) recommends use of the term *reference values* and related terms, such as *reference individual, reference limit, reference interval,* and *observed values*. **Reference values** are results of a certain type of quantity obtained from a single individual or group of individuals corresponding to a stated description, which must be spelled out and made available for use by others.

A short description of qualifiers associated with the term *reference values,* such as *health-associated reference values* (close to what was understood by the obsolete term *normal values*), is convenient. Other examples of such qualifying words are (1) *diabetic patient*, (2) *hospitalized diabetic patient*, and (3) *ambulatory diabetic patient*. These short descriptions prevent the common misunderstanding that reference values are associated only with health.

A further distinction is made between subject-based and population-based reference values. *Subject-based reference values* are previous values from the same individual, obtained when the individual was in a defined state of health.

Population-based reference values are those obtained from a group of systematically defined reference individuals and are usually the values referred to when the term *reference values* is used with no qualifying words.

Selection of Reference Individuals

A set of explicit criteria should be used to determine which individuals should be included in the group of **reference individuals**. Such criteria include (1) statements describing the source population, (2) specifications of criteria for health, and (3) the disease of interest. The selection of reference individuals is based essentially on the application of these defined criteria to the entire group of examined candidates. The required characteristics of the reference values determine which criteria should be used in the selection process. As examples, the left-hand column of Table 5.1 provides some examples of criteria that should be considered when *excluding* individuals in the production of health-associated reference values.

Ideally the group of reference individuals should be a *random sample* of all individuals in the parent population who meet the **selection criteria**. However, a strictly random sampling scheme is impossible to obtain in most situations for a variety of practical reasons. For example, it would imply the examination and application of selection criteria to the entire population (thousands or millions of individuals) and the random selection of a subset of individuals among those accepted. Therefore using the best reference sample obtained after all practical considerations have been taken into account is necessary. Data then should be used and interpreted with due caution because of the possible bias introduced by the nonrandomness of the sample selection process.

Often, separate reference values for sex, age group, and other criteria are necessary. Thus it is important to define the *partitioning* criteria for the subclassification of the set of selected reference individuals. Some examples are provided in the right-hand column of Table 5.1. It should be noted that some criteria can be used as either exclusion or partitioning criteria, depending on the nature of the study. In practice, each partition could require as many as 120 samples; therefore the number of partitions should usually be kept as small as possible to obtain sufficient sample sizes for the derivation of valid statistical estimates.

Age and sex are the most frequently used criteria for partitioning because several analytes vary significantly among different age and gender groups. Age may be categorized by equal intervals (e.g., by decades) or by intervals that are narrower in the periods of life where greater variation is observed. In addition, the use of qualitative age groups (e.g., [1] postnatal, [2] infancy, [3] childhood, [4] prepubertal, [5] pubertal, [6] adult, [7] premenopausal, [8] menopausal, and [9] geriatric) often may be appropriate. Height and weight also have been used as criteria for the categorization of children.

Specimen Collection

Preanalytical standardization of (1) preparation of individuals before sample collection, (2) the sample collection itself, and (3) handling of the sample before analysis may eliminate or minimize bias or variation from these factors. These steps may reduce biological "noise" that otherwise may conceal important biological "signals" of disease, risk, or treatment effect.

TABLE 5.1 Examples of Exclusion and Partitioning Criteria

Exclusion Criteria	Partitioning Criteria
Age	Age
Alcohol intake	Blood group
Blood donation (recent)	Circadian variation
Drug abuse	Ethnicity
Exercise intensity (recent)	Exercise intensity (recent)
Fasting vs. nonfasting	Fasting vs. nonfasting
Sex	Sex
Hospitalization (recent)	Menstrual cycle (by stage)
Hypertension	
Illness (recent)	
Lactation	
Obesity	Obesity
Occupation	Posture (when sampled)
Oral contraceptives	
Pregnancy	Pregnancy (by stage)
Prescription drugs	Prescription drugs
Recent transfusion	

The magnitudes of preanalytical sources of variation clearly are not equal for different analytes (see Chapter 3). Therefore one may argue that one should consider only those factors that cause unwanted variation in the biological quantity for which reference value production is intended. Body posture during sample collection is, for instance, highly relevant for the establishment of reference values for nondiffusible analytes, such as albumin in serum, but is irrelevant for diffusible ones, such as serum sodium.

However, performing separate studies to allow for different preanalytical conditions for each constituent is impractical. In addition, several constituents are typically analyzed in the same clinical specimens. For these reasons, standardized procedures are recommended for sample collection, taking into account the requirements that will enable all the constituents under study to be measured accurately.

A special problem is caused by drug ingestion before sample collection. A distinction may be made between indispensable and dispensable medications. The latter category of drugs always should be avoided for at least 2 days before specimen collection. The use of indispensable drugs, such as contraceptive pills or essential medication, may be a criterion for exclusion or partition.

Analytical Procedures and Quality Control

Essential components of the required definition of a set of reference values are specifications concerning (1) analysis method, including information on equipment, reagents, calibrators, types of raw data, and calculation method; (2) quality control (see Chapter 7); and (3) reliability criteria (see Chapter 2). Specifications should be carefully described so

that another investigator will be able to reproduce the study and evaluate comparability of the reference values with values obtained by the methods used for production of the patient's values in a routine laboratory. To ensure comparability between reference and observed values, the same analytical method should be used. Alternatively (or in addition), one can establish comparability of methods and populations by analyzing 20 samples from reference individuals and ensuring that no more than two values fall outside the proposed limits.

Statistical Treatment of Reference Values

After the analysis of the reference specimens is performed, the reference values are subjected to a statistical treatment, which includes (1) partitioning of the reference values into appropriate groups, (2) inspection of the distribution of each group, (3) identification of outliers, and (4) determination of reference limits.

Partitioning of Reference Values

The subset of reference individuals and the corresponding reference values may be partitioned according to sex, age, and other characteristics (see Table 5.1). **Partitioning** is also known as *stratification*, *categorization*, or *subgrouping*, and its results are called *partitions*, *strata*, *categories*, *classes*, *or subgroups*. Such partitioning gives rise to narrower and potentially more appropriate reference intervals. For example, creatinine reference intervals for adult males and adult females overlap but are different because of the typical differences in muscle mass; combining them into a single interval would obscure those differences. Various statistical criteria for partitioning have been suggested, and all feature the need to collect sufficient data to allow evaluation of the partitions separately and then, if appropriate, to combine them. For example, one may test for differences in means or in standard deviations (SDs) of the separate distributions. However, note that differences in means or differences in variability may be *statistically* significant and still too small *clinically* to justify replacing a single overall reference interval with several class-specific intervals. Harris and Boyd and Lahti and coworkers have developed other criteria for partitioning and statistical methods for this purpose.

In the following sections a homogeneous reference distribution is assumed to exist—either the complete sample distribution (if partitioning is unnecessary) or a subclass distribution after partitioning.

Inspection of Distribution

It is advisable to display the reference distribution graphically and subsequently to inspect it. A histogram, as shown in Fig. 5.1, is prepared manually or by a computer program. Examination of the histogram serves as a safeguard against the misapplication or misinterpretation of statistical methods, and it may provide valuable information about the data. The following characteristics should be sought in an examination of the distribution:

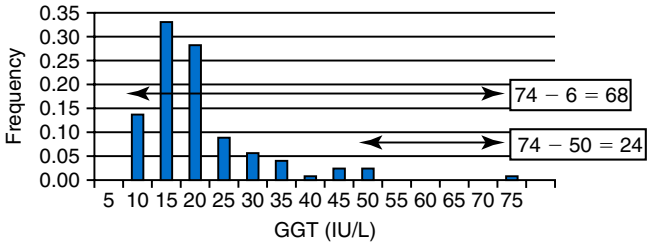


Fig. 5.1 Observed distribution of 124 γ-glutamyltransferase *(GGT)* values in serum *(IU/L)*. This distribution is clearly not Gaussian; it appears skewed to the right. The *upper arrow* indicates the range of observed values (highest − lowest, or 74 − 6 = 68); the *lower arrow* indicates the difference between the highest value and the next highest value (74 − 50 = 24). Because the quotient (24/68 = 0.35) exceeds 0.33, Dixon's range test indicates that the highest value is an outlier and therefore is omitted from all further analyses.

1. Highly deviating values (outliers) may represent erroneous values.
2. Bimodal or polymodal distributions have more than one peak and may indicate that the distribution is nonhomogeneous because of the mixing of two or more distributions. If nonhomogeneity is the case, the criteria used to select reference individuals should be reevaluated or partitioning of the values according to age, sex, or other relevant factors attempted.
3. The shape of the distribution may be asymmetrical (skewed) or more or less peaked than the symmetrical and bell-shaped Gaussian distribution (non-Gaussian kurtosis).
4. Visual inspection may provide initial estimates of the location of reference limits that are useful as checks on the validity of computations.

Identification and Handling of Outliers

An outlier is an erroneous value that deviates significantly from the proper reference values. Visual inspection of a histogram is a reliable method for identification of possible outliers. However, the inspector must keep in mind that values near the farthest point on the long tail of a skewed distribution may easily be misinterpreted as outliers. If the distribution is positively skewed, inspection of a histogram displaying the logarithms of the values may aid in the identification of outliers. Some outliers may be identified by statistical tests, but no single method will detect outliers in every situation that may occur. Two main problems are often encountered:

1. Many tests assume that the type of the true distribution is known before the tests are used. Some tests specifically require that the distribution be Gaussian. However, biologic distributions are very often non-Gaussian, and their types seldom are known in advance. The Dixon-Reed range test, described in IFCC's recommendation, is relatively robust and involves identification of the extreme value as an outlier if the difference between the two highest (or lowest) values in the distribution exceeds one-third of the range of all values (see Fig. 5.1).
2. Several tests for outliers assume that the data contain only a single outlier. Thus the range test may fail in the presence of several outliers.

A method published in 2005 may provide a solution to both of these problems. The algorithm involves mathematically transforming the data so as to approximate a Gaussian distribution, calculating the range of the central 50% of the resulting distribution, and then subtracting 150% of this value from the 25th percentile and adding 150% of this value to the 75th percentile. Any values beyond these limits are considered outliers.

Deviating values identified as possible outliers should not be discarded automatically. Values should be included or excluded on a rational basis. The records of the suspect values should be checked and any errors corrected. In some cases, deviating values should be rejected because noncorrectable causes have been found, such as previously unrecognized conditions that qualify individuals for exclusion from the group of reference individuals.

Determination of Reference Limits

In clinical practice an observed patient's value usually is compared with the corresponding **reference interval**, which is bounded by a pair of reference limits. This interval, which may be defined in different ways, is a useful condensation of the information carried by the total set of reference values.

The terms *reference limits* and *clinical decision limits* should not be confused. Reference limits describe the reference distribution and provide information about the observed variation of values in the selected set of reference individuals. Thus comparison of new values with these limits conveys only information about similarity to the given set of reference values. In contrast, clinical decision limits provide optimal separation among clinical categories. Such limits usually are based on analysis of reference values from several groups of individuals (e.g., healthy individuals, patients with relevant diseases) and thus are used for differential diagnosis. Alternatively, such values are established scientifically on the basis of outcome studies and guide decisions on treatment. For an example of decision limits currently in widespread use, the reader is referred to American College of Cardiology/American Heart Association guidelines for cholesterol.

As discussed earlier, the term *reference range* has been used for the term *reference interval,* but this use should be discouraged because the statistical term *range* denotes the difference (a single value!) between maximum and minimum values in a distribution.

Categories of reference intervals include (1) tolerance interval, (2) prediction interval, and (3) interpercentile interval. The choice from among these three may be important for certain systematically defined statistical problems, but, in practice, their numerical differences are negligible when based on at least 100 reference values.

The interpercentile interval is (1) simple to estimate, (2) more commonly used, and (3) recommended by the IFCC. It is defined as an interval bounded by two percentiles of the reference distribution. A percentile denotes a value that divides the reference distribution such that a specified percentage of its values has magnitudes less than or equal to the limiting value. For example, if 47 IU/L is the 97.5 percentile of serum γ-glutamyltransferase (GGT) values, then 97.5% of the values are equal to or less than this value.

The definition of the reference interval as the central 95% interval bounded by the 2.5 and 97.5 percentiles is an arbitrary but common convention (i.e., 2.5% of the values are cut off in both tails of the reference distribution). Another size or an asymmetrical location of the reference interval may be more appropriate in particular cases.

Of importance is the degree of uncertainty associated with a given percentile as an estimate of a population value; the magnitude of this uncertainty depends on the size of the sample, which increases when the number of observations is low. If the assumption of random sampling is fulfilled, determination of the confidence interval of the percentile (i.e., the limits within which the true percentile is located with a specified degree of confidence) is possible. The 0.90 confidence interval of the 97.5 percentile (upper reference limit) for serum GGT may, for example, be 39 to 50 IU/L. The true percentile would be expected in this interval with a confidence limit of 0.90 if all serum GGT concentrations in the total **reference population** were measured. The theoretical minimum sample size required for estimation of the 2.5 and 97.5 percentiles is 40 values, but usually at least 120 reference values are required to obtain reliable estimates.

The interpercentile interval has been determined by both parametric and nonparametric statistical techniques. The **parametric** method for the determination of percentiles and their confidence intervals assumes a certain type of distribution, and it is based on estimates of population parameters, such as the mean and SD. For example, a parametric method is used if the true distribution is believed to be Gaussian and reference limits (percentiles) are determined as the values located 2 SDs less than and greater than the mean. Most parametric methods in fact are based on the Gaussian distribution. If the reference distribution has another shape, mathematical functions that transform data to approximately Gaussian shape may be used. In contrast, the **nonparametric** method makes no assumptions concerning the type of distribution and does not use estimates of distribution parameters. The percentiles are determined simply by cutting off the required percentage of values in each tail of the subset reference distribution.

When the results obtained by these two methods are compared, the estimates of the percentiles usually are very similar. In general, the simple and reliable nonparametric method is preferable to the parametric method.

POINTS TO REMEMBER

The preferred method to establish reference intervals is nonparametric analysis, which

- Requires 120 reference values per partition
- Can be applied to any distribution of values without transformation; specifically, does not require a Gaussian distribution, as most other methods do
- Involves simple mathematical skills
- Is the method endorsed by CLSI and IFCC

TABLE 5.2 Nonparametric Confidence Intervals of Reference Limits[a]

Sample Size	RANK NUMBERS Lower	Upper
119–132	1	7
133–160	1	8
161–187	1	9
188–189	2	9
190–218	2	10
219–248	2	11
249–249	2	12
250–279	3	12
280–307	3	13
308–309	4	13
310–340	4	14
341–363	4	15
364–372	5	15
373–403	5	16
404–417	5	17
418–435	6	17
436–468	6	18
469–470	6	19
471–500	7	19

[a]The table shows the rank numbers of the 0.90 confidence interval of the 2.5 percentile for samples with 119 to 500 values. To obtain the corresponding rank numbers of the 97.5 percentile, subtract the rank numbers in the table from (n = 1), where n is the sample size.
From International Federation of Clinical Chemistry and Laboratory Medicine (IFCC), Expert Panel on Theory of Reference Values. Approved recommendation on the theory of reference values.

Nonparametric method. Several nonparametric methods are available, but those based on ranked data are simple and reliable and allow nonparametric estimation of the confidence intervals of the percentiles.

The steps in a nonparametric procedure are as follows:

1. Sort the n reference values in ascending order of magnitude, and rank the values. The minimum value has rank number 1, the next value number 2, and so on, until the maximum value, rank n, is reached. Consecutive rank numbers should be given to two or more values that are equal ("ties"). Spreadsheet software such as EXCEL is often used to sort and rank this type of data.
2. Compute the rank numbers of the 2.5 and 97.5 percentiles as $0.025(n + 1)$ and $0.975(n + 1)$, respectively.
3. Determine the percentiles by finding the original reference values that correspond to the computed rank numbers, provided that the rank numbers are integers. Otherwise, interpolation between the two limiting values is necessary.
4. Finally, determine the confidence interval of each percentile by using the binomial distribution. Table 5.2 facilitates this step for the 0.90 confidence interval of 2.5 and 97.5 percentiles. The bounding rank numbers for each percentile may be located in the table.

Tables 5.3A and 5.3B show an example of the nonparametric determination of percentiles using the serum GGT values shown in Fig. 5.1.

TABLE 5.3A γ-Glutamyltransferase Values Used in the Nonparametric Determination of Reference Intervals

GGT Value (IU/L)	Frequency	Rank Order
6	1	1
7	2	2, 3
8	6	4–9
9	4	10–13
10	4	14–17
11	9	18–26
12	7	27–33
13	7	34–40
14	9	41–49
15	9	50–58
16	8	59–66
17	11	67–77
18	8	78–85
19	5	86–90
20	3	91–93
21	2	94, 95
22	2	96, 97
23	2	98, 99
24	2	100, 101
25	3	102–104
26	2	105, 106
27	1	107
28	1	108
29	2	109, 110
30	1	111
32	2	112, 113
34	2	114, 115
35	1	116
39	1	117
42	2	118, 119
45	1	120
47	1	121
48	1	122
50	1	123

GGT, γ-Glutamyltransferase.

Parametric method. The parametric method is more complicated than the nonparametric method and usually requires the use of a computer statistics program when large data sets are processed. The parametric method assumes that the true distribution is Gaussian. Therefore it is absolutely critical to test the goodness-of-fit level of the reference distribution to a hypothetical Gaussian distribution. A simple test is examination of the cumulative frequency plotted on Gaussian probability paper (Fig. 5.2B); the plot should be close to a straight line if the distribution is Gaussian. In addition, many statistical computer programs have goodness-of-fit tests (e.g., tests based on coefficients of skewness and kurtosis, the Kolmogorov-Smirnov test, the Anderson-Darling test).

If the reference distribution does not differ significantly from the Gaussian distribution, the 2.5 and 97.5 percentiles are estimated by values approximately 2 SDs on each side of the mean or more precisely:

$$2.5 \text{ Percentile } \bar{x} - 1.96 \times \text{SD}$$

TABLE 5.3B Nonparametric Determination of Reference Interval[a]

Calculation of Rank Numbers of Percentiles	
Lower	0.025 (123 + 1) = 3.1 (i.e., Rank #3)
Upper	0.975 (123 + 1) = 120.9 (i.e., Rank #121)
Original Values Corresponding to These Rank Numbers	
Lower limit (2.5 percentile)	7 IU/L
Upper limit (97.5 percentile)	47 IU/L
Rank Numbers and Values of the 0.90 Confidence Limits	
Lower Reference Limits	
Rank numbers (see Table 5.2)	#1 and #7
Values	6 and 8 IU/L
Upper Reference Limits	
Rank numbers (see Table 5.2)	(123 + 1) − 7 = #117 and (123 + 1) − 1 = #123
Values	39 and 50 IU/L
Summary	
Lower reference limit	7 (6–8) IU/L
Upper reference limit	47 (39–50) IU/L

[a]The table shows an example using the γ-glutamyltransferase (GGT) results listed in Table 5.3A.

$$97.5 \text{ Percentile } \bar{x} + 1.96 \times \text{SD}$$

The 0.90 confidence interval of each percentile is estimated by the following two limits:

$$\text{Lower confidence limit} = \text{percentile limit} - 2.81 \times \frac{\text{SD}}{\sqrt{n}}$$

$$\text{Upper confidence limit} = \text{percentile limit} + 2.81 \times \frac{\text{SD}}{\sqrt{n}}$$

If the reference distribution is non-Gaussian, mathematical transformation of data may adjust the shape to approximate the Gaussian distribution. One frequent observation of interest is that logarithmically transformed values of a distribution with a long right tail (positively skewed) fit the Gaussian distribution rather closely (see Fig. 5.2D). In other cases, square roots of the values better approximate the Gaussian distribution. This information provides the basis for the common use of logarithmic and square root transformations when reference limits are estimated, as described in the following section. If these two functions fail to transform data to fit a Gaussian distribution, more general transformations can be used. Such functions are described in other relevant literature.

To apply the parametric procedure to transformed data, the process is very similar, as shown in the following steps:

1. Transform data with the logarithmic function $y = \log_{10}(x)$ (or $y = \ln[x]$) or by using square roots: $y = \sqrt{x}$. Then test the fit to the Gaussian distribution by using the methods described previously. If both transformations fail, then more general functions, which usually are more complicated, or the simple nonparametric method previously described should be used.
2. Then compute the *mean* ($\bar{y}$) and the standard deviation (SD_y) of the transformed data.
3. Next, estimate the percentiles and their confidence intervals in the transformed scale by using the formulas presented previously, substituting $\bar{y}$ for $\bar{x}$ and SD_y for SD.

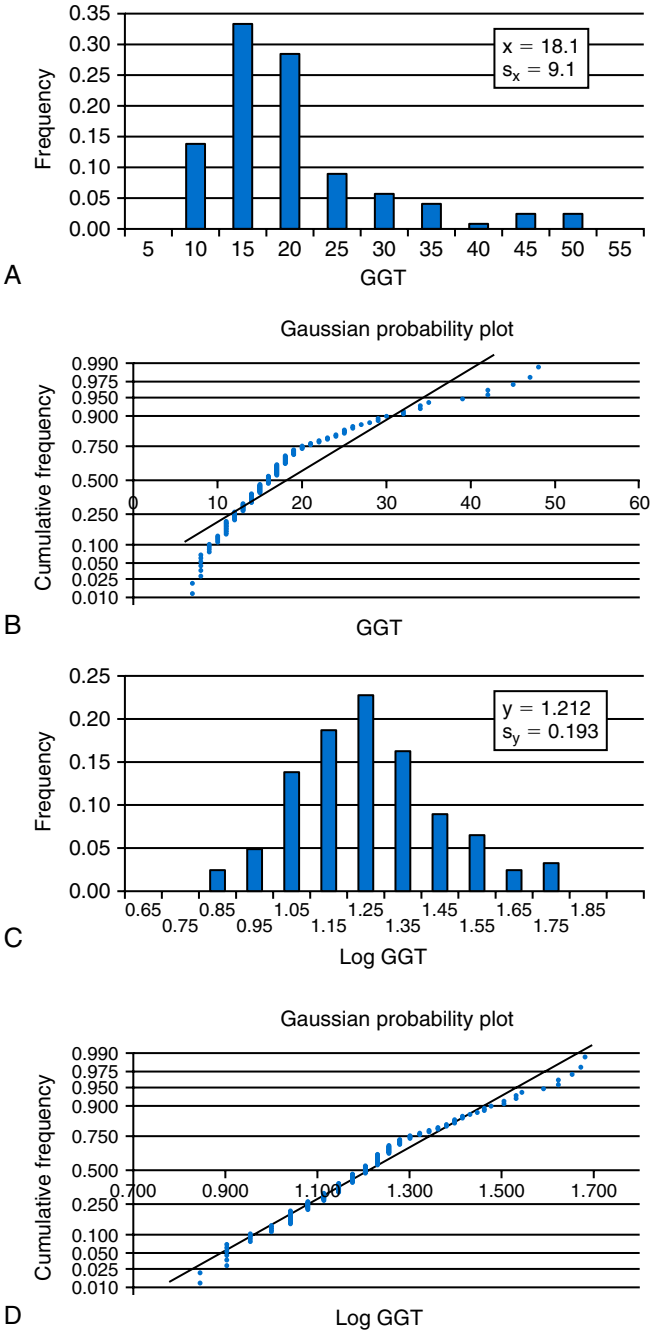


Fig. 5.2 Distribution of 123 remaining γ-glutamyltransferase (*GGT*) values from reference subjects. (A) A histogram of the original, untransformed data. (B) The cumulative frequency of the data from (A) plotted on Gaussian probability paper. (C) A histogram of the logarithmic transformed data. (D) The cumulative frequency of the data from (C), plotted on Gaussian probability paper.

TABLE 5.4 Summary of Glutamyltransferase Reference Interval Determination by Three Methods

Method	Lower Limit (Confidence Interval)	Upper Limit (Confidence Interval)	Values Below Lower Limit	Values Above Upper Limit
Nonparametric	7 (6–8)	47 (39–50)	1	2
Parametric-untransformed data	0 (−2 to 2)	36 (34–38)	0	7
Parametric-transformed data	7 (6–8)	40 (35–44)	1	6

The table summarizes the 95% reference intervals and associated 90% confidence limits generated by each of three methods for the same data set. The numbers of observed values deemed lower and higher than the corresponding interval for each method are given in the last two columns. Because the original data are positively skewed, note that the parametric techniques generate intervals that are biased low. Note too that the parametric technique on untransformed data has a lower confidence interval, which is actually less than 0.

4. Finally, reconvert the percentiles and their confidence intervals to the original data scale by using inverse functions—antilogarithms or squares, respectively.

As shown in Fig. 5.2D, the mean and the SD of the serum GGT values in Fig. 5.1 after logarithmic ($\log_{10}$) transformation are $\bar{y}$ = 1.212 and SD_y = 0.193. The 2.5 percentile is:

$$1.212 - 1.96 \times 0.193 = 0.835$$

$$2.5 \text{ percentile} = 10^{0.835} = 6.84$$

The lower reference limit of serum GGT thus is 7 IU/L. The 0.90 confidence interval of this percentile is:

$$1.212 - 2.81 \times \left(0.193/\sqrt{123}\right) = 0.786$$

$$\text{Lower confidence limit} = 10^{0.786} = 6.1$$

$$1.212 + 2.81 \times \left(0.193/\sqrt{123}\right) = 0.884$$

$$\text{Upper confidence limit} = 10^{0.884} = 7.7$$

that is, 6 to 8 IU/L. The 97.5 percentile (and its 0.90 confidence interval) is by the same method found to be 40 IU/L (35 to 44 IU/L).

Table 5.4 demonstrates that the nonparametric method and the parametric method (using the transformed data) result in very similar estimates of reference limits (percentiles).

Other methods for calculating reference limits. Other methods have been recommended for calculating reference limits, including the so-called bootstrap and robust methods. Neither of these methods makes assumptions about the underlying distribution; it needs not be Gaussian. Both require the use of computer software because they involve numerous iterations and somewhat complicated calculations. A brief discussion of these two methods is available elsewhere.

USE OF REFERENCE VALUES

In practice, interpreting medical laboratory data requires comparison of the patient's values with reference values.

Presentation of an Observed Value in Relation to Reference Values

An observed value (patient's value) may be compared with reference values. This comparison is often similar to hypothesis testing, but it is seldom statistical testing in the strict sense. Thus it is advisable to consider the reference values as the yardstick for a less formal assessment than hypothesis testing.

The clinician or health care provider should be supplied with as much information about the reference values as necessary for the interpretation. Reference intervals for all laboratory tests may be presented to clinicians or health care providers in a booklet, together with information about the analysis methods and their imprecision, along with descriptions of the reference values. A convenient presentation of the observed value and the reference interval on the same report sheet can be helpful to the busy clinician or health care provider. For example, the reference intervals may be preprinted on report forms, or the computer system may select the appropriate age- and sex-specific reference interval from the database and print it next to the test result or in graphical form. In some countries, there are specific recommendations for the uniform formatting of laboratory reports.

An observed value may be classified as low, usual, or high, depending on its location in relation to the reference interval. On reports, a convenient practice is to flag unusual results (e.g., through use of the letters *L* and *H* for low and high, respectively).

Another popular method of classification is expressing the observed value by a mathematical distance measure. For example, the well-known SD unit, or normal equivalent deviation, is such a measure. It is calculated as the difference between the observed value and the mean of the reference values divided by their SD. However, this measure is unreliable if the distribution of values is skewed. Values beyond approximately 2 SD imply that the value is beyond the central 95% of the reference interval. Indeed, by using the SD

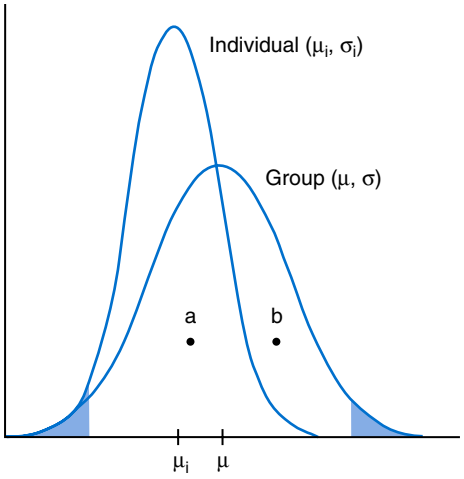


Fig. 5.3 The relationship between population-based and subject-based reference distributions and reference intervals. The example is hypothetical, and the two distributions are, for simplicity, Gaussian. Hypothetical means and standard deviations are μ and σ (for the population) and μ_i and σ_i (individual, i). (Modified from Harris, E. K. [1974]. Effects of intra- and interindividual variation on the appropriate use of normal ranges. *Clinical Chemistry, 20*, 1536.)

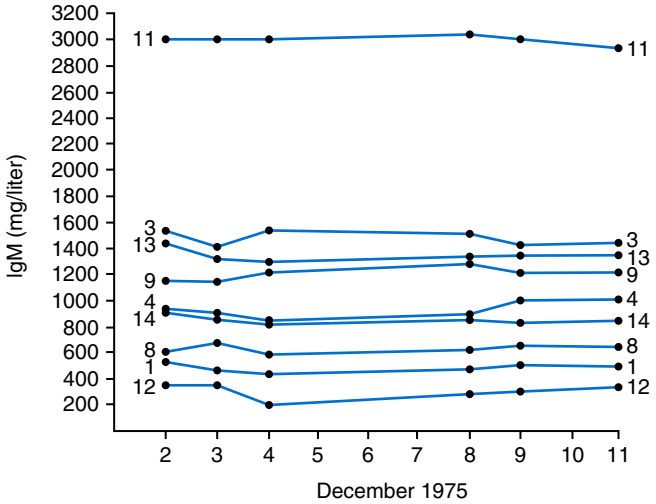


Fig. 5.4 Serial immunoglobulin M *(IgM)* values over several days from reference individuals. Note that intraindividual variability is very small compared with interindividual variability. (From Statland, B. E., Winkel, P., Killingsworth, L. M. [1976]. Factors contributing to intraindividual variation of serum constituents: 6. Physiological day-to-day variation in concentrations of 10 specific proteins in sera. *Clinical Chemistry, 22*, 1635–1636.)

unit deviation value, one determines the percentile of the observed value (e.g., values greater than 3.0 SD occur in only less than 0.15% of people in the reference distribution).

Subject-Based Reference Values

Fig. 5.3 illustrates the inherent problem associated with **population-based reference values**. The figure shows two hypothetical reference distributions. One represents the common reference distribution based on single samples obtained from a group of several different reference individuals. It has a true (hypothetical) mean μ and an SD of σ. The other distribution is based on several samples collected over time in a single individual, the *i*th individual. Its hypothetical mean is μ_i and the SD, σ_i.

If an observed value is located outside the subject's 2.5 and 97.5 percentiles (i.e., the personal or **subject-based reference interval**), the cause may be a change in biochemical status, suggesting the presence of disease. Fig. 5.3 demonstrates that such an observed value still may be within the population-based reference interval. The sensitivity of the latter interval to changes in a subject's biochemical status depends accordingly on the location of the individual's mean μ_i relative to the common mean μ and to the relative magnitudes of the corresponding SDs, σ_i and σ. A mean μ_i close to μ and a small σ_i relative to σ may conceal the individual's changes entirely within the population-based reference interval.

Two specific examples may help to clarify this concept. Fig. 5.4 depicts immunoglobulin M (IgM) values from several healthy individuals over the course of several days. As illustrated, intraindividual differences are small as compared with interindividual differences. Even though the population-based reference interval might extend from 200 to 1600 mg/L, in practice it would be unusual (abnormal) for any patient's IgM value to change by more than 200 mg/L, even if the value remained within the population-based reference interval. Similarly, it is well known that any given patient's serum creatinine value is reasonably constant, and this is related both to glomerular filtration rate (GFR) and to lean muscle mass. If the latter is constant, changes in GFR are inversely proportional to the serum creatinine (see Chapters 21 and 35), that is, even though a typical (population-based) reference interval for serum creatinine might extend from 62 to 106 µmol/L (0.7 to 1.2 mg/dL), a change from 65 to 105 µmol/L in a given patient would be distinctly abnormal, representing loss of almost half of the GFR.

At least one solution is known for the problem of the limitations of population-based reference intervals for certain tests. With this solution, the subject's previous values, obtained in a well-defined state of health, are used as the reference for any future value. Application of subject-based reference values becomes more feasible as "health screening" by laboratory tests and computer storage of results become available to large sections of the general population.

Transferability of Reference Values

Determining reliable reference values for each test in the laboratory's repertoire is a major task that is often far beyond the capabilities of the individual laboratory. Therefore it would be convenient if reference values generated in another laboratory could be used. This is especially important when ethical considerations limit the number of available individuals (e.g., when pediatric reference values are produced). Then cooperative establishment of reference values may be necessary.

A major prerequisite for transfer of reference values is that the populations must be comparable, and no major ethnic, social, or environmental differences must be noted between them. If they are not comparable, a separate reference interval study must be done.

Other factors that should not be overlooked include adherence to explicit, standardized protocols for (1) qualifying reference individuals, (2) preparing those individuals for specimen collection, and (3) performing specimen collection.

Analytical Issues

In practice, even if the populations are comparable and preanalytical standards are met, the problem of analytical **transferability** remains. The optimal, but usually very unrealistic, situation assumes that analytical methods, including their calibration and quality assurance, are identical in the laboratories. A more pragmatic approach involves (1) standardization of analytical protocols, (2) common calibration, (3) design of a sufficiently efficient external quality control scheme, and (4) the use of mathematical transfer functions if results still are not directly comparable.

Other Sources for Reference Limits

Sources for reference limits include manufacturers' package inserts, peer-reviewed publications, and multicenter trials. An advantage of using manufacturers' package inserts is that the analytical issues just mentioned are largely addressed. However, comparability of the patient populations and of other preanalytical issues must be ensured. For example, the creatine kinase (CK) upper reference limits cited in a package insert based on a Caucasian population underestimated, by several fold, the upper reference limits for blacks and Asians.

As to peer-reviewed publications, when a laboratory considers adopting such reference limits, both method and population comparability must be addressed, but these tasks may be considerably easier for the laboratory than performing its own reference interval study. For example, the CALIPER initiative established, using the techniques described earlier, pediatric reference intervals, partitioned by age, sex, and ethnicity, for 40 different assays. As the authors emphasized, the published reference limits were specific to the methods they used.

A third source for reference limits is becoming increasingly popular—multicenter trials. In contrast to the CALIPER initiative, in which all the analyses were performed by a single laboratory, these studies seek to pool data from many laboratories spread over large regions and potentially among different countries. Although this decreases the number of reference individuals needed to be recruited from each laboratory, it also typically increases the number of analytical methods involved for each assay. Despite global efforts at harmonization and standardization, these multicenter trials have repeatedly shown that current methods do not always produce interchangeable results, and therefore methods to ensure method comparability are still needed.

Verification of Transfer

Whether a laboratory adopts reference values from (1) a package insert, (2) another laboratory, or (3) a multicenter trial, it is important that the laboratory verifies the appropriateness of those values for its own use. This verification serves as the final check that the laboratory has implemented the analytical method correctly and that the laboratory's own population is comparable with that used for the original reference value study.

Comparison of a locally produced, small subset of values with the large set produced elsewhere using traditional statistical tests often is not appropriate because the underlying statistical assumptions are not fulfilled and the sample sizes are unbalanced. A reasonably practical alternative has been recommended by the Clinical and Laboratory Standards Institute (CLSI): With a sample size of 20 reference values, one verifies the appropriateness of a proposed reference interval as long as no more than two values are outside the proposed limits.

POINTS TO REMEMBER

Sources for reference limits (other than establishing one's own)

- Manufacturers' package inserts[a]
- Peer-reviewed literature[a]
- Multicenter trials[a]

[a] In all cases the use of these outside sources should be confirmed by a (short) verification study. See text.

Process of Selecting Reference Limits

Laboratories are responsible for, and usually take the lead in, providing reference limits for their test results. However, there are many potentially complicated decisions involved in the process, and the final selection of reference limits will influence the decisions of many physicians about their patients. As a result a multidisciplinary group should be involved in making decisions, such as whether to perform a reference interval study or to transfer intervals from another source; if the decision is made to do a local study, whether partitioning will be necessary (which will affect the number of subjects) and which preanalytical variables are relevant; and if limits will be transferred from another source, whether the methods and populations from that source are comparable.

Clinical Sensitivity, Specificity, and Predictive Value

When a clinician or a health care provider uses a laboratory test to help establish a diagnosis (as opposed to following a trend or evaluating the effectiveness of treatment), knowing the test's sensitivity and specificity can assist with proper interpretation (see Chapter 2). The **clinical sensitivity** of an assay is the fraction of those subjects with a specific disease that the assay correctly detects. The **clinical specificity** is the fraction of those individuals without the disease that the assay correctly detects. Fig. 5.5A illustrates pertinent definitions and formulas.

Changing the decision limit of an assay affects both clinical sensitivity and specificity. Consider the case when the disease group has higher assay values than the healthy group (Fig. 5.6). Values greater than the decision limit are classified as positive; those at or less than are negative. Moving the upper decision limit to a lower value increases the clinical sensitivity—but at the cost of decreased clinical specificity. Thus increased true-positive detection is traded for an increase in the number of false-positive results. This trade-off occurs in most tests performed in medicine.

Formulas

	Patients with positive test result	Patients with negative test result	
Patients with disease	True positives (TP)	False negatives (FN)	*Clinical sensitivity* $= \frac{TP}{TP + FN}$
Patients without disease	False positives (FP)	True negatives (TN)	*Clinical specificity* $= \frac{TN}{FP + TN}$
	Total positives $= TP + FP$	*Total negatives* $= FN + TN$	
	PV (positive test) $= \frac{TP}{TP + FP}$	*PV (negative test)* $= \frac{TN}{(FN + TN)}$	

A

Test X: Disease prevalence = 50%

	Patients with positive test result	Patients with negative test result	
100	95	5	*Sensitivity* $= 95/(95 + 5) = 95\%$
100	10	90	*Specificity* $= 90/(90 + 10) = 90\%$
	Positives $= 95 + 10 = 105$	*Negatives* $= 5 + 90 = 95$	
	PV (positive) $= 95/(95 + 10) = 90\%$	*PV (negative)* $= 90/(90 + 5) = 95\%$	

B

Test X: Disease prevalence = 5%

	Patients with positive test result	Patients with negative test result	
500	475	25	*Sensitivity* $= 95\%$
9500	950	8550	*Specificity* $= 90\%$
	Positives $= 475 + 950 = 1425$	*Negatives* $= 25 + 8550 = 8575$	
	PV (positive) $= 475/1425 = 5\%$	*PV (negative)* $= 8550/8575 = 100\%$	

C

Fig. 5.5 (A) A basic 2 × 2 table *(bold lines)* facilitates understanding of the concepts of sensitivity, specificity, and predictive value. In the left-hand column, all patients with positive test results are tabulated; in the right-hand column, all patients with negative test results are tabulated. In the top row, patients with the disease under study are divided by their test results; likewise, the bottom row *(shading)* divides people without the disease by their test results. The top left-hand corner then represents patients with disease who have positive results, true positives. The other three boxes are as labeled. As shown, the clinical sensitivity is calculated using the top row; the clinical specificity is calculated using the bottom row. The predictive value of a positive test is calculated using the data in the left-hand column; the predictive value of a negative test, using the right-hand column. (B) The calculations described in A are done on test X, whose sensitivity and specificity are 95% and 90%, respectively. For this example, the prevalence of the population tested is 50%, reflected in the fact that 100 people with disease and 100 people without disease are tested. Of note, this is frequently the prevalence used when tests are first described in the literature. As shown, the predictive value of a positive test is 90%. (C) The same calculations are done on the same test X, but the prevalence used is a more realistic, but still quite high, 5%, reflected in the fact that 500 people have the disease and 9500 do not. Although the sensitivity and specificity remain unchanged (95% and 90%, respectively), the predictive value of a positive test is now just 33%, that is, the likelihood that a patient with a positive test result in this population actually has the disease is 33%, or, in other words, 67% of the positive results are false positives.

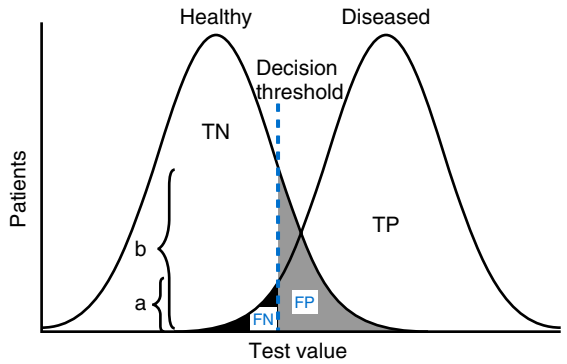


Fig. 5.6 Simulated distributions of healthy and diseased populations. Note that at the shown decision threshold, the probability of a subject with disease *(a)* is much less than the probability of a healthy subject *(b)*. *FN,* False negatives; *FP,* false positives; *TN,* true negatives; *TP,* true positives.

A separate, perhaps more important and more practical, issue that faces clinicians is this: Given a positive result, what is the likelihood that a patient actually has the disease? The **predictive value** of a positive test answers this question. The predictive value of a test combines disease prevalence with the test's sensitivity and specificity. **Prevalence** is the proportion of the population (or of those being tested) with the disease. The predictive value of a positive test is the number of true-positive results divided by the total number of positive results (true-positive and false-positive results combined). As reflected in Fig. 5.5B and C, the proportion of true-positive and false-positive results is a function of the prevalence in the population and of the sensitivity and specificity of the test in question.

As shown in Fig. 5.5A, the predictive value of a negative test follows in a similar way. It answers the question, "Given a negative result, how likely is it that a patient does not actually have the disease?" It has been used very effectively in many situations, for example, to reliably exclude, in the appropriate clinical context, deep venous thrombosis and/or pulmonary embolism.

When "Normal" May Appear to Be "Abnormal" and Vice Versa

Some, if not many, clinical laboratory test results are related to other clinical laboratory test results. For example, consider the relatively common situation of a patient with a low serum albumin. Because of the relationship between total calcium and albumin in serum (see Chapter 39), a total serum calcium concentration in the healthy reference interval might actually be pathologically high in this patient, and a total serum calcium concentration less than the lower reference limit might be healthy.

As another example, consider prostate-specific antigen (PSA) concentration in a post-prostatectomy patient or thyroglobulin in a post-thyroidectomy patient. In these cases, it would be "healthy" to have abnormally low (undetectable) concentrations of these measurands, and it would be distinctly abnormal to have "healthy" (detectable) levels. In these cases the problem is that the traditional reference population is not appropriate.

In other words, even when reference limit studies are done well, one needs to remember there are dependencies that can render those reference limits misleading.

REVIEW QUESTIONS

1. A "subject-based" reference value is:
   - **a.** obtained from a group of systematically defined reference individuals.
   - **b.** the type of value referred to when the term *reference value* is used with no qualifying words.
   - **c.** based on several samples collected over time in a single individual.
   - **d.** a random sample of all individuals in the parent population who fulfill the selection criteria.
2. True-negatives ÷ (false-positives + true-negatives) × 100 is the formula for determining:
   - **a.** clinical sensitivity.
   - **b.** clinical specificity.
   - **c.** predictive value.
   - **d.** reference value.
3. The proportion of a population that has the particular disease being studied for establishment of reference values is referred to as the:
   - **a.** prevalence.
   - **b.** predictive value.
   - **c.** positive value.
   - **d.** clinical sensitivity.
4. When reference values are established, the criteria used to determine which individuals should be included in the group of reference individuals are referred to as:
   - **a.** exclusion criteria.
   - **b.** inclusion criteria.
   - **c.** partition criteria.
   - **d.** selection criteria.
5. Calculate the predictive value of a test in which 220 tested individuals with positive test results actually have the disease and 45 tested individuals with positive test results do not have the disease.
   - **a.** 16.9%
   - **b.** 66%
   - **c.** 83%
   - **d.** 120%
6. An example of an exclusion criterion when establishing reference values would be a(n):
   - **a.** individual's age.
   - **b.** risk factor such as obesity.
   - **c.** individual's ethnic origin.
   - **d.** individual's sex.

7. In the selection of reference individuals, subclassifying a set of these individuals into homogeneous groups is referred to as:
   a. partitioning.
   b. excluding.
   c. transferring.
   d. including.
8. The number of true positive results divided by the sum of true positive results plus false positive results is referred to as:
   a. clinical sensitivity of a positive result.
   b. clinical specificity of a positive result.
   c. prevalence of the disease.
   d. predictive value of a positive result.
9. When reference limits are determined, the statistical method that assumes that the true distribution of reference values is a Gaussian (normal) distribution is the:
   a. nonparametric method.
   b. parametric method.
   c. interpercentile interval method.
   d. predictive value.

SUGGESTED READINGS

Brewster, L. M., Mairuhu, G., Sturk, A., et al. (2007). Distribution of creatine kinase in the general population: implications for statin therapy. *American Heart Journal*, *154*, 655–661.

Clinical and Laboratory Standards Institute. (2010). *Defining, establishing, and verifying reference intervals in the clinical laboratory. CLSI Document C28-A3c*. Wayne, Pa: Clinical and Laboratory Standards Institute.

Colantonio, D. A., Kyriakopoulou, L., Chan, M. K., et al. (2012). Closing the gaps in pediatric laboratory reference intervals: a CALIPER database of 40 biochemical markers in a healthy and multiethnic population of children. *Clinical Chemistry*, *58*, 854–868.

Ferré-Masferrer, M., Fuentes-Arderiu, X., Alvarez-Funes, V., Güell-Miró, R., & Castiñeiras-Lacambra, M. J. (1997). Multicentric reference values: shared reference limits. *European Journal of Clinical Chemistry and Clinical Biochemistry*, *35*, 715–718.

Fraser, C. G. (2001). *Biological variation: from principles to practice*. Washington, DC: AACC Press, 15–17.

Horowitz, G., & Jones, G. R. D. (2017). Establishment and use of reference values. In N. Rifai, A. R. Horvath, & C. T. Wittwer (Eds.). *Tietz textbook of clinical chemistry and molecular diagnostics* (6th ed.) (pp. 170–194). St Louis: Elsevier.

Harris, E. K., & Boyd, J. C. (1995). *Statistical bases of reference values in laboratory medicine*. New York: Marcel Dekker.

Horn, P. S., Feng, L., Li, Y., & Pesce, A. J. (2001). Effect of outliers and nonhealthy individuals on reference interval estimation. *Clinical Chemistry*, *47*, 2137–2145.

Horowitz, G., & Jones, G. R. D. (2017). Establishment and use of reference values. In N. Rifai, A. R. Horvath, & C. T. Wittwer (Eds.). *Tietz textbook of clinical chemistry and molecular diagnostics* (6th ed.) (pp. 170–194). St Louis: Elsevier.

International Federation of Clinical Chemistry and Laboratory Medicine (IFCC), Expert Panel on Theory of Reference Values. Approved recommendation on the theory of reference values.

Part 1. (1987). The concept of reference values. *Journal of Clinical Chemistry and Clinical Biochemistry*, *25*, 337–342.

Part 2. (1987). Selection of individuals for the production of reference values. *Journal of Clinical Chemistry and Clinical Biochemistry*, *25*, 639–644.

Part 3. (1988). Preparation of individuals and collection of specimens for the production of reference values. *Journal of Clinical Chemistry and Clinical Biochemistry*, *26*, 593–598.

Part 4. (1991). Control of analytical variation in the production, transfer, and application of reference values. *European Journal of Clinical Chemistry and Clinical Biochemistry*, *29*, 531–535.

Part 5. (1987). Statistical treatment of collected reference values: determination of reference limits. *Journal of Clinical Chemistry and Clinical Biochemistry*, *25*, 645–656.

Part 6. (1987). Presentation of observed values related to reference values. *Journal of Clinical Chemistry and Clinical Biochemistry*, *25*, 657–662.

Kearon, C., Ginsberg, J. S., Douketis, J., Turpie, A. G., Bates, S. M., Lee, A. Y., et al. (2006). An evaluation of D-dimer in the diagnosis of pulmonary embolism: a randomized trial. *Annals of Internal Medicine*, *144*, 812–821.

Lahti, A., Petersen, P. H., Boyd, J. C., Rustad, P., Laake, P., & Solberg, H. E. (2004). Partitioning of nongaussian-distributed biochemical reference data into subgroups. *Clinical Chemistry*, *50*, 891–900.

Rustad, P., & Felding, P. (Eds.). (2004). Transnational biological reference intervals: procedures and examples from the Nordic Reference Interval Project. *Scandinavian Journal of Clinical and Laboratory Investigation*, *64*, 265–441.

Rustad, P., Felding, P., Franzson, L., et al. (2004). The Nordic Reference Interval Project 2000: recommended reference intervals for 25 common biochemical properties. *Scandinavian Journal of Clinical and Laboratory Investigation*, *64*, 271–284.

Solberg, H. E. (2004). The IFCC recommendation on estimation of reference intervals. The RefVal Program. *Clinical Chemistry and Laboratory Medicine*, *42*, 710–714.

Solberg, H. E., & Grasbeck, R. (1989). Reference values. *Advances in Clinical Chemistry*, *27*, 1–79.

Solberg, H. E., & Lahti, A. (2005). Detection of outliers in reference distributions: performance of Horn's algorithm. *Clinical Chemistry*, *51*, 2326–2332.

Stone, N. J., Robinson, J. G., Lichtenstein, A. H., et al. (2014). 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults. *Journal of the American College of Cardiology*, *63*, 2889–2934.

Yamamoto, Y., Hosogaya, S., Osawa, S., et al. (2013). Nationwide multicenter study aimed at the establishment of common reference intervals for standardized clinical laboratory tests in Japan. *Clinical Chemistry and Laboratory Medicine*, 51, 1663–1672.

6

Specimen Collection, Processing, and Other Preanalytical Variables

Doris M. Haverstick, Patricia M. Jones

OBJECTIVES

1. Define the following terms:
   - Anticoagulant
   - Controllable preanalytical variable
   - Hemolysis
   - Order of draw
   - Phlebotomy
   - Plasma
   - Preanalytical variable
   - Preanalytical error
   - Serum
   - Uncontrollable preanalytical variable
   - Venipuncture
2. Give two examples of a preanalytical error and two examples of an uncontrollable preanalytical variable.
3. List the types of biologic specimens that are analyzed in a clinical laboratory.
4. Summarize the steps that are performed by a phlebotomist in obtaining a blood sample by venipuncture; state the preferred site of venous blood collection, including the practice used when an intravenous line is present.
5. List the general effects on analytes caused by the following:
   - Pumping a fist before venipuncture
   - Stress
   - Collection order (first tube, second tube, etc.)
   - Time of collection related to diurnal variation
   - Prolonged venous occlusion with a tourniquet
   - Not filling collection tube completely
6. Discuss order of draw for multiple blood specimens, including the order required for collecting multiple tubes of blood, color of stopper and associated additive, need for tube filling and inversion, and reason for filling tubes in a specific order.
7. Describe the skin puncture collection technique, including methods of stimulating blood flow; state the reasons for collecting a specimen using skin puncture; and describe the collection procedure for obtaining a blood spot for molecular genetic testing.
8. Explain the difference between serum and plasma.
9. Compare the difference in composition, if any, between serum and plasma specimens of the following analytes:
   - Calcium
   - Cholesterol
   - Albumin
   - Creatinine
   - Total protein
   - Glucose
   - Potassium
10. State how the following anticoagulants prevent blood from coagulating:
    - Heparin
    - EDTA
    - Citrate
    - Oxalate
    - Iodoacetate
11. State the appropriate analytical situations for using evacuated tubes containing various additives; state reasons why blood collected in the same anticoagulants cannot be used for certain analyses.
12. Describe the hemoglobin concentration at which hemolysis is observable in plasma; describe how hemolysis affects measurement of certain analytes and use of certain analytical procedures.
13. List three types of urine specimens and their use in clinical analysis; list two methods of urine preservation and discuss the use of each.
14. Outline the procedure for collecting a timed urine specimen.
15. List the clinical chemistry analyses performed on the following specimen types and the collection procedure name, if any:
    - Feces
    - Cerebrospinal fluid (CSF)
    - Synovial fluid
    - Amniotic fluid
    - Chorionic villus
    - Pleural, pericardial, and ascitic fluids
    - Saliva
    - Buccal cells
    - Hair and nails
16. Summarize each of the four aspects of specimen handling; discuss these in relation to specimen identification on various containers, centrifugation of blood samples, sample storage on ice/at −20°C/at 4°C, protection from light, and transport guidelines.

KEY WORDS AND DEFINITIONS

Additives Compounds added to biologic specimens to prevent them from clotting or to preserve the constituents of a specimen.

Anticoagulant Any substance that prevents blood from clotting.

Chorionic villus sampling A prenatal test to detect birth defects that is performed at an early stage of pregnancy and involves retrieval and examination of tissue from the chorionic villi. Also called *chorionic villus biopsy.*

Coagulation (clotting) The sequential process by which the multiple coagulation factors of blood interact in the coagulation cascade, resulting in formation of an insoluble fibrin clot.

Diurnal variation Variation that occurs in the amount of a substance during a 24-hour period.

Hemolysis Disruption of the red cell membrane causing release of hemoglobin and other components of red blood cells.

Phlebotomist One who practices phlebotomy; the individual drawing a specimen of blood.

Phlebotomy The puncture of a blood vessel to collect blood; literally, "the letting of blood in the treatment of disease."

Plasma The noncellular component of anticoagulated whole blood; plasma contains clotting factors.

Preanalytical errors Factors that affect specimens before tests are performed and that can lead to error if not controlled; they are classified as controllable or uncontrollable.

Preservative A substance or preparation added to a specimen to prevent changes in the constituents of a specimen.

Serum The watery portion of blood that remains after coagulation has occurred; it is obtained after centrifugation.

Sharps Any object which could readily puncture or cut the skin of an individual when encountered.

Sharps container A container designed for the disposal of sharps; required and regulated by the Occupational Safety and Health Administration (OSHA).

Skin puncture Collection of capillary blood usually from a pediatric patient by making a thin cut in the skin, usually at the heel of the foot.

Specimen A sample or portion of body fluid or tissue collected for examination, study, or analysis.

Venipuncture All of the steps involved in obtaining an appropriate and identified blood specimen from an individual's vein.

Venous occlusion Obstruction of the return of venous blood to the heart and distention of the veins; in phlebotomy, this is a temporary blockage caused by application of pressure, usually from a tourniquet.

Proper collection, processing, storage, and transport of common sample types used for diagnostic testing are critical to the provision of quality test results. Each step involved, as well as factors associated with the patient, can be the source of errors that cause inaccurate results. Minimizing these errors will result in more reliable information for use by health care professionals in providing quality patient care.

This chapter provides a review of the most common specimen types and discusses how they are (1) collected, (2) identified, (3) processed, (4) stored, and (5) transported. Differences between adult and pediatric collection are highlighted.

PATIENT IDENTIFICATION

Before any specimen is collected, the phlebotomist must confirm the identity of the patient using two or three items of identification (e.g., name, medical record number, date of birth). In specialized situations, such as paternity testing or other tests of medicolegal importance, establishment of a chain of custody for the specimen may require that additional patient identification, such as a photograph, be provided.

Identification must be an active process, with the patient stating his or her name and the phlebotomist verifying information either on the patient's wrist band or on the test requisition form or computer order. In the case of pediatric patients the parent or guardian should be present and be asked to actively provide identification of the child in a manner to prevent a yes or no answer from an often distracted parent, such as: "Please tell me the name of your child."

TYPES OF SPECIMENS

Types of biologic **specimens** that are analyzed in clinical laboratories include (1) whole blood; (2) serum; (3) plasma; (4) urine; (5) feces; (6) saliva; (7) other body fluids such as spinal, synovial, amniotic, pleural, pericardial, and ascitic fluids; and (8) cells and various types of solid tissue. The World Health Organization and the Clinical and Laboratory Standards Institute (CLSI) have published several guidelines for collecting many different types of specimens under standardized conditions (Table 6.1).

Blood

Blood for analysis may be obtained from veins, arteries, or capillaries. Venous blood is usually the specimen of choice. Arterial puncture is used mainly for blood gas analyses. In young children and for many point-of-care tests, skin puncture is frequently used to obtain capillary blood. The process of collecting blood is known as **phlebotomy** (from *phleb,* which means vein, and *tome,* to cut or incise) and should always be performed by a trained **phlebotomist**.

Venipuncture

In the clinical laboratory, **venipuncture** is defined as all of the steps involved in obtaining an appropriate and identified blood specimen from a patient's vein (CLSI H3-A6, see Table 6.1).

Preliminary steps. Prior to collection, the patient should be asked about latex allergies, and if necessary the phlebotomist

should secure nonlatex equipment. Any signed consent forms or special requisitions should be completed.

Before collection, the phlebotomist should dress in personal protective equipment (PPE), such as an impervious gown and gloves. PPE and adherence to standard precautions limit the spread of infectious disease from one patient to another and promote the safety of the phlebotomist. PPE is often brightly colored in pediatric institutions, where small children may be frightened of anyone in a white coat or gown. When collecting specimens from a patient in isolation in a hospital, the phlebotomist must often dress in specialized PPE. The extent of the precautions required varies with the nature of the patient's illness and the institution's policies and bloodborne pathogen plan.

TABLE 6.1 Clinical and Laboratory Standards Institute Documents Related to Specimen Collection, Processing, and Transport

Document Name	Document Number
Accuracy in patient and sample identification	GP33-P
Blood collection on filter paper for newborn screening programs	LA4-A5
Body fluid analysis for cellular composition	H56-A
Collection, transport, and processing of blood specimens for testing plasma-based coagulation assays and molecular hemostasis assay	H21-A5
Collection, transport, preparation, and storage of specimens for molecular methods	MM13-A
Procedures and devices for the collection of diagnostic capillary blood specimens	H04-A6
Procedures for the collection of arterial blood specimens	H11-A4
Procedures for the collection of diagnostic blood specimens by venipuncture	H3-A6
Procedures for the handling and transport of diagnostic specimens and etiologic agents	H5-A3
Protection of laboratory workers from occupationally acquired infections	M29-A3
Routine urinalysis and collection, transportation, and preservation of urine specimens	GP16-A3
Selecting and evaluating a referral laboratory	GP9-A
Tubes and additives for venous blood specimen collection	H1-A5

If appropriate, the phlebotomist should verify that the patient has fasted, identify what medications are being taken or have been discontinued as required, and determine any other relevant information. The patient should be comfortable, seated or supine, and should have been in this position for as long as possible before the specimen is drawn. At no time should venipuncture be performed on a standing patient.

Infants and young children may need to be held to prevent movement. A young child may be held sitting upright in a parent's lap, with the parent helping to support and hold the patient and arm still (Fig. 6.1). An infant's blood is often drawn with the infant in a supine position, and the infant may be swaddled in a blanket, or a papoose board may be used to restrain movement.

Either of the patient's arms should be extended in a straight line from the shoulder to the wrist. An arm with extensive scarring or a hematoma at the intended collection site should be avoided. If a woman has had a mastectomy, arm veins on that side of the body should not be used. If a woman has had double mastectomies within 6 months, a vein on the back of the hand or ankle should be used. Outside of 6 months, blood should be drawn from the arm of the side on which the first procedure was performed.

Before performing a venipuncture, the phlebotomist should estimate the volume of blood to be drawn and select the appropriate number and types of tubes for the tests requested. This may be facilitated by computer-generated collection recommendations and should be designed to collect the minimum amount necessary for testing. Estimating volume of blood to be drawn is especially critical in a pediatric setting. Because an average-weight newborn infant has a total blood volume of approximately 350 mL, blood collection in the pediatric population should not exceed recommended volumes for the pediatric patient's weight. Iatrogenic blood loss can lead to unnecessary blood transfusions and increased risk of exposure to bloodborne pathogens.

Fig. 6.1 Holding a child for venipuncture. (Modified from World Health Organization. [2010]. WHO guidelines on drawing blood: best practices in phlebotomy. *Pediatric and Neonatal Blood Sampling*. Geneva: World Health Organization. http://www.ncbi.nlm.nih.gov/books/NBK138647.)

In addition to tubes, an appropriate needle must be selected. The most commonly used sizes for adults are 19 to 22 gauge. (The larger the gauge number, the smaller the bore.) The usual choice for an adult with normal veins is 20 gauge. In pediatric patients, 23- to 25-gauge needles are most commonly used, with 23-gauge being the preferred size because infant veins are very small. Even for larger volumes of blood, rarely will a needle larger than a 21 gauge be used because it will not fit into the vein easily. A needle is typically 1.5 inches (3.7 cm) long, but 1-inch (2.5-cm) needles, usually attached to a winged or butterfly collection set, are also commonly used in pediatrics and geriatrics. Finally, the phlebotomist should ensure that all post-draw safety devices are in place, including convenient and safe access to proper disposal devices for contaminated needles and associated devices and appropriate post–blood draw supplies (gauze and bandage).

Location. The median cubital vein in the antecubital fossa, or crook of the elbow, is the preferred site for collecting venous blood in adults because the vein is large and close to the surface of the skin. Veins on the back of the hand or at the ankle may be used, although these should be avoided in people with diabetes or with poor circulation. However, in infants and children younger than 2 years old, collection from superficial veins is recommended, and dorsal hand veins are preferred over the median cubital vein. An arm containing a cannula or an arteriovenous fistula should not be used without consent of the patient's physician.

Preparation of the site. The area around the intended puncture site should be cleaned with institutionally approved cleanser. Commonly used materials are a prepackaged alcohol swab, a gauze pad saturated with 70% isopropanol, or a benzalkonium chloride solution. Cleaning of the puncture site should be done with a circular motion from the site outward. The skin should be allowed to dry in the air. No alcohol or cleanser should remain on the skin because traces may increase patient discomfort and cause hemolysis and invalidate test results.

Timing. The time at which a specimen is obtained is important for blood constituents that undergo marked **diurnal variation** (e.g., corticosteroids, iron), those for which a fasting sample has been requested, and those used to monitor drug therapy. In each case the timing should match the conditions under which reference intervals or clinical decision points were determined. Furthermore, timing is important in relation to specimens for alcohol or drug measurements in association with medicolegal considerations.

Venous occlusion. After the skin is cleaned, a blood pressure cuff or a tourniquet is applied 4 to 6 inches (10 to 15 cm) above the intended puncture site for adults. This obstructs the return of venous blood to the heart and distends the veins (**venous occlusion**). When a blood pressure cuff is used, it is usually inflated to approximately 60 mm Hg (8.0 kPa). If a dorsal hand vein is being accessed, no tourniquet is used. The phlebotomist applies enough pressure with the hand holding the patient's wrist and hand to occlude and distend the vein.

It is rarely necessary to leave a tourniquet in place for longer than 1 minute after venous access is secured, but even within this short time, the composition of blood changes. Although the changes that occur in 1 minute are slight, marked changes have been observed after 3 minutes for some chemistry analytes. The composition of blood drawn first is most representative of the composition of circulating blood and the least affected by fluid shifts where protein-bound components and other large molecules will be concentrated; water-soluble smaller molecules such as electrolytes may be less affected. The first-drawn specimen should therefore be used for analytes such as calcium and other analytes that are both protein bound and pertinent to critical medical decisions. A uniform procedure for the order of draw for tests should therefore be established (see later). If it is possible to collect only a small volume of blood, the priority of which tests to perform should be established in consultation with the ordering physician.

Two special notes on the collection process: (1) Pumping of the fist before venipuncture should be avoided because it causes an increase in plasma potassium, phosphate, and lactate concentrations. Increased lactate lowers blood pH and causes the plasma ionized calcium concentration to increase. (2) The stress associated with blood collection can have effects on patients at any age. Thus plasma concentrations of analytes affected by stress, such as cortisol, thyroid-stimulating hormone, and growth hormone, may increase. Stress occurs particularly in young children who are frightened, struggling, and held in physical restraint. Collection under these conditions may cause adrenal stimulation, leading to an increased plasma glucose concentration, or may create increases in the serum activities of enzymes that originate in skeletal muscle.

Order of draw for multiple blood specimens. In a few patients, backflow from blood tubes into veins occurs owing to a decrease in venous pressure but can be minimized if the arm is held downward. When collecting multiple specimens with an evacuated tube system, one of the primary concerns is to prevent cross-contamination between tubes. To minimize these and other problems, blood should be collected into tubes in the order outlined in Table 6.2.

Collection with evacuated blood tubes. Evacuated blood tubes are usually considered to be safer, less expensive, more convenient, and easier to use than syringes. Tubes vary by the type of **additive** and the volume of the tube. The different types of additives are identified by the color of the stopper used. Color coding of specimen collection tubes is not yet harmonized and may vary according to manufacturers (CLSI H1-A5, see Table 6.1). Serum or plasma separator tubes are available that contain an inert, polymer gel material that settles like a disk between cells and supernatant when the tube is centrifuged. Release of intracellular components into the supernatant is thus generally prevented. Most importantly, these separator tubes may be used as primary containers from which serum or plasma can be directly aspirated by a number of analytical instruments, avoiding aspiration of red blood cells (RBCs) or possible errors of patient or sample identification during aliquoting. Increasingly, tubes are sold for special applications, such as RNA isolation. As with all specimen collection containers, these less common tubes must be validated by each laboratory before use if not approved by the manufacturer for the specific analysis to be conducted.

TABLE 6.2 Recommended Order of Draw for Multiple Specimen Collection

Stopper Color	Contents	Inversions
Yellow	Sterile media for blood culture	8
Royal blue	No additive	0
Clear	Nonadditive; discard tube if no royal blue used	0
Light blue	Sodium citrate	3–4
Gold/red	Serum separator tube	5
Red/red, orange/yellow, royal blue	Serum tube, with or without clot activator, with or without gel	5
Green	Heparin tube with or without gel	8
Tan (glass)	Sodium heparin	8
Royal blue	Sodium heparin, sodium EDTA (trace metal free)	8
Lavender, pearl white, pink/pink, tan (plastic)	EDTA tubes, with or without gel	8
Gray	Glycolytic inhibitor	8
Yellow (glass)	ACD for molecular studies and cell culture	8

ACD, Acid citrate dextrose; *EDTA,* ethylenediaminetetraacetic acid. Modified from information in Clinical and Laboratory Standards Institute. (2003). *Tubes and additives for venous blood specimen collection: CLSI-approved standard H1-A5* (5th ed.). Wayne: Clinical and Laboratory Standards Institute; and Kiechle, F. L. (Ed.), (2005). *So you're going to collect a blood specimen: an introduction to phlebotomy* (11th ed.). Northfield: College of American Pathologists.

Blood collected into a tube containing one additive should never be transferred into other tubes because the first additive may interfere with tests for which a different additive is specified. For example, ethylenediaminetetraacetic acid (EDTA) contamination can cause an erroneously reported hyperkalemia or hypocalcemia when an inappropriate tube type is analyzed.

After the skin has been cleaned, the needle is guided gently into the patient's vein; when the needle is in place, the tube is pressed forward into the holder to puncture the stopper and release the vacuum. As soon as blood begins to flow into the tube, the tourniquet is released without moving the needle. The tube is filled until the vacuum is exhausted. It is critically important that the evacuated tube be filled completely. Many additives, particularly for coagulation testing, are provided at concentrations in the tube based on a "full and proper" collection; both short and too-full draws can be a source of **preanalytical error** because they can significantly affect the established testing parameters that are based on a properly collected sample. Vacuum tubes should never be opened and filled from a syringe or other source.

Blood collection with a syringe. Syringes are customarily used for patients with difficult veins and for blood gas analysis. The syringe with the needle attached should be aligned with the vein to be entered and the needle pushed into the vein with the bevel up and at an angle to the skin of approximately 15 degrees. When the initial resistance of the vein wall is overcome, forward pressure on the syringe is eased, and the blood is withdrawn by very gently pulling back the plunger of the syringe.

After filling the syringe and completing the collection, if the sample needs to be transferred to an evacuated tube, the same needle, a new needle, or a transfer device should be used to puncture the cap of the tube. The tube should be allowed to fill passively using its vacuum. Vigorous withdrawal of blood into a syringe during collection or forceful transfer from the syringe to the receiving vessel may cause hemolysis of blood, making the sample invalid for testing.

Completion of collection. When blood collection is complete and the needle withdrawn, a dry gauze pad should be held tightly over the puncture site with the arm raised to stop residual bleeding and promote the clotting process; the pad can be held in place firmly by a bandage or by a nonadhesive strap. The needle is covered, and the needle and the tube holder are immediately discarded into a **sharps container**. In the event that a winged (butterfly) set is used, the wings are either pushed forward to cover the needle or a button is pressed to a spring that retracts the needle. All used supplies should be discarded in a hazardous waste receptacle.

Tubes should then be labeled per institutional policy. Most institutions have a written procedure prohibiting the advance labeling of tubes because of the potential for mislabeling, one of the most common sources of preanalytical error. Many US institutions recommend showing the labeled tube to the patient to further confirm correct identification.

Venipuncture in children. The techniques for venipuncture in children and adults are similar. However, children are likely to make unexpected movements, and assistance in holding them still is often desirable. Although a syringe or an evacuated blood tube system may be used to collect specimens, a syringe with a winged butterfly collection set is more commonly used in younger children. The pressure on a syringe can be more easily controlled by the phlebotomist than the vacuum pressure in an evacuated tube, thus preventing the pressure from pulling small veins closed. In the pediatric population, alternative collection through skin puncture is often used.

Skin Puncture

Skin puncture is an open collection technique in which the skin is punctured by a lancet and a small volume of blood is collected into a microdevice. In practice, skin puncture is used in situations in which (1) sample volume is limited (e.g., pediatric applications), (2) repeated venipunctures have resulted in severe vein damage, or (3) patients have been burned or bandaged and veins therefore are unavailable for venipuncture. This technique is also commonly used when the sample is to be applied directly to a testing device in a point-of-care testing situation or to filter paper. It is most often performed on the tip of the finger or the heels of infants. In an infant younger than 6 months, the lateral or medial plantar surface of the foot should be used for skin puncture (Fig. 6.2). These areas are the fleshiest part of the foot in an infant, with the

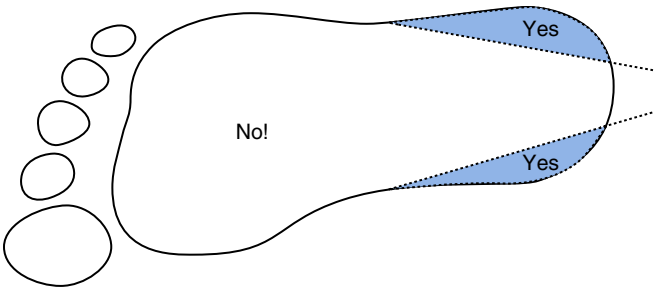


Fig. 6.2 Acceptable sites for skin puncture to collect blood from an infant's foot. (Modified from Blumenfeld, T. A., Turi, G. K., & Blanc, W. A. [1979]. Recommended site and depth of newborn heel punctures based on anatomical measurements and histopathology. *Lancet, 1,* 230–233. Reprinted with permission from Elsevier.)

TABLE 6.3 Tip Lengths in Finger- and Heel-Stick Devices

Collection Type and Age of Use	Tip Length (mm)
Heel stick: premature infants	0.85
Heel stick: term infants to 6 months old	1.25
Heel stick: 6 months to 1 year	1.25
Finger stick: walking to 8 years	1.5
Finger stick: >8 years	1.75–2.2

most distance between the skin surface and the underlying bone, decreasing the risk of inadvertently hitting bone and causing a bone infection. The back of the heel and the toes should not be used. Blood collection from anywhere on the foot should be avoided on ambulatory patients; thus, when an infant starts walking, a heel stick should no longer be performed. Lancets are made specifically for a heel stick or a finger stick, and they should not be used interchangeably because they have different tip lengths (Table 6.3). The finger-stick procedure should not be performed on infants younger than 6 months because no commercially available device punctures shallow enough to avoid bones. The complete procedure for collecting blood from infants using skin puncture is described in CLSI H04-A6 (see Table 6.1).

To collect by skin puncture, the phlebotomist first cleans the skin with an approved cleaning solution; once the skin is dry, it is quickly punctured by a sharp stab with a lancet. To minimize the possibility of infection, a different site should be selected for each puncture. The finger should be held in such a way that gravity assists collection of blood at the fingertip and the lancet held to make the incision as close to perpendicular to the fingernail as possible. Massage of the finger to stimulate blood flow should be avoided because it causes the outflow of debris and tissue fluid, which does not have the same composition as plasma. To improve circulation of the blood, the finger or heel may be warmed by application of a warm, wet washcloth or a specialized device, such as a heel warmer, for 3 minutes before cleaning and the lancet is applied. Warming properly will not only cause the capillary blood to be free flowing and improve the ability to collect the sample, but the analytes in the sample will also approach arterial blood values. In a cold heel, the values more approximate

TABLE 6.4 Order of Draw for Skin Puncture: Capillary Blood

Usage or Additive	Tube Top Color
Blood gases (heparin)	Microhematocrit tubes
Ethylenediaminetetraacetic acid	Lavender
Heparin	Green
Other additives	Light blue, gray
Nonadditives	Red, tiger, yellow

venous blood. The first drop of blood is wiped off, and subsequent drops are transferred to the appropriate collection tube by gentle contact. Filling should be done rapidly to prevent clotting, and introduction of air bubbles should be prevented.

A variety of collection tubes are commercially available for capillary blood collection. Samples should be collected by dripping or capillary action from the puncture site into the small tubes. Drop-by-drop collection and scooping along the skin with the edge of the tube should be avoided because both practices increase hemolysis. The correct order of filling of these devices differs from evacuated blood tubes because the concerns are different. The anticoagulant tubes, especially the EDTA tube for the complete blood count (CBC), are collected first rather than last. The serum tubes are collected last (Table 6.4).

For collection of blood specimens on filter paper for molecular genetic testing and neonatal screening (CLSI LA4-A5, see Table 6.1), the skin is cleaned and punctured as described previously. The first drop of blood is wiped away and the filter paper is gently touched against a large drop of blood that is allowed to soak into the paper to fill the marked circle. Only a single application per circle should be made. The procedure is repeated to fill all the circles. The filter paper should be properly labeled and then air-dried for 2 to 3 hours before storage in a labeled paper envelope.

Arterial Puncture

Arterial puncture requires considerable skill and is usually performed only by physicians or specially trained personnel. Preferred sites of arterial puncture are, in order, the (1) radial artery at the wrist, (2) brachial artery in the elbow, and (3) femoral artery in the groin, with sites in the arm used most often. The proper technique for arterial puncture is described in CLSI H11-A4 (see Table 6.1).

Anticoagulants and Preservatives for Blood

Serum is defined as that portion of blood that remains after **coagulation** has occurred and the cells have been removed and is the specimen of choice for many analyses. Samples are collected into tubes with no additive or with a clot activator and must be allowed to complete the coagulation process before further processing. **Plasma** is defined as the noncellular component of anticoagulated whole blood after the cellular components have been removed. Heparinized plasma is increasingly being used for routine chemistry testing to decrease turnaround time because it is not necessary to wait for the blood to clot. However, considerable differences may be observed between the concentrations of certain analytes

(such as total protein) in serum and in plasma. For molecular diagnostics, anticoagulated whole blood or plasma is more likely to be the specimen of choice for either genomic DNA isolation from the white blood cells (WBCs) or viral identification and quantification from plasma.

For any assay provided for clinical use, manufacturers specify the appropriate sample type(s) for which they have validated the assay. Use of different sample types must be validated by laboratory personnel before use.

Heparin. Heparin is the most widely used **anticoagulant** for chemistry testing. It is available as sodium, potassium, lithium, and ammonium salts, all of which can adequately prevent coagulation by accelerating the action of antithrombin III, which neutralizes thrombin. This prevents the formation of fibrin from fibrinogen. Heparin is a naturally occurring anticoagulant and has the disadvantages of high cost and a more temporary action of anticoagulation than is attained by the chemicals discussed later. The use of lithium or ammonium heparin is unacceptable for lithium and ammonia measurements, respectively.

It should be noted that heparin is unacceptable for most tests performed using polymerase chain reaction (PCR) because of inhibition of the polymerase enzyme by this large molecule. DNA can be extracted from heparinized samples, but amplification may be reduced, and the effect of heparin on any molecular diagnostic assay should be assessed as part of a method validation study.

Ethylenediaminetetraacetic acid. EDTA is a chelating agent of divalent cations such as Ca^{2+} and Mg^{2+} that is particularly useful for (1) hematologic examinations including transfusion medicine applications, (2) measurement of intracellular drugs such as cyclosporine or tacrolimus, (3) HbA_{1c} analysis, (4) isolation of genomic DNA, and (5) qualitative and quantitative virus determinations by molecular techniques. It is used most commonly as the tripotassium salt. EDTA prevents coagulation by binding calcium, which is essential for the clotting mechanism.

Because it chelates calcium, magnesium, and iron, EDTA is unsuitable for specimens for these analyses. In addition, EDTA, probably by chelation of required metallic cofactors, inhibits alkaline phosphatase and creatine kinase activities.

Sodium fluoride. Sodium fluoride (NaF) is a weak anticoagulant that is often added as a preservative for blood glucose and lactate. As a preservative, it is often used together with another anticoagulant such as potassium oxalate. It exerts its **preservative** action by inhibiting the enzyme systems involved in glycolysis, although such inhibition is not immediate, and a certain amount of degradation occurs during the first hour after collection. Without an antiglycolytic agent, the blood glucose concentration decreases approximately 100 mg/L (0.56 mmol/L) per hour at 25°C. The rate of decrease is faster in newborns because of the increased metabolic activity of their erythrocytes and in leukemic patients because of the high metabolic activity and number of the WBCs. Because of the delay in onset of action, recent protocols recommend placing the tube on ice until the sample can be separated to ensure accurate glucose measurements. Newer formulations containing both NaF and citrate or EDTA immediately inhibit glycolysis and preserve glucose in the specimen for at least 24 hours.

Citrate. Sodium citrate solution, in a ratio of 1 part to 9 parts of blood, is widely used for coagulation studies. The correct ratio of blood to anticoagulant is critical to achieve proper coagulation measurements because the anticoagulant effect is reversed by addition of standard amounts of Ca^{2+} that are based on a proper collection volume. Because citrate chelates calcium, it is unsuitable as an anticoagulant for specimens for measurement of this element. It also inhibits aminotransferases and alkaline phosphatase but stimulates acid phosphatase when phenylphosphate is used as a substrate.

Acid citrate dextrose. When both the form and function of the cellular components are required, such as for cytogenetic testing, samples are generally collected into acid citrate dextrose (ACD) anticoagulant. There are two ACD tube designations: ACD A and ACD B. These differ only by the concentrations of the additives. Both enhance the vitality and recovery of WBCs for several days after collection of the specimen.

Oxalates. Sodium, potassium, ammonium, and lithium oxalates inhibit blood coagulation by forming rather insoluble complexes with calcium ions.

Oxalate inhibits several enzymes, including acid and alkaline phosphatases, amylase, and lactate dehydrogenase (LDH), and may cause precipitation of calcium as the oxalate salt.

Iodoacetate. Sodium iodoacetate is an effective antiglycolytic agent (with the same caveats mentioned earlier) and a substitute for NaF. It inhibits creatine kinase but appears to have no notable effects on other clinical tests.

Influence of Site of Collection on Blood Composition

Blood obtained from different sites differs in composition. In general, skin puncture blood is more similar to arterial blood than to venous blood, depending on the collection condition as described earlier. Thus there are no clinically significant differences between freely flowing capillary blood and arterial blood in pH, PCO_2, PO_2, and oxygen saturation. The PCO_2 of venous blood is up to 6 to 7 mm Hg (0.8 to 0.9 kPa) higher. Venous blood glucose is as much as 70 mg/L (0.39 mmol/L) less than capillary blood glucose.

Collection of Blood From Intravenous or Arterial Lines

When blood is collected from a central venous catheter or arterial line, it is necessary to ensure that the composition of the specimen is not affected by the fluid that is infused into the patient. With clinical approval, fluid may be shut off using the stopcock on the catheter, and 10 mL of blood is aspirated through the stopcock and discarded before the specimen is withdrawn. In pediatric patients, 10 mL of blood is often not feasible, so less volume is aspirated, although the goal is still to aspirate approximately 3 times the dead space of the line before collecting the sample for testing. Aspirating this blood

and clearing the lines is equally important for molecular diagnostics and coagulation testing because the stopcock is often heavily saturated with heparin. Collection of samples for therapeutic drug monitoring should not be done from the line used for the infusion, irrespective of the time since infusion or amount of blood aspirated, because the drug may adhere to the line and leak into the collected sample.

In theory, blood may be collected from the veins of an arm below an IV line without interference from the fluid being infused because retrograde blood flow does not occur in the veins and the fluid that is infused must first circulate through the heart and return to the tissue before it reaches the sampling site. However, collection from the arm without the IV line is strongly recommended.

Hemolysis

Hemolysis is defined as the disruption of the RBC membrane, resulting in the release of hemoglobin, and may be the consequence of intravascular events (in vivo hemolysis) or may occur subsequent to or during blood collection (in vitro hemolysis). Serum and plasma show visual evidence of hemolysis when the hemoglobin concentration exceeds 50 mg/dL (7.7 μmol/L). Slight hemolysis has little effect on most but not all test values. Test interference at this concentration may be observed on those constituents that are present at a higher concentration in erythrocytes than in plasma, such as activities or concentrations of LDH, aspartate transaminase (AST), potassium, magnesium, and phosphate. Most manufacturers provide data on the effects of hemolysis on the analytical performance of individual tests, and this should be evaluated in the selection of individual methods. Laboratory personnel must define at what level of hemolysis results should be held and not reported to prevent poor clinical action on unreliable test results.

In molecular diagnostic testing, hemoglobin may interfere with the amplification reaction. In some situations the isolation of nucleic acid is sufficiently selective that free hemoglobin from the ruptured cells is removed and will not cause a problem. However, with hemolyzed blood, alternative or additional extraction methods are usually needed to ensure that RNA is fully and accurately transcribed and that the greatest amplification of DNA is achieved.

Urine

The timing of urine specimen collection is dictated by the tests to be performed. Untimed or random specimens are suitable for some chemical tests; usually, urine specimens must be collected over a predetermined interval of time, such as 4, 12, or 24 hours. A clean, early morning, fasting specimen is usually the most concentrated specimen and thus is preferred for microscopic examinations and for the detection of abnormal amounts of constituents, such as proteins, or of unusual compounds, such as chorionic gonadotropin. The clean timed specimen is one obtained at specific times of the day or during certain phases of the act of micturition. Although bacterial examination of the first 10 mL of urine voided is most appropriate to detect urethritis, the midstream specimen is best for investigating bladder disorders. The double-voided specimen is the urine excreted during a timed period after complete emptying of the bladder; it is used, for example, to assess glucose excretion during a glucose tolerance test. Similarly, in some metabolic disorders, urine must be collected during or immediately after symptoms of the disease appear.

When they are to be tested for their alcohol and drugs of abuse content, urine specimens are collected under rigorous conditions, often including enhanced patient identification and special collection facilities where water that could be used for urine dilution is restricted or colored, for example. It is necessary for laboratory personnel to be aware of the relevant legal requirements for this type of testing and plan in advance how these samples and the supporting paperwork will be collected and handled.

Catheter specimens are used for microbiologic examination in critically ill patients and in those with urinary tract obstruction but should not normally be obtained just for examination of chemical constituents. The suprapubic tap specimen is a useful alternative because the tap is unlikely to cause infection. After appropriate cleaning of the skin over the full bladder, a 22-gauge spinal needle is passed through a small wheal made by a local anesthetic. The bladder is penetrated and the urine withdrawn into the syringe.

Even though tests in the clinical laboratory are not usually affected by lack of sterile collection procedures, the patient's genitalia should be cleaned before each voiding to minimize the transfer of surface bacteria to the urine.

Currently, urine is an uncommon specimen type in the molecular diagnostic laboratory for genomic testing, although some laboratories use urine samples for bladder cancer screening and monitoring of therapy for bladder cancer. However, urine is frequently used for molecular testing for infectious agents, such as *Chlamydia,* a common sexually transmitted organism, or BK virus, associated with potential rejection or failure of transplanted kidneys. Because most requests involve a specific organism, an untimed or random urine specimen collected into a sterile container with no preservative is usually acceptable.

Timed Urine Specimens

The collection period for timed specimens should be long enough to minimize the influence of short-term biologic variations. When specimens are to be collected over a specified period of time, the patient's close adherence to instructions to the collection procedure and restrictions to diet or drug ingestion is critically important and a common source of a preanalytical variable. The bladder must be emptied at the time the collection is to begin and this urine discarded. Thereafter all urine must be collected until the end of the scheduled time, including emptying of the bladder at the end of the collection period. Precautions should be taken to prevent fecal contamination of the urine. When the collection is made over several hours, urine should be passed into a separate container at each voiding and then emptied into a larger container for the complete specimen. This two-step procedure prevents the danger of patients

Fig. 6.3 Urine collection device used in children.

splashing themselves with a preservative, such as acid that may be included in the collection device. The large container generally should be stored at 4°C during the entire collection period.

Urine should not be collected at the same time for two or more tests requiring different preservatives. Removal of aliquots during the collection procedure is not permissible even when the volume removed is measured and corrected because excretion of most compounds varies throughout the day. Appropriate information regarding the collection, including warnings with respect to handling of the specimen, should appear on the container label.

When a timed collection is complete, the specimen should be delivered without delay to the clinical laboratory, where the volume should be measured by using graduated cylinders or by weighing the container and the urine when preweighed or uniform containers are used. The mass in grams may be reported as if it were the volume in milliliters.

Before a specimen is transferred into small containers for each of the ordered tests, it must be thoroughly mixed to ensure homogeneity because the specific gravity, volume, and composition of the urine all may vary throughout the collection period and settling may occur.

Collection of Urine From Children

Collection of any type of urine specimen from an infant is difficult, with timed collections being the most problematic. The approved method for collecting a random urine specimen from an infant involves a process known as bagging. A plastic bag (e.g., U-Bag, Hollister, Chicago, IL) is placed around the infant's genitalia after cleaning and then is held in place by a mild adhesive (Fig. 6.3). The baby's diaper is reapplied to help hold the bag in place. As soon as voiding has occurred, the bag containing the urine sample is removed and emptied into a regular urine collection cup. In infants, a random or spot urine collection for something as simple as a urinalysis may require either catheterization or a suprapubic tap collection. In infants and very young children requiring a rare 24-hour urine collection, hospitalization and catheterization are often required to obtain a complete collection.

Urine Preservatives

Preservatives for urine (CLSI GP16-A3, see Table 6.1) are usually added to reduce bacterial action or chemical decomposition or to solubilize constituents that otherwise might precipitate out of solution. Some specimens should not have *any* preservatives added because of the possibility of interference with analytical methods.

One of the most acceptable forms of preservation of urine specimens is refrigeration immediately after collection; it is even more successful when combined with chemical preservation. Acidification to a pH less than 3 is widely used to preserve 24-hour specimens for determination of calcium, steroids, adrenaline, noradrenaline, and vanillylmandelic acid. However, precipitation of urates will occur, making the sample unsuitable for measurement of uric acid.

A mild base, such as sodium bicarbonate or a small amount of sodium hydroxide, is used to preserve porphyrins, urobilinogen, and uric acid. pH should be adjusted to between 8 and 9.

Feces

Small aliquots of feces are frequently analyzed to detect the presence of "hidden" or occult blood. Guaiac-based occult blood screening is subject to many interferences causing both false-positive and false-negative results. Results should be interpreted cautiously unless the intent is to identify current bleeding in any portion of the gastrointestinal (GI) tract in an emergent patient. Newer immunochemical-based tests, so-called iFOB tests, have decreased the effects of food intake greatly for the purpose of identifying a lower GI tract bleed that may indicate the presence or possibility of malignant growth. However, iFOB methods will not detect upper GI bleeds. In either case, tests for occult blood should be done on aliquots of excreted stools rather than on material obtained on the glove of a physician doing a rectal examination because this procedure may cause enough bleeding to produce a positive result.

In newborns the first specimen from the bowel (meconium) may be used for detection of maternal drug use during the gestational period, which requires specific attention to the details of collection and identification similar to the chain of custody procedure for urine collection discussed earlier. Feces from infants and children may be screened for tryptic activity or for increased fecal fat concentrations, both of which can be indicators of cystic fibrosis. Fecal material is also commonly collected in childhood for the detection of parasites (ova and parasites [O & P]), enteric disease organisms such as *Salmonella* and *Shigella*, and viruses, all of which are useful in sorting out the differential diagnosis of diarrhea. Fecal testing is also used for helping to determine causes of malabsorption.

Usually, no preservative is added to the feces, but the container should be kept refrigerated throughout the collection period, and care should be taken to prevent contamination from urine. When the collection is complete, the container and feces are weighed, and the mass of excreted feces is calculated. Results are generally reported as activity per unit weight.

Other Body Fluids

Specimens may be collected for analysis from a range of different body fluids. These include cerebrospinal fluid (CSF), pleural fluid, ascitic fluid, pericardial fluid, amniotic fluid, synovial fluid, and others.

Chorionic Villus Sampling

Chorionic villus sampling (CVS) allows for earlier (gestational period 10 to 12 weeks) diagnosis of inherited genetic disorders than is possible with amniotic fluid (gestational period 15 to 20 weeks) analysis. CVS is the technique of inserting a catheter or needle into the placenta and removing some of the chorionic villi, or vascular projections, from the chorion. This tissue has the same chromosomal and genetic makeup as the fetus and can be used to test for disorders that may be present in the fetus. After culture and growth for 3 weeks, DNA is extracted for analysis.

Buccal Cells

Collection of buccal cells (cells of the oral cavity of epithelial origin) has been identified as providing an excellent source of genomic DNA, as well as being less invasive than collection of blood. It is particularly useful for collecting cells with the patient's genomic DNA when the patient has had blood transfusions and thus another person's DNA or after bone marrow transplantation when the circulating blood cells are derived wholly or partially from the donor of the bone marrow. Two methods are used commonly to collect buccal cells: rinsing with mouthwash and using swabs or cytobrushes.

Rinsing of the oral cavity generally provides a higher yield of cells than can be obtained by using swabs. Mouthwash solutions high in phenol and ethanol are destructive to recovered cells and should be avoided. It is necessary for laboratory staff to validate a list of acceptable solutions.

For swabs, a sterile Dacron or rayon swab with a plastic shaft is preferred because calcium alginate swabs or swabs with wooden sticks may contain substances that inhibit PCR-based testing. After collection, the swab or cytobrush should be stored in an airtight plastic container or immersed in liquid, such as phosphate-buffered saline or viral transport medium.

Solid Tissue

Traditionally, the solid tissue most often analyzed in the clinical laboratory was malignant tissue from the breast for estrogen and progesterone receptors. Another traditional use is measurement of liver iron or copper to assist with the diagnosis of hemochromatosis or Wilson disease, respectively.

The same procedure may be used to obtain and prepare solid tissue for elemental or for toxicologic analysis; however, when trace element determinations are to be made, all materials used in the collection or handling of the tissue should be made of plastic or materials known to be free of contaminating trace elements.

Somatic gene analysis such as T and B cell clonality and the identification of possible clinically actionable mutations in malignant tissue (e.g., *KRAS* mutations) are now proving to be of increasing importance for clinicians in both diagnosis and direction of appropriate therapeutic options for the patient. For these studies, the molecular diagnostic laboratory often receives tissue that has been formalin fixed and paraffin embedded (FFPE) rather than fresh tissue because the request for further testing is generally made after the pathologist's diagnosis of the particular malignancy. In general, neutral buffered formalin, containing no heavy metals, will not interfere with amplification reactions. DNA can still be extracted from tissue embedded in paraffin, but the DNA will be degraded to low-molecular-weight fragments. In most cases, segments of DNA will amplify in a PCR reaction, but Southern blot methods are problematic because most require high-molecular-weight DNA.

Hair and Nails

Currently, the use of hair or nails in molecular diagnostics is limited to forensic analysis (genomic DNA identification). Hair and fingernails or toenails have been used for trace metal and drug analyses, with the potential advantage of timing of exposure if separate segments of longer hair are analyzed, although no standards for such testing currently exist. Use of such samples requires laboratory staff to validate the processes.

HANDLING OF SPECIMENS FOR ANALYSIS

Steps that are important for obtaining a valid specimen for analysis include (1) identification, (2) preservation, (3) separation and storage, and (4) transport.

MAINTENANCE OF SPECIMEN IDENTIFICATION

Although the collection of an acceptable specimen is a key aspect of excellent testing, proper identification of the specimen must be maintained at each step of the testing process to ensure that the correct result is reported for the correct patient at all times. The minimum information on any label associated with a specific specimen should include the patient's name, location, and identifying number, as well as the date and time of collection. All labels should conform to the laboratory's stated requirements. In the United States, there is no specific labeling used for samples from patients with infectious diseases. Universal precautions should always be used, meaning that all specimens should be treated as if they are potentially infectious with the following caveat. The exception will be samples from patients suspected to have known, high-risk pathogens (e.g., hemorrhagic viruses such as Ebola) that must have separate, pre-prepared sample handling protocols that may or may not involve the routine laboratory. Proactive procedures must be in place, including by whom and where such samples might be analyzed, unless the sample can be rendered safe (e.g., by heat treatment), in which case again standard universal precautions should be used.

In practice, every specimen container must be adequately labeled, even if the specimen must be placed in ice or if the container is so small that a label cannot be placed along the tube, as might happen with a capillary blood tube. Direct labeling of a capillary blood tube by folding the label like a flag around the tube is recommended. For any specimen submitted in a screw-cap test tube or cup, the label should be placed on the cup or tube directly, not just on the cap.

It is critical that samples be positively identified through all steps of processing and analysis. Aliquoting a sample from the primary collection container to one or more other containers configured for the instrumentation requires close attention to proper labeling and tracking of the sample identifiers to ensure samples are not switched. Good work practice includes "piece work" in which only a single patient's samples are in the work area at one time, the area is clean with no old labels present, and the worker is not disturbed. Because the majority of samples received from pediatric patients are in microtubes, most samples need to be aliquoted, poured into a microsampling device, or hand loaded onto instruments that use whole blood, because these systems are not made to deal with such small tubes. This extra handling of the specimens offers more opportunities for error and thus requires more strict attention to detail and analysis of the possible risks during design of the process.

Preservation of Specimens

Specimens must be properly treated both during transport to the laboratory and from the time the serum, plasma, or cells have been separated until analysis. For some tests, specimens must be kept at 4°C from the time the blood is drawn until the specimens are analyzed or until the serum or plasma is separated from the cells to minimize metabolism and degradation of sample components. Transfer of these specimens to the laboratory must be done by placing the specimen container on ice; a bag of ice is used to surround the sample to keep it cool while the sample and label are not subjected to possible water contamination.

For all test constituents that are thermally labile, serum and plasma should be separated from cells in a refrigerated centrifuge. Some specimens may need to be protected from both daylight and fluorescent light to prevent photodegradation, although the use of plastic, rather than glass, tubes has decreased this preanalytical variable.

For specimens that are collected in a remote facility with infrequent transportation by courier to a central laboratory, proper specimen processing must be done in the remote facility so that appropriately separated and preserved plasma or serum is delivered to the laboratory.

Add-on Requests

In the interest of preventing additional phlebotomies and to assess a clinical situation from a specimen collected at a specific time, many physicians request an "add-on" test (i.e., for the laboratory staff to perform a test on a sample already in the laboratory and processed). This is especially true in specimens collected from pediatric patients and for patients from an emergency department, where additional testing from the time of presentation with specific symptoms may be needed after a clinical diagnosis has been made or narrowed by the clinician. Each laboratory staff must establish its own guidelines for what will be allowed in what time frame.

Separation and Storage of Specimens

Plasma or serum should be separated from cells as soon as possible and certainly within 2 hours for some but not all analytes. However, premature separation of serum may permit continued formation of fibrin, which can clog sampling devices in testing equipment. If it is impossible to centrifuge a blood specimen within 2 hours, the specimen should be held at room temperature rather than at 4°C to decrease any effect on potassium measurement caused by leakage from the RBCs by inhibition of the Na^+,K^+-ATPase pump. For most plasma samples used for molecular diagnostics, the plasma should be removed from the primary tube promptly after centrifugation and held at −20°C. In all instances of freezing a sample, frost-free freezers should be avoided because they have a wide temperature swing during the freeze-thaw cycle.

Primary specimen tubes should always be centrifuged with the original cap in place. Such containment reduces evaporation, which occurs rapidly in a warm centrifuge with the air currents set up by centrifugation. Caps on the original tube also prevent aerosolization of infectious particles. Specimen tubes with requested testing for volatiles, such as ethanol, *must* have the initial cap in place while they are spun to prevent release of the volatile compound and result in an artificially reduced measurement. Centrifuging specimens with the cap in place also maintains anaerobic conditions, which are important in the measurement of carbon dioxide and ionized calcium.

Transport of Specimens

Hemolysis may occur in pneumatic tube systems unless the tubes are completely filled and movement of the blood tubes inside the specimen carrier is prevented. The pneumatic tube system should be designed to eliminate sharp curves and sudden stops of specimen carriers. However, with many systems, the plasma hemoglobin concentration may be increased, and the serum activity of RBC enzymes, such as LDH and AST, may also be increased. There are also tests that cannot be transported to the laboratory via a pneumatic tube system such as platelet function testing where samples must be hand delivered to the laboratory.

Although the remaining discussion uses the specific example of referral laboratory testing, many of the issues discussed, such as regulations related to shipping, are also relevant to a laboratory that receives specimens from outlying clinics. This may involve validating specific transport or storage conditions that are in conflict with existing CLSI recommendations. Before a referral laboratory is used for any tests, the quality of its work should be verified by the referring laboratory. Guidelines for selection and evaluation of a referral laboratory have been published (GP 9-A, see Table 6.1). For laboratories accredited by the College of American Pathologists (CAP), it is a requirement that the referring laboratory validate that the referral laboratory is CLIA'88 certified by obtaining a copy of the Clinical Laboratory Improvement Act (CLIA) certificate before specimens are shipped. For molecular diagnostic testing, this is of particular importance because often the latest genetic test being requested by a physician has not yet been moved from research interest status to patient care status and may not be available in a CLIA-certified laboratory.

Specimen type and quantity and specimen handling requirements of the referral laboratory must be observed, and in laboratories operating under CLIA'88 regulations, test results reported by a referral laboratory must be identified as

such when they are filed in a patient's medical record. Most reference laboratories have lower minimum volume requirements for pediatric specimens than for adult specimens, but these lower minimums may preclude being able to retest the sample if there is a problem with the initial analysis. The tube and transport condition for the specimen should be constructed such that the contents do not escape if the container is exposed to extremes of heat, cold, or sunlight. Reduced pressure may be encountered during air transport, together with vibration, and specimens should be protected from these adverse conditions by a suitable container. Variability in temperature is a significant factor causing instability of test constituents.

Polypropylene and polyethylene containers are usually suitable for specimen transport. Containers must be leakproof and should have a Teflon-lined screw cap that does not loosen under the variety of temperatures to which the container may be exposed. The materials of both stopper and container must be inert and must not have any effect on the concentration of the analyte.

For transport of frozen or refrigerated specimens, a Styrofoam container should be used. The container should be vented to prevent buildup of carbon dioxide under pressure and a possible explosion. Solid carbon dioxide (dry ice) is the most convenient refrigerant material for keeping specimens frozen. The amount of dry ice required in a container depends on the size of the container, the efficiency of its insulation, and the length of time for which the specimens must be kept frozen.

Various laws and regulations apply to the shipment of biologic specimens. Although such regulations theoretically apply only to etiologic agents (known infectious agents), all specimens should be transported as if the same regulations apply. Airlines deem dry ice a hazardous material; therefore the transport of most clinical laboratory specimens is affected and those who package the specimens should be trained in the appropriate regulations, such as those published by the International Air Transport Association.

The various modes of transport of specimens influence the shipping time and cost, and each laboratory staff needs to make its own assessment as to adequate service. The objective is to ensure that the properly collected, processed, and identified specimen arrives at the testing facility in time and under the correct storage conditions so that the analytical phase can proceed.

CONCLUSION

Accurate test results (always the right result for the right patient) begin and end with the integrity of the sample being tested. Integrity can be assured only by proper preparation of the patient; choice of sample container; collection of the sample; and finally transport, processing, and storage of the collected sample, with each step maintaining accurate identification. For these reasons, best laboratory practice demands attention to detail and following appropriate protocols. It is incumbent on laboratory professionals to fully delineate their processes in complete policies and procedures that not only cover the routine and correct procedures but also cover the unusual. These should include how to handle the process when the system breaks down and steps are not properly performed and may need to be addressed case by case by the laboratory director. Finally, laboratory staff should fully validate protocols that may not be covered under normal procedures or that may deviate from local regulatory guidelines (e.g., the US Food and Drug Administration or CLIA). Preanalytical variables can be lessened or even avoided if these steps are followed.

POINTS TO REMEMBER

- Proper identification of the patient is essential; samples should be properly labeled at all steps if separated from the primary collection container.
- Collection of all samples must be in the correct primary container and never transferred to another primary container.
- Attention to the details related to processing of the collected sample (time, temperature, special handling) should always follow validated policies.
- The accurate result for any patient's sample (that will be acted on by the clinician) depends on adherence to all policies and procedures.

REVIEW QUESTIONS

1. Identify the incorrect specimen collection scenario below.
   - **a.** Ask a parent to name a pediatric patient.
   - **b.** Ask an adult patient to state his or her name and another identifier.
   - **c.** Show the patient or parent the labeled tube after collection is complete.
   - **d.** Check the room and bed number of the patient to be collected and proceed with collection.
   - **e.** Contact a translation service for a patient unable to speak the local language.
2. Identify the correct tube (additive)/biochemical action below:
   - **a.** Lavender (EDTA)/anticoagulated specimen that will limit effects on cell morphology
   - **b.** Light blue (ACD)/chelation of calcium to allow effective coagulation testing
   - **c.** Red (no additive, no gel)/serum sample that can be sampled directly by many analyzers
   - **d.** Light green/no gel/sodium–heparin tube prevents any further cellular activity that might affect analytes such as glucose
   - **e.** Dark green/lithium heparin/requires prompt removal of the plasma from the cell pellet

3. Patients should not pump their fists after the tourniquet is applied because
   a. it may or may not cause venous stasis not under the control of the collector.
   b. it will affect the chemical interaction between the tube additive and the patient's blood.
   c. metabolic activity of the blood and muscles will affect analytes such as potassium and calcium through effects on pH.
   d. it will affect the distribution of analytes between plasma and serum.
   e. it will make it more difficult to find the patient's vein.
4. For urine collections, which of the following is true?
   a. Samples should be kept at room temperature for a timed collection.
   b. Special identification of the patient may be required for forensic purposes.
   c. Identification of added preservative does not need to be communicated to the patient because all are safe.
   d. For a 24-hour collection, the patient may collect into a rinsed milk jug and then pour the sample into the provided container at the end of the collection period.
   e. No special handling is required because of collection preservative if the sample is sent to another laboratory.
5. Additional testing requested by the clinician treating the patient (add-on testing) should always be carried out, when
   a. the sample was not stored correctly per local validated storage policies.
   b. the patient is not identified correctly.
   c. the request is for a labile compound that may have been affected by storage
   d. the sample needs to be retrieved from a waste container.
   e. the request is for a test that has been validated on the sample submitted.

SUGGESTED READINGS

Arena, J., Emparanza, J. I., Nogues, A., & Burls, A. (2005). Skin to calcaneous distance in the neonate. *Archives of Disease in Childhood. Fetal and Neonatal Edition*, *90*(4), F328–F331.

Cornes, M. P., Ford, C., & Gama, R. (2008). Spurious hyperkalemia due to EDTA contamination: common and not always easy to identify. *Annals of Clinical Biochemistry*, *45*, 601–603.

Fokker, M. (2014). Stability of glucose in plasma with different anticoagulants. *Clinical Chemistry and Laboratory Medicine*, *52*, 1057–1060.

Garza, D., & Becan-McBride, K. (2015). Capillary blood specimens. In D. Garza, & K. Becan-McBride (Eds.), *Phlebotomy handbook: blood specimen collection from basic to advanced* (9th ed.) (pp. 359–380). Upper Saddle River, NJ: Pearson Prentice Hall.

Garza, D., & Becan-McBride, K. (2015). Pediatric and geriatric procedures. In D. Garza, & K. Becan-McBride (Eds.), *Phlebotomy handbook: blood specimen collection from basic to advanced* (9th ed.) (pp. 405–439). Upper Saddle River, NJ: Pearson Prentice Hall.

Green, A. M. I., & Gray, J. (2003). *Neonatology & laboratory medicine*. London, UK: ACB Venture Publications.

Haverstick, D. M, Brill, L. B., Scott, M. G., et al. Preanalytical variables in measurement of free (ionized) calcium in lithium heparin-containing blood collection tubes. *Clinica Chimica Acta*, 403, 102–104.

Hoeltke, L. B. (2006). The challenge of phlebotomy. In L. B. Hoeltke (Ed.), *The complete textbook of phlebotomy* (3rd ed.) (pp. 227–248). Clifton Park, NJ: Thomas Delmar Learning.

Jones, P. M. (2010). Pediatric clinical biochemistry: why is it different? In D. J. Dietzen, M. J. Bennett, & E. C. C. Wong (Eds.), *Biochemical and molecular basis of pediatric disease* (4th ed.) (pp. 1–9). Washington DC: AACC Press.

Kiechle, F. L. (Ed.). (2005). *So you're going to collect a blood specimen: an introduction to phlebotomy* (11th ed.). Northfield, Ill: College of American Pathologists.

Mikesh, L. M., & Bruns, D. E. (2008). Stabilization of glucose in blood specimens: mechanism of delay in fluoride inhibition of glycolysis. *Clinical Chemistry*, *54*, 930–932.

Mullins, G. R., Harrison, J. H., & Bruns, D. E. (2016). Smartphone monitoring of pneumatic tube system-induced hemolysis. *Clinica Chimica Acta*, *462*, 1–5.

Pupek, A., Matthewson, B., Whitman, E., Fullarton, R., & Chen, Y. (2017). Comparison of pneumatic tube system with manual transport for routine chemistry, hematology, coagulation and blood gas tests. *Clinical Chemistry and Laboratory Medicine*, *55*(10), 1537–1544.

Ridefelt, P., Akerfeldt, T., & Helmersson-Karlqvist, J. (2014). Increased plasma glucose levels after change of recommendation from NaF to citrate blood collection tubes. *Clinical Biochemistry*, *47*, 625–628.

WHO Guidelines on Drawing Blood: Best practices in phlebotomy. Pediatric and neonatal blood sampling. (2010). World Health Organization (Chapter 6) <http://www.ncbi.nlm.nih.gov-/books/NBK138647/>.

Young, D. S. (2007). *Effects of preanalytical variable on clinical laboratory tests* (3rd ed.). Washington, DC: AACC Press.

Zhang, D. J., Elswick, R. K., Miller, W. G., & Bailey, J. L. (1998). Effect of serum-clot contact time on clinical chemistry laboratory results. *Clinical Chemistry*, *44*(6), 1325–1333.

7

Quality Management

*W. Greg Miller, Sverre Sandberg**

OBJECTIVES

1. Define the following:
   - Coefficient of variation
   - External quality control
   - Internal quality control
   - Levey-Jennings chart
   - Mean
   - Proficiency testing
   - Quality control plan
   - Quality control rules
   - Sigma metric
   - Standard deviation (SD)
   - SD interval
2. Explain how to establish a target value and SD for use with a quality control material in the following situations: a new lot of quality control material replaces an existing lot for an analyte; a new measurement procedure replaces an existing procedure for the same analyte; a new measurement procedure for a new analyte is introduced to the laboratory.
3. Explain how to establish rules for evaluating quality control sample results to confirm that results for patient samples from a measurement procedure are acceptable to make medical decisions.
4. Explain how to verify quality control (QC) target values following a reagent lot change.
5. Explain the components of a quality control plan and how the plan is established.
6. Explain the actions taken when a failed quality control result is obtained.
7. Explain how to evaluate results from external quality assessment (EQA) or proficiency testing.
8. Explain why peer group evaluation is used in EQA and what information a laboratory gets about its measurement procedure performance from the data.
9. Explain the limitations in using peer group mean values to assess the relationship of results among different measurement procedures.
10. Explain the value of EQA that uses commutable samples.
11. Explain how quality control information is documented and reviewed by laboratory staff.

KEY WORDS AND DEFINITIONS

Coefficient of variation Also called *relative standard deviation*; calculated as the standard deviation (SD) divided by the mean and multiplied by 100 to express in percent.

External quality assessment Also called *external quality control* or *proficiency testing;* an assessment process in which samples that simulate patient specimens are received from an external organization and results compared to those from other laboratories or from a reference method to determine that a measurement procedure's performance meets preestablished criteria to be suitable for use in medical decisions.

Internal quality control Term used to refer to quality control as defined here to confirm that results from a measurement procedure are suitable for use.

Levey-Jennings chart A graphical display with observed control values plotted on the *y*-axis and time (typically in days) shown on the *x*-axis. The *y*-axis shows the target value with plus and minus 1, 2, and 3 SDs indicated. Control limits can be indicated but are not practical to show on the chart when multiple control rules are used for evaluation of quality control results.

Mean The arithmetic average of a series of numbers such as results of repeated measurements of a quality control sample. The mean of a series of replicate results for a quality control material is used as the target value for future measurements of that quality control material.

Proficiency testing Another term for external quality assessment.

*Some of this material was previously published in McPherson, R. A., & Pincus, M. R. (Eds.), (2016). *Henry's clinical diagnosis and management by laboratory methods* (23rd ed.). St. Louis: Elsevier.

Quality assessment Protocols to confirm that the laboratory service meets the needs of medical providers who use laboratory results for medical decisions. Quality assessment includes preanalytical, analytical, and postanalytical components of quality.

Quality control Statistical and/or nonstatistical check protocols, typically using quality control samples that are surrogates for the patient samples being tested, to assess that results from a measurement procedure meet preestablished criteria to be suitable for use in medical decisions.

Quality control plan A document that describes the organization and operation of the internal quality control process.

Quality control rules A set of rules that results for one or more quality control samples must meet to make a decision that the measurement procedure performance is producing results for patient samples that are suitable for use in medical decisions.

Sigma metric A value that expresses the variation in performance of a measurement procedure relative to the allowable variability in results to be suitable for medical decisions expressed in SD units. Six-sigma performance means that six SDs of measurement procedure variation fit within the allowable limits for acceptable performance.

Standard deviation A statistical value that estimates the dispersion of replicate values around a mean value. A SD assumes a Gaussian distribution of values that is typically observed for numeric quality control results.

Standard deviation interval The number of SDs an individual result is from a mean value calculated as mean minus individual value divided by SD. The SD interval can be positive or negative relative to the mean value.

The purpose of a clinical laboratory test is to provide information on the pathophysiological condition of an individual patient to assist with diagnosis, therapy, or to assess risk for a disease. **Internal quality control**, frequently referred to as **quality control** (QC), is a statistical sampling approach performed by a laboratory on a regular schedule to assess performance of its measuring systems. External QC, frequently referred to as **external quality assessment** (EQA) or **proficiency testing** (PT) is an approach to assess measuring system performance by comparison of results for samples measured by a group of different laboratories. Both types of QC are addressed in this chapter (see later section for EQA/PT).

Internal QC evaluates a measurement procedure by periodically measuring a QC sample for which the expected result is known in advance. If the result for a QC sample is within acceptable limits of the known value, the measurement procedure is verified to be performing as expected, and results for patient samples can be reported with high probability that they are suitable for clinical use. If a QC result is not within acceptable limits, the measurement procedure is not performing correctly, there is a high probability that results for patient samples are not suitable for clinical use, and corrective action is necessary. Patient sample measurements could need to be repeated when the measurement procedure has been restored to its stable performance condition. If erroneous results have already been reported before an error condition is identified, a corrected report must be issued.

Measurement procedures fall into one of two general categories from a **QC plan** perspective. One type of procedure is a "batch" measurement process in which the results for patient samples and QC samples are completed before the results are reported. The other type of procedure is a "continuous" measurement process in which patient sample results are reported during the interval between QC samples measurements. For continuous measurement procedures, there is a possibility that erroneous results have already been reported if an error condition is identified by the next QC samples measurements. In either category, QC procedures only identify error conditions present at the point in time when a QC sample is measured.

The design of a QC plan must consider the analytical performance capability of a measurement procedure and the risk of harm to a patient that might occur if an erroneous laboratory test result is used for a clinical care decision.

POINTS TO REMEMBER

Internal Quality Control

- The primary role of internal QC is to confirm that patient results are correct and to identify possible errors before they will affect patient care decisions.

MEASUREMENT PROCEDURE PERFORMANCE AS A PREREQUISITE FOR A QUALITY CONTROL PLAN

Analytical Bias and Imprecision

Fig. 7.1 illustrates the meaning of bias and imprecision for a measurement procedure. The horizontal axis represents the numeric value for an individual result, and the vertical axis represents the number of repeated measurements with the same value made on samples of a QC material. The *blue line* shows the dispersion of results for repeated measurements of the same QC material, which is the random imprecision of the measurement. The **standard deviation** (SD) is a measure of expected imprecision in a measurement procedure when it is performing within specifications. The **mean** of repeated measurements of a QC sample becomes the expected value for that QC sample.

Fig. 7.1B illustrates that if a systematic bias (error) occurs in the measurements, the mean value shifts to another value. Note that the imprecision is the same as before the bias occurred because it is unlikely, although not impossible, that a change in imprecision would occur at the same time as a bias shift. The primary purpose of measuring QC samples is to statistically evaluate the measurement procedure to

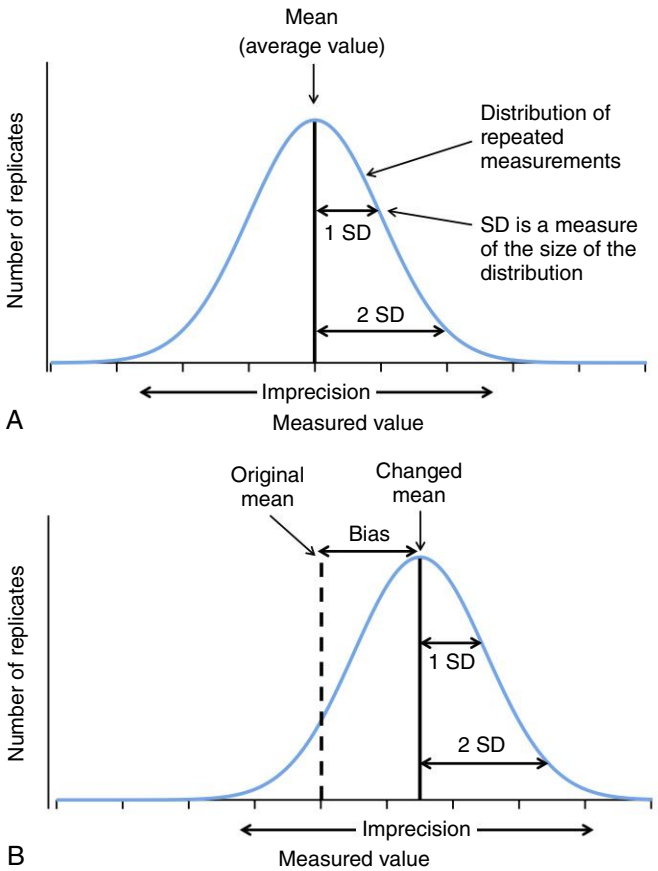


Fig. 7.1 (A) Distribution of results showing the mean value and expected imprecision for repeated measurements of a quality control sample. (B) Bias when a change in calibration has occurred. *SD*, Standard deviation. (From Miller, W. G. [2016]. Quality control. In: *Henry's clinical diagnosis and management by laboratory methods* [23rd ed.]. Philadelphia: Elsevier.)

verify that it continues to perform within the specifications consistent with its acceptable expected stable condition or to identify that a change in performance occurred that needs to be corrected. QC result acceptance criteria are based on the probability for an individual QC result to be different from the variability in results expected when the measurement procedure is performing in a stable condition within its specifications.

Fig. 7.2 shows a **Levey-Jennings chart** that is the most common presentation for evaluating QC results. This format shows each QC result sequentially over time and allows a quick visual assessment of performance. Assuming the measurement procedure is performing in a stable condition, the mean value represents the target (or expected) value for the QC result, and the SD lines represent the expected imprecision. Assuming a Gaussian (normal) distribution of imprecision, the results should be distributed uniformly around the mean with results observed more frequently closer to the mean than near the extremes of the distribution. The data show fluctuations in performance at different intervals and illustrate the importance of calculating the SD over a long enough interval to appropriately include all sources of variability in the measurement procedure. Note that a small number of results in Fig. 7.2 are greater than 2 SDs, and three results slightly exceed 3 SDs, which is expected for an approximately Gaussian distribution of imprecision. The number of results expected within the **standard deviation intervals** (SDIs) is:

- ±1 SD = 68.3% of observations
- ±2 SD = 95.4% of observations
- ±3 SD = 99.7% of observations

Interpretation of an individual QC result is based on its probability to be part of the expected distribution of results

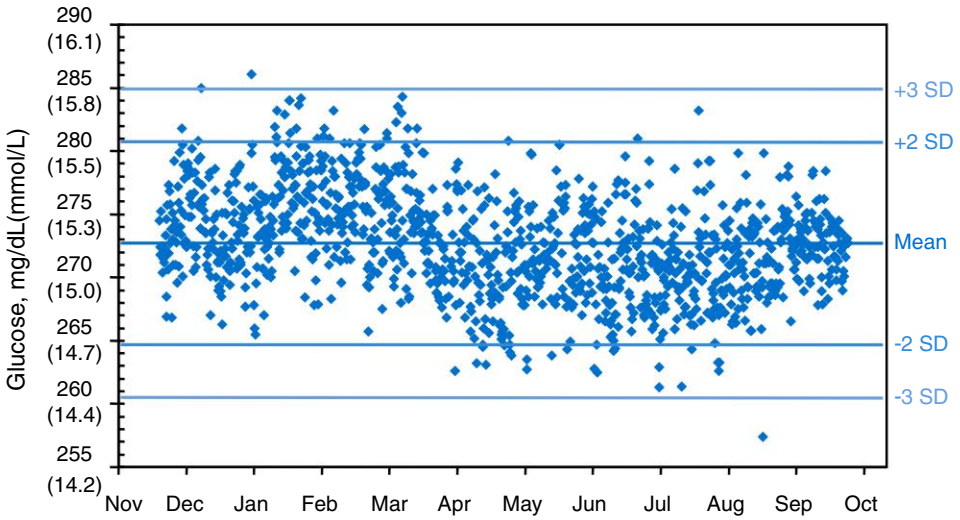


Fig. 7.2 Levey-Jennings chart of quality control (QC) results (n = 1232) for a single lot of QC material used over a 10-month period. *SD*, Standard deviation. (Reprinted with permission from Miller, W. G., & Nichols, J. H. [2011]. Quality control. In: W. A. Clarke [Ed.], *Contemporary practice in clinical chemistry* [2nd ed.]. Washington, DC: AACC Press.)

for the measurement procedure when the procedure is performing correctly. Note that evaluation of individual QC results may be performed by computer algorithms without visual examination of a Levey-Jennings chart.

Performance of a Measurement Procedure for Its Intended Medical Use

It is necessary to determine how the performance of a measurement procedure relates to the medical requirements for interpreting results to determine the frequency to measure QC samples and the criteria to use to evaluate the QC results. The **sigma metric** is commonly used to assess how well a measurement procedure performs relative to the medical requirement. For laboratory measurements, the sigma metric is calculated as:

$$\text{Sigma} = \frac{(TE_a - |\text{bias}|)}{SD}$$

where TE_a is the total error allowed based on medical requirements, and absolute value of bias and SD refer to performance characteristics of the measurement procedure. The SD is estimated from the QC data. It is critically important that the estimate of SD be made using QC data that represent all or most components of variability that occur over an extended time period. The bias is difficult for a laboratory to estimate because it is difficult to evaluate if a particular measurement procedure has a bias compared with a reliable estimate of a true value such as a reference measurement procedure. For internal QC, a laboratory is usually interested to determine if a bias has occurred compared with the condition established by calibration of a measurement procedure, and if a bias has occurred it will be corrected. Consequently, the bias is usually assumed to be zero for calculating sigma.

TE_a represents the measurement procedure performance required for medical decisions based on a test result. TE_a can be estimated using three models. The preferred model (model 1) is to set a performance specification based on an outcome study such as the impact of analytical performance on the clinical outcome. Outcome studies are only available for a small number of analytes and typically included in clinical laboratory practice guidelines.

Model 2 bases the TE_a on a fraction of the within and between individual biological variations of the measurand. (For additional information on biological variation refer to Chapter 4.) This model minimizes the ratio of the "analytical noise" to the "biologic signal," with an assumption that a small ratio will identify measurement procedure performance that relates to the medical requirements. Tables of optimal, desirable, and minimal TE_a based on biological variation are available.

Model 3 bases the performance specifications on the "state of the art," which is the performance capability of a measurement procedure usually derived from QC data. The laboratory should consult with clinical care providers to agree on an appropriate TE_a for the patient population served.

Because sigma assumes a Gaussian or normal distribution for repeated measurements, the probability of a defect (i.e., an erroneous laboratory result) can be predicted. The term *six-sigma* refers to a condition when the variability in the measurement process is sufficiently smaller than the medical requirement that erroneous results are very uncommon. Fig. 7.3A shows a six-sigma test that has the TE_a limits 6 SDs away from the center point of the distribution of variability in measurements. A small amount of bias or increased imprecision will have little influence on the number of erroneous results produced, and the risk of producing an erroneous result even with some loss of performance is very low. Consequently, less stringent QC is suitable.

Fig. 7.3B shows a three-sigma measurement procedure that has the TE_a limits 3 SDs away from the center point of the expected distribution of variability in measurements. In the three-sigma situation, a small amount of bias or increased imprecision will cause the number of erroneous results to increase substantially. More frequent QC and more stringent acceptance criteria will allow the laboratory to identify when small changes in performance occur so they can be corrected to minimize the risk of harm to a patient from erroneous results being acted on to make clinical care decisions.

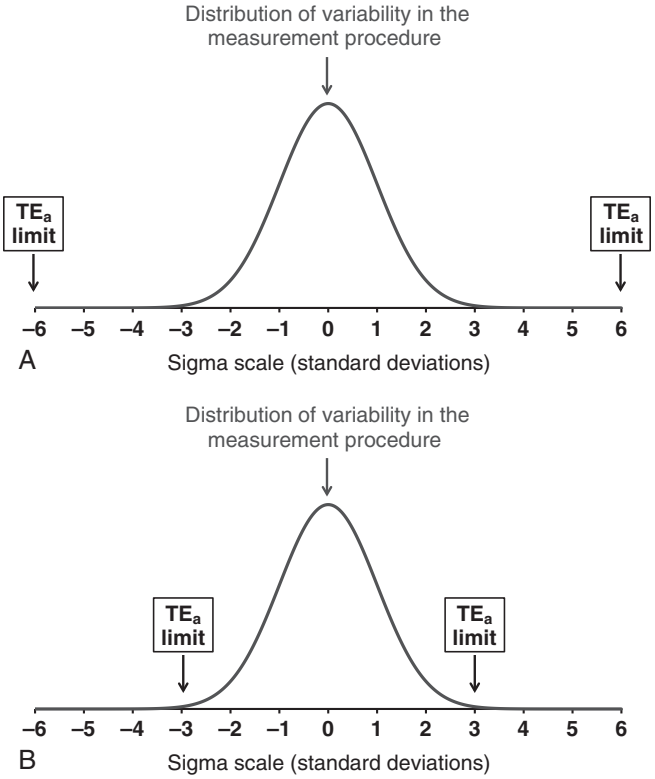


Fig. 7.3 Test performance relative to the sigma scale to describe how well performance meets medical requirements expressed as the allowable total error *(TEa)*. (A) A six-sigma measurement procedure. (B) A three-sigma measurement procedure. (From Miller, W. G. [2016]. Quality control. In: *Henry's clinical diagnosis and management by laboratory methods* [23rd ed.]. Philadelphia: Elsevier.)

POINTS TO REMEMBER

- The performance characteristics of a measurement procedure when it is performing in a stable in-control condition must be known.
- The allowable total error for a measurement procedure must be established based on requirements for using a laboratory result in patient care decisions.
- The sigma metric represents the probability that a given number of erroneous results may occur when the test measurement procedure is performing to its specifications.

DEVELOPING A QUALITY CONTROL PLAN AND IMPLEMENTING INTERNAL QUALITY CONTROL PROCEDURES

Selection of Quality Control Materials

In general, two different concentrations of QC materials are necessary for adequate statistical QC. For quantitative measurement procedures, QC materials should be selected to provide analyte concentrations that represent clinical decision values over the analytical measuring interval of the measurement procedure. More than two concentrations of QC materials may be needed when there are several important medical decision values. In practice, laboratories are frequently limited by concentrations available in commercial QC products. For procedures with extraction or other pretreatment steps, controls must be used to include any pretreatment steps. For qualitative tests, QC values should assess the stability of a threshold for a classification decision.

The QC materials must be manufactured to provide a stable product that can be used for an extended time period when possible. Use of a single lot for a year or more allows reliable interpretive criteria to be established that will minimize the effort and uncertainty associated with QC lot changes.

Limitations of Quality Control Materials

An important limitation of most QC materials, which also applies to EQA or PT materials, is called noncommutability with patient samples. Fig. 7.4A shows that a commutable QC material gives a result that closely agrees with results for authentic patient samples with the same amount of analyte when measured by different procedures. Fig. 7.4B shows that results for a noncommutable QC material have a different relationship than observed for patient samples when measured by different procedures. QC and EQA/PT materials are typically noncommutable with patient samples because the serum or other biologic fluid matrix is usually altered from that of a patient sample during product manufacturing, for example, by use of partially purified human and nonhuman additives to achieve desired concentrations and various stabilization additives and processes. The impact of the matrix alteration on the recovery of a measurand is not predictable and is frequently different for different lots of QC material, for different lots of reagent within a given measurement procedure, and for different measurement procedures. Because of the noncommutability limitation, special procedures are required when changing lots of reagent or comparing QC results among two or more measurement procedures.

A second limitation of QC materials is gradual deterioration of the analyte during storage and after reconstitution, thawing, or vial opening.

Frequency to Measure Quality Control Samples

The frequency to measure QC samples depends on several considerations. Regulatory requirements may specify a minimum frequency for QC measurements. The instructions for use from measurement procedure manufacturers may specify a minimum or recommended QC frequency. The risk of action being taken before a measurement error of clinical importance is detected is an important consideration for more frequent QC measurements than based on regulatory requirements or manufacturer's recommendations.

Analytical Stability of the Measurement Procedure

The more stable the measurement procedure, the less frequently a QC evaluation needs to be performed. Some

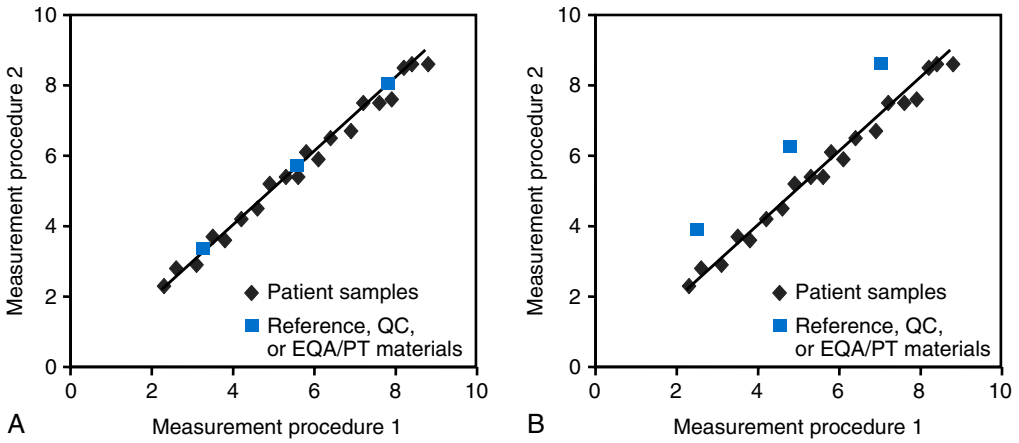


Fig. 7.4 Illustration of commutable and noncommutable materials. (A) Commutable materials *(blue squares)* have the same relationship between two measurement procedures as observed for patient samples *(black diamonds)*. (B) Noncommutable materials have a different relationship than observed for patient samples. *EQA,* External quality assessment; *PT,* proficiency testing; *QC,* quality control.

measurement procedures, particularly point-of-care (POC) devices, have been designed with sophisticated built-in control procedures to mitigate the risk that an erroneous result may be produced. These measurement systems may be sufficiently stable and self-monitored to justify reduced frequency of traditional QC sample testing.

Risk of Harm to a Patient and Number of Patients Who May be at Risk

More frequent QC sampling is appropriate to avoid the situation of discovering a measurement procedure defect many hours after a physician has made a clinical treatment or nontreatment decision based on an erroneous result. For example, QC sampling performed on a 24-hour cycle might be performed at 9 a.m. If QC results indicate a measurement procedure problem, the erroneous condition could have started at any time during the previous 24 hours. If the problem had occurred at 3 p.m. the previous day, erroneous results could have been reported for 18 hours, likely putting a large number of patients at risk of an inappropriate care decision.

The Clinical and Laboratory Standards Institute (CLSI) has published guideline EP23 addressing risk-based QC procedures. The document provides guidance on how to develop a QC plan based on evaluation of risk of harm to a patient and assessment of the effectiveness of risk mitigation procedures. In general terms the laboratory director makes a judgment that suitable built-in and laboratory-applied controls are in place and that a result has a high probability to be correct at the time it is reported for clinical use.

Event-Based Quality Control Sample Measurement

It is necessary when using a continuous measurement system to measure QC samples before and after scheduled events such as recalibration or maintenance that may alter the current performance condition. Each of these operations is intended to restore the measurement conditions to optimal specifications and to correct for any calibration drift or component deterioration that may have occurred. If QC samples are not measured before such scheduled events, a laboratory will not know if an error condition may have occurred since the last time QC samples were measured. It is also necessary to measure QC samples after these events to verify that the operations were performed correctly and that measurement procedure performance meets specifications before restarting to measure patient samples.

Establishing the Quality Control Target Value and Standard Deviation That Represent a Stable Measurement Operating Condition

QC target values and acceptable performance limits are established to optimize the probability to detect a measurement defect that is large enough to have an impact on clinical care while minimizing the frequency of "false alerts" caused by statistical limitations of the criteria used to evaluate QC results.

Quality Control Material Target Value

The generally accepted minimum protocol for target value assignment is to use the mean from a minimum of 10 measurements of the QC material on 10 different days when the measurement procedure is correctly calibrated and performing to its specifications. Because all sources of variability cannot be captured in 10 measurements, it is recommended to update the target value after more data have been acquired during use of the QC material. If a 10-day protocol is not possible (e.g., if an emergency replacement of a lot of QC material is necessary), a provisional target value can be established with fewer data but should be updated when additional QC results are available.

Quality Control Material Standard Deviation

The SD must represent the variability expected for a measurement procedure over an extended time interval to include all sources of variability when its performance meets its specifications. Measurement variability has short time interval components (such as pipetting volume), gradual changes (such as pipet seal deterioration or coating of cuvette or electrode surfaces), or variable and sometimes long intervals (such as calibration cycles, reagent replenishment, and maintenance procedures). An SD that represents stable measurement performance can usually be estimated from the cumulative SD over a 6- to 12-month period for a single lot of QC material because most expected sources of variation are likely to be represented. Fig. 7.5 illustrates the fluctuation in SD that occurred when calculated for monthly intervals compared with the relatively stable value observed for the cumulative SD after a period of 6 months. Note that the cumulative SD is not the average of the monthly values but is the SD determined from all individual results obtained over a time interval since the lot of QC material was first used. If the imprecision expected during normal stable operation is underestimated, the acceptable range for QC results will be too small, and the false-alert rate will be unacceptably high.

When a measurement procedure has been established in a laboratory and a new lot of QC material is being introduced,

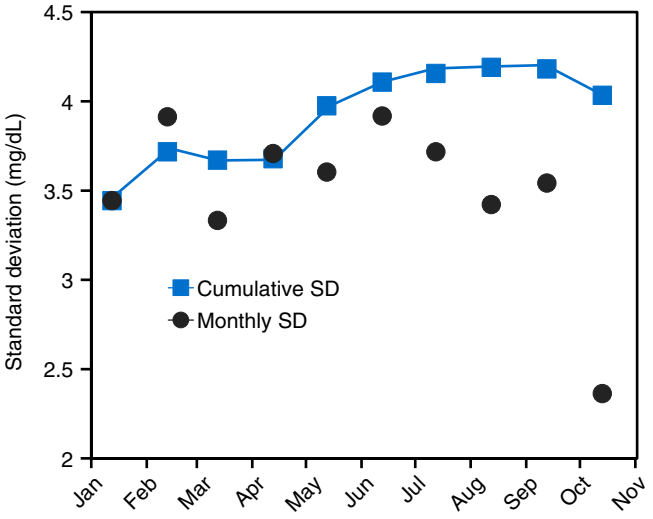


Fig. 7.5 Cumulative standard deviation *(SD)* versus single monthly values calculated from the data in Fig. 7.2. (From Miller, W. G. [2016]. Quality control. In: *Henry's clinical diagnosis and management by laboratory methods* [23rd ed.]. Philadelphia: Elsevier.)

TABLE 7.1 Abbreviation Nomenclature for Quality Control Evaluation Rules

Rule	Meaning	Detects
1_{2S}	One observation exceeds 2 SDs from the target value. The 1_{2S} rule is not recommended except for low sigma measurement procedures because it has a high false-alert rate.	Bias or imprecision
1_{3S}	One observation exceeds 3 SDs from the target value.	Bias or large imprecision
2_{2S} ($2_{2.5S}$)	Two sequential observations, or observations for two QC samples measured at approximately the same time, exceed 2 SDs (or 2.5 SDs) from the target value in the same direction.	Bias
2 of 3_{2S}	Two observations for three QC samples measured at approximately the same time exceed 2 SDs from the target value in the same direction. Note that this type of rule is used when three QC materials are used for a measurement procedure.	Bias
R_{4S}	Range between observations for two QC samples measured at approximately the same time, or for two sequential observations of the same QC sample, exceeds 4 SDs.	Imprecision
10_x or 10_m	Ten sequential observations for the same QC sample are on the same side of the target value (x or mean). The 10_x rule is not recommended because it has an excessive false-alert rate.	Not recommended
8_{1S} ($8_{1.5S}$)	Eight sequential observations for the same QC sample exceed 1 SD (or 1.5 SD) in the same direction from the target value.	Bias trend
CUSUM	CUSUM of SDI for the current and previous results.	Bias trend
EWMA	EWMA for the current and previous results with newer results having more influence (weight).	Bias trend

CUSUM, Cumulative sum; *EWMA*, exponentially weighted moving averages; *QC*, quality control; *SD*, standard deviation; *SDI*, standard deviation interval.
From Miller, W. G. (2016). Quality control. In: *Henry's clinical diagnosis and management by laboratory methods* (23rd ed.). Philadelphia: Elsevier.

the target value for the new lot of QC material is used along with the well-established SD from the previous lot. This practice is appropriate because in most cases, measurement imprecision is a property of the measurement procedure and equipment used and is unlikely to change with a different lot of QC material.

When a new measurement procedure replaces an existing procedure, the SD for the existing procedure can in many cases be used as the initial SD for the new measurement procedure until the SD for the new measurement procedure has been established. An assumption is made that the SD for the existing measurement procedure was appropriate to ensure the results were suitable for use in medical decisions, consequently that SD is likely to be suitable for QC decisions for the new measurement procedure.

If a measurement procedure for a new analyte is introduced, there will not be any historical performance information. The SD is based on QC data obtained during the measurement procedure validation. A minimum of 20 observations on different days is recommended for the initial estimate of the SD. This initial estimate of SD will likely be an underestimate because it does not include all sources of variability and must be updated when sufficient QC results have been accumulated to include most sources of variability.

Quality Control Materials With Preassigned Values

Some QC materials are provided by the measurement procedure manufacturer with preassigned target values and acceptable ranges intended to confirm that the measurement procedure meets the manufacturer's specifications. Such assigned values may be used to verify the manufacturer's specifications. However, it is recommended that both the target value and the SD should be reevaluated and assigned by the laboratory after adequate replicate results have been obtained because the QC interpretive rules used in a single laboratory should reflect performance for the measurement procedure in that laboratory.

QC materials with assigned target values and SDs are also available from third-party manufacturers (i.e., manufacturers not affiliated with the measurement procedure manufacturer) and typically have values that are applicable to specifically stated measurement procedures and reagent lots to accommodate the influence of noncommutability. However, a laboratory should determine a target value and SD that reflect the operating conditions in that laboratory to ensure optimal assessment of QC results to identify an erroneous measurement condition.

Establishing Rules to Evaluate Quality Control Results

The acceptable range and rules for interpretation of QC results are based on the probability of detecting an analytical error condition with an acceptably small false-alert rate.

The conventional way to express QC interpretive rules is by using an abbreviation nomenclature popularized among clinical laboratories by Westgard and summarized in Table 7.1. Note that fractional SDIs can be used as in the $2_{2.5S}$ and $8_{1.5S}$ examples and that combinations of numbers of controls and limits can be used as appropriate for QC interpretive rules. Trend detection procedures such as cumulative sum (CUSUM) or exponentially weighted moving average (EWMA) are recommended if supported by an available computer system because they are more powerful for detecting trends than approaches based on counting the number of sequential observations exceeding a specified SDI. A trend rule can be set to give an alert as a warning that may not require discontinuing testing but indicate that a problem is developing that should be investigated.

TABLE 7.2 Empirical Multirule for the Quality Control Data Presented in Fig. 7.2

Multirule Components	Type of Variability Detected
1_{3S}	Imprecision or bias
$2_{2.5S}$	Bias
R_{4S}	Imprecision
$8_{1.5S}$	Bias trend

From Miller, W. G. (2016). Quality control. In: *Henry's clinical diagnosis and management by laboratory methods* (23rd ed.). Philadelphia: Elsevier.

It is recommended to improve the efficiency of QC interpretive rules by combining two or more rules and applying them simultaneously as multirule criteria. For example, the $1_{3S}/2_{2S}$ multirule identifies an error condition if one control exceeds ±3 SD from the target value or if two controls exceed ±2 SDs in the same direction from the target value. The $1_{3S}/2_{2S}$ multirule has a low false-alert rate but improved probability to detect a bias error of a given magnitude.

In practice, empirical judgment is frequently used to establish acceptance criteria (rules) to evaluate QC results based on data acquired over a long enough period of time to adequately estimate the expected variability when a measurement procedure is working correctly. An empirical approach can be used by obtaining a set of QC data that represents stable measurement procedure performance over a time interval expected to include most sources of variability. Using those data, the false-alert rate for a rule can be determined, and bias errors of different magnitudes can be added to estimate the ability of a rule, or a combination of rules, to identify that error. Table 7.2 gives an example of an empirically developed multirule for a 6-sigma measurement procedure. Such control rules should allow the laboratory to detect errors before they are of a magnitude that will affect clinical decisions. A 10_x rule was not used because it would have increased the false-alert rate by 10.6%. A 10_x rule or other rule that counts the number of sequential QC results on one side of the target value is not recommended because this condition typically does not indicate a problem with clinical interpretation of patient results when the magnitude of the difference from the target value is small. Counting the number of sequential results that exceed a larger SD from the target value, such as $8_{1.5S}$ in this example, is more likely to represent a measurement condition that might need investigation.

For measurement procedures with small sigma values, small deviations from the expected performance need to be identified. Consequently, more stringent QC practices need to be used such as selecting rules such as 1_{2S} that will give an alert at smaller error conditions, using additional rules in a multirule set, measuring QC more frequently, using more than two QC samples, and not releasing patient results until QC assessment is complete for the time interval during which patient samples were measured. More stringent **QC rules** will have more false alerts, but this is an unavoidable cost when lower sigma measurement procedures are used.

Specifying the Quality Control Plan

The preceding subsections describe the considerations for each component in a QC plan. The laboratory director is responsible for considering the components, making judgments regarding the considerations, and approving the final plan for each analyte measured in a laboratory. A plan for internal QC specifies the following components:

- The number of QC samples to be measured and the approximate concentrations of analytes in those controls
- The target value for each QC sample
- The SD for each QC sample to be used in the QC rules
- The rules for evaluating the QC results
- The frequency to test the QC samples

Considerations for Point-of-Care Testing

Internal QC of POC instruments offer extra challenges compared with those addressed at the central laboratory. The main reasons are that POC instruments are often operated by persons without laboratory training; they often use methodologic principles that are different from those in the central laboratory; they often have "built-in" controls; and the number of measurements can be small, making the use of QC samples in the traditional way expensive. In cartridge-based and some strip-based instruments, the manufacturer often has placed the technology in the cartridge or strip together with QCs, and in some cases, QC rules are built in so that patient results cannot be reported unless the QC is "satisfactory." The instrument is then merely an electronic reader that often has incorporated an "electronic quality control" that verifies the electronics of the measurement procedure. The electronic instrument checks do not verify the reagents in the cartridges or strips, and unless each cartridge has internal QC materials, the reagent cartridges or strips should be checked at delivery and then at intervals (e.g., with the arrival of a new shipment or lot or at a suitable interval such as weekly or monthly depending on the cartridges/strips).

Not all POC instruments include enhanced QC features. In these cases, one must rely on daily liquid QC performed by the operator. The limitation of using the liquid QC sample in this situation is that it only checks if one disposable cartridge or strip meets the performance specifications. This limitation requires an assumption that all devices in a lot were manufactured uniformly and will perform equivalently.

How internal QC should be performed and supervised also depend on the location of the POC instruments. In a hospital, it is now possible with real-time bidirectional connectivity between the POC devices and the central laboratory to transfer both patient and QC results and to set lock-out parameters for conformance to a QC protocol. As technology advances, the general trend is for more sophisticated POC devices with built-in control systems to be incorporated to minimize or prevent the possibility for an incorrect result.

Corrective Action When a Quality Control Result Indicates a Measurement Problem

A QC alert occurs when a QC result fails an evaluation rule, which indicates that an analytical problem may exist. A QC

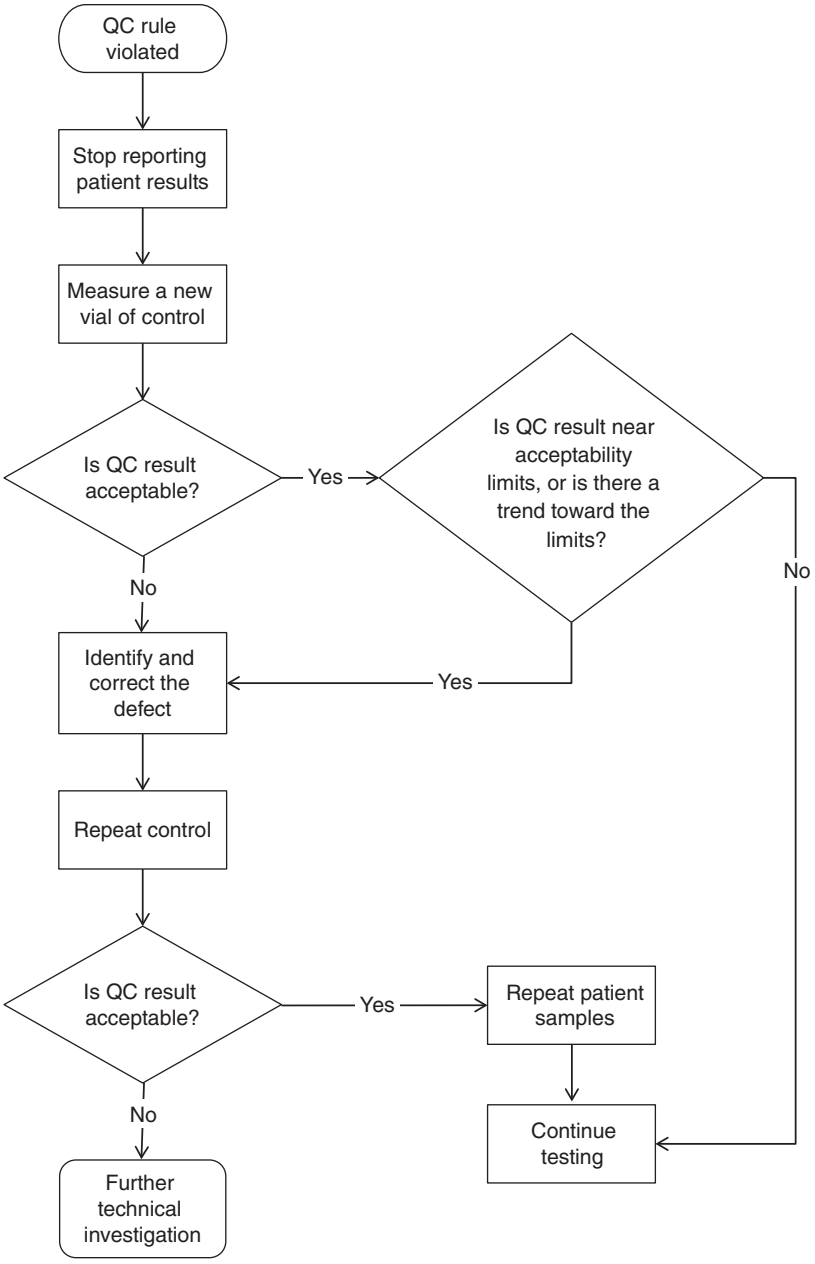


Fig. 7.6 Generalized troubleshooting sequence showing the initial steps after an unacceptable quality control *(QC)* result. The details of troubleshooting the defect may be different for different rules violations or if more than one unacceptable QC result was obtained.

alert means there is a high probability that the measurement procedure is producing results that have the potential to be unreliable for patient care and testing must be stopped until the problem is resolved. Fig. 7.6 presents a generalized troubleshooting sequence. QC materials can deteriorate after opening because of improper handling and storage or because of unstable analytes. Thus repeating the measurement on a new vial of the QC material is a useful step to determine if the alert was caused by deteriorated QC material rather than by a measurement procedure problem. In this situation, if the result for the new QC sample is acceptable, testing of patient samples can resume. One caution when the repeat QC result is near acceptability limits is to consider whether the repeat and original results are essentially the same. It is not acceptable to repeat the QC until a value happens to be just within the acceptable limit. In this situation the probability is high that a measurement problem exists, and this possibility should be investigated. In addition, current and preceding QC results should be examined for a trend in bias that indicates a measurement issue that needs to be corrected. These precautions in evaluating repeat results for a new QC sample can be challenging or impossible for automated evaluation by computer systems, thus requiring the laboratory technologist to be vigilant in reviewing results.

When repeat testing of a new QC sample does not resolve the alert situation, the instrument and reagents should be

inspected for component deterioration, empty reagent containers, mechanical problems, and so on. In many cases, it will be necessary to recalibrate. When the problem is identified and corrected, QC samples should be measured to verify the correction, and all patient samples since the time of the last acceptable QC results, or the time when the error condition occurred, should be measured again. The laboratory director must establish acceptable criteria to determine if the repeat results agree adequately to permit reporting of original results without issuing a corrected report. Otherwise, corrected results must be reported. The criteria for acceptability of repeated tests are based mainly on the TE_a described earlier with consideration of the measurement procedure performance characteristics.

It may be difficult to establish the time when an error condition occurred. One approach is to repeat every few samples back to the time of the last acceptable QC results. The repeated results are then compared with acceptable criteria for repeated results agreement to identify a point in time when the error condition occurred. When selecting the samples to repeat, it is important to ensure that a substantial representation of the potentially erroneous samples is repeated and that samples at a concentration consistent with that of the unacceptable QC are represented. Alternatively, groups of 10 patient samples can be repeated, again ensuring that samples at a concentration consistent with that of the unacceptable QC are represented, until all repeat results in at least two sequential groups are within acceptable criteria for repeated results. When the point at which the error condition was likely to have occurred is identified, all patient samples must be repeated from that point until the unacceptable QC result was obtained. Any assessment of the point at which an error condition occurred by repeating selected patient samples has a risk to incorrectly identify that point, and laboratories are encouraged to repeat enough patient samples to have confidence in the assessment.

Verifying Quality Control Evaluation Parameters After a Reagent Lot Change

Changing reagent lots can cause a shift in QC results when there is no change in results for patient samples. Because the matrix-related noncommutability between a QC material and a reagent can change with a different reagent lot, QC results are not a reliable indicator of a measurement procedure's performance for patient samples after a reagent lot change. In the example in Fig. 7.7, QC values for the high-concentration control shifted after the change to a new lot of reagents, but there was no change in results for the low control. A comparison of results for a panel of patient samples assayed using the new and old reagent lots showed equivalent results. Consequently, the change in QC values for the high-concentration material was due to a difference in matrix-related noncommutability bias between the QC material and each of the reagent lots.

It is necessary to use patient samples to verify the consistency of results between old and new lots of reagents because of the unpredictability of a matrix-related noncommutability bias being present for QC materials. Fig. 7.8 presents a protocol

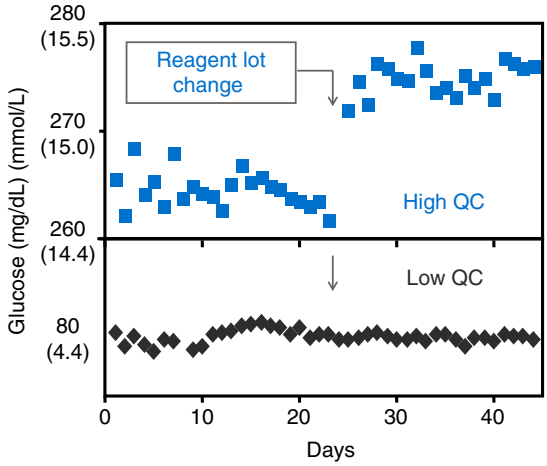


Fig. 7.7 Levey-Jennings plot showing impact of a reagent lot change on matrix bias with quality control *(QC)* samples. (Modified with permission from Miller, W. G., & Nichols, J. H. [2010]. Quality control. In: W. A. Clarke [Ed.], *Contemporary practice in clinical chemistry* [2nd ed.]. Washington, DC: AACC Press.)

to verify or adjust QC material target values after a reagent lot change. A group of patient samples and the QC samples are measured using both the current (old) and new reagent lots. The first step is to verify that results for a group of patient samples measured with the new reagent lot are consistent with results from the current (old) lot. The patient sample results, not the QC results, provide the basis for verifying that the new reagent lot is acceptable for use. If a problem is identified, the calibration of the new reagent lot must be investigated and corrected, or the new reagent lot may be defective and should not be used. When evaluating the patient results, keep in mind that the calibration of the old reagent lot may have drifted and should be verified before concluding that the new reagent lot is not giving acceptable results for the patient samples.

The number of patient samples to use for verifying the performance of a new reagent lot will depend on the measuring interval, the imprecision of a measurement procedure, and the concentrations at which clinical decisions are made. CLSI document EP26 recommends a minimum of three patient samples and more patient samples depending on the number of important clinical decision concentrations and the imprecision of a measurement procedure. This CLSI guideline includes a statistical analysis to determine if a difference in patient results is less than a critical difference that would represent risk for an inappropriate patient care decision based on a particular laboratory test result. An alternate approach is to select 5 to 10 patient samples that span the measuring interval and use a difference plot to evaluate average performance over the interval of concentrations represented by the patient samples. The laboratory must establish acceptance criteria for the agreement of patient results for old and new lot measurements consistent with the relatively small number of samples used, the analytical performance characteristics of a measurement procedure, and the clinical requirements for interpreting results.

When the results for patients are acceptable, the second step in Fig. 7.8 evaluates results for each QC material to

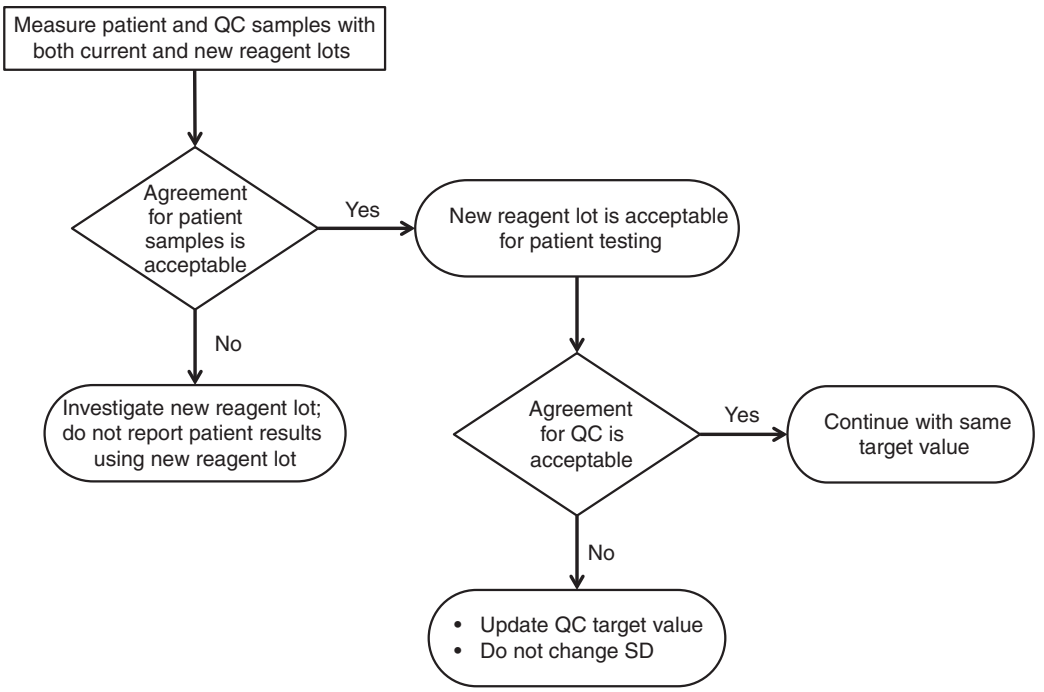


Fig. 7.8 Process for assessment of potential matrix-related noncommutability bias on quality control *(QC)* samples after a reagent lot change. *SD,* Standard deviation. (From Miller, W. G. [2016]. Quality control. In: *Henry's clinical diagnosis and management by laboratory methods* [23rd ed.]. Philadelphia: Elsevier.)

determine if its target value is correct for use with the new lot of reagent(s). If the target value has changed, it must be adjusted to correct for the change in matrix-related noncommutability bias between old and new lots of reagent(s). This adjustment keeps the expected variability centered around the QC target value so that QC interpretive rules will remain valid.

Failure to make a target value adjustment will cause subsequent QC results to be evaluated incorrectly, as illustrated in Fig. 7.9. The shift in target value would cause some of the results shown by *blue squares* to exceed the old upper QC rule limit, when in reality there is no defect in patient results because the increase in QC results is caused by the matrix-related noncommutability bias with the new reagent lot. Similarly, the increased magnitude of the gap to the old lower QC rule limit will permit a low bias condition, as shown by the *blue square* points at sequence number 29 and 30, to be undetected. In most cases the SD for a QC material will be the same with any lot of reagent(s).

Note that a reagent lot induced matrix-related noncommutability bias change in the numeric values for the QC results will cause an incorrect increase in the cumulative SD if all results are used for the calculation. For this reason, it is recommended to use the cumulative SD from a single reagent lot or the pooled SD from more than one reagent lot when determining the SD to use for interpreting QC rules.

Experience in clinical laboratories has shown that there are changes, other than reagent lot changes, in measurement procedures that can also affect the QC values but not the results for patient samples. Such changes could be caused by instrument component replacement or other causes. In theory, there should be an assignable cause for such effects, but such a cause is not always identifiable. In practice, any condition that affects QC results but does not affect patient results is treated in the same manner as described for reagent lot changes. The important QC principle is that if the results for patient samples are consistent between the two conditions, then the target value for the QC sample should be adjusted, if necessary, to reflect its value under the new condition. Failure to adjust the QC target value will cause inappropriate acceptability criteria to be used for evaluating the QC results.

Review of Quality Control Data and the Effectiveness of the Quality Control Plan

The immediate use of QC data is to determine if the results for patient samples can be reported for use in clinical care decisions, as described in the preceding sections. In addition, QC data must be reviewed by laboratory management on a regular schedule. Typical review schedules are weekly by senior technologists or supervisors and at least monthly by the laboratory director. However, the laboratory director or supervisor should promptly review items such as reagent or calibrator lot change validations, changes in QC target values associated with reagent lot or other changes, EQA/PT results review, and other occurrences that may affect quality of the laboratory results.

The weekly review process should determine that correct follow-up of any QC alerts was conducted, that all patient samples that may have had erroneous results were repeated, that any corrected reports were issued, and that the process was properly documented in QC records. The monthly review should include any issues identified by the weekly

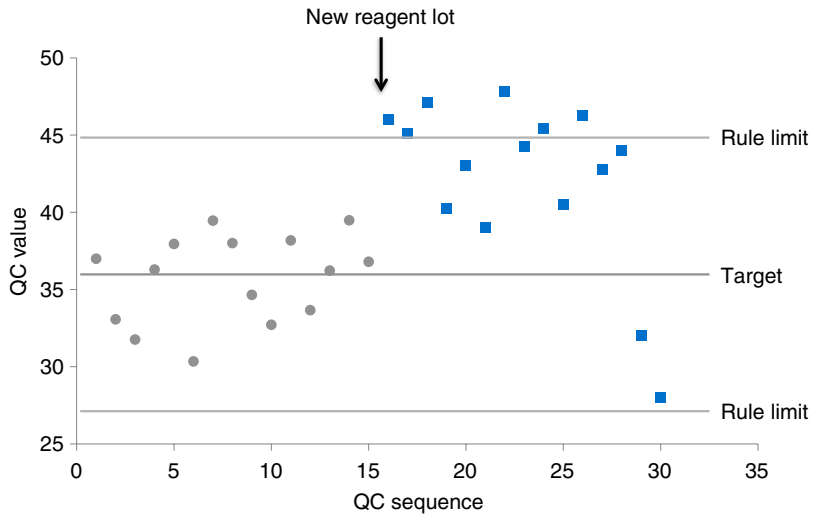


Fig. 7.9 Illustration of the influence on the failure rate for a quality control *(QC)* rule when failing to adjust the target value for a matrix-related shift. QC results before a new reagent lot are shown as *gray circles* and after the new reagent lot as *blue squares*. (From Miller, W. G. [2016]. Quality control. In: *Henry's clinical diagnosis and management by laboratory methods* [23rd ed.]. Philadelphia: Elsevier.)

review process, as well as examination of the Levey-Jennings chart or a computer-based report, to identify trends or changes in assay performance that may need to be addressed before they have effects on clinical care decisions. Note that automated systems to assist in the review of QC data are acceptable, and individual Levey-Jennings charts do not need to be examined every month. The monthly review should also include any adjustments made to QC parameters or the QC plan for a measurement procedure during the month.

POINTS TO REMEMBER

- Internal QC samples are measured along with patient samples.
- The target value and SD expected for a QC sample are established by the laboratory.
- Results from QC samples are evaluated using interpretive rules that are established after considering the probability for false alerts and the probability for detecting errors that represent a risk of harm to a patient.

EXTERNAL QUALITY ASSESSMENT OR PROFICIENCY TESTING

EQA/PT is used to evaluate measurement procedure performance by comparing a laboratory's results with those of other laboratories for the same set of samples. An EQA/PT provider circulates a set of samples among a group of laboratories that can be several thousand in larger programs. Each laboratory measures the EQA/PT samples as if they were patient samples and reports the results to the EQA/PT provider for evaluation. The EQA/PT provider establishes target values for the EQA/PT samples and determines if the results for an individual laboratory are in close enough agreement with the target value to be consistent with acceptable measurement procedure performance.

EQA/PT is not available for some analytes because a particular measurement procedure may be new to the clinical laboratory or is not commonly performed or because analyte stability makes it difficult to include in an EQA/PT sample. In these situations the laboratory should use an alternate approach to periodically verify acceptable performance of the measurement procedure. CLSI guideline GP27 provides approaches for verifying measurement procedure performance when formal EQA/PT is not available.

Internal QC material manufacturers may provide a data analysis service that compares results from different laboratories using the same QC material by calculating group statistics for performance evaluation. As with EQA/PT evaluation, this type of interlaboratory QC data analysis allows a laboratory to verify that it is producing QC results that are consistent with those of other laboratories using the same measurement procedure. This information can be helpful for troubleshooting measurement procedure issues and for assessing performance of a new measurement procedure being introduced to a laboratory.

External Quality Assessment or Proficiency Testing Programs That Use Commutable Samples

EQA/PT programs that use commutable samples are preferred whenever available. Commutable samples are typically prepared by using an individual donor's specimen or by pooling clinical patient samples with minimal processing or additives to avoid alteration of the sample matrix. When commutable EQA/PT samples can be prepared, the results reflect what would be expected if individual patient samples were sent to each of the different laboratories. Thus agreement among different laboratories and measurement procedures (harmonization) can be correctly evaluated.

The agreement between an individual laboratory result and a reference measurement result gives an assessment of correct calibration for the laboratory. The agreement between an individual laboratory result and an all-methods mean gives an assessment of harmonization of results with other measurement procedures. The agreement between a measurement procedure group mean value and the reference measurement result gives an assessment of trueness and calibration traceability for the measurement procedure group. In addition, the agreement between a measurement procedure group mean value and an all results mean gives an assessment of harmonization of results. The information for measurement procedure groups is of particular interest to the producers of measurement procedures and can be used as part of a surveillance program for the calibration traceability scheme.

External Quality Assessment or Proficiency Testing Programs That Use Noncommutable Samples

The materials commonly used for EQA/PT samples are derived from blood, urine, or other body fluids but are altered in the process to manufacture EQA/PT samples such that the matrix is modified and the samples frequently do not have the same measurement characteristics as observed for unaltered clinical patient samples. In addition, some EQA/PT samples (e.g., urine, cerebrospinal fluid, or blood gas) are prepared as synthetic materials that are not derived from patient fluids. Consequently, many EQA/PT samples, as for QC samples, are noncommutable with authentic patient samples. The results for a noncommutable EQA/PT sample will have a different relationship in their numeric values between different measurement procedures and sometimes for different reagent lots within a measurement procedure than would be observed for patient samples.

It is a common practice for EQA/PT providers to organize results for noncommutable samples into "peer groups" of measurement procedures that represent similar technology expected to have the same result for a noncommutable EQA/PT sample. The mean or median value of the peer group results is the target value. Because the peer group mean value may be influenced by a matrix-related noncommutability bias, that mean value can be used only to evaluate laboratories using the same or very similar measurement procedures and cannot be used to evaluate laboratories using other measurement procedures or to evaluate if results from different measurement procedures agree with each other. If an individual laboratory's results agree with those of the peer group, the individual laboratory can conclude that the measurement procedure was performing in conformance with the manufacturer's specifications. An individual laboratory cannot use results for noncommutable EQA/PT samples to verify that a measurement procedure is calibrated correctly to be traceable to the reference system for an analyte.

However, even within a peer group using the same measurement procedure, differences can occur because of different reagent lots used by the measurement procedures in different laboratories because the matrix of the EQA/PT material can influence the results from different reagent lots when the patient samples give similar results. Therefore, in some cases, reagent lots should be registered as part of the EQA/PT reporting, and even reagent lot–specific target values may need to be assigned.

Reporting External Quality Assessment or Proficiency Testing Results When One Measurement Procedure Is Adjusted to Agree With Another Measurement Procedure

It is good laboratory practice to adjust the calibration of different measurement procedures for the same measurand used within a large hospital system that can have several satellite laboratories or within a collection of several hospitals with the same management structure, so that the results for patient samples are consistent, irrespective of which measurement procedure is used. Such harmonization of results is important for uniform use of reference intervals and decision thresholds within a hospital or clinic system. It is important to report EQA/PT results such that they can be properly evaluated against the peer group target value. The peer group target value will reflect the measurement procedure calibration established by the measurement procedure manufacturer. For an individual laboratory's EQA/PT result to be evaluated against the peer group mean, that individual result must be reported to the EQA/PT provider after removing any calibration adjustments so that the reported result is consistent with the manufacturer's nonadjusted calibration.

The most convenient way to remove a calibration adjustment is to first measure the EQA/PT samples with the calibration adjustment applied to the measurement procedure, as would be the usual measurement process for patient samples. After the measurement, the EQA/PT results should be adjusted "in reverse" by mathematically removing the calibration adjustment factors, and the results should be reported to the EQA/PT provider with any adjustment factors removed. One should not recalibrate the instrument with a new set of calibrators for the purpose of measuring the EQA/PT samples because this practice would violate regulations requiring the EQA/PT material to be measured in the same manner as patient samples. This process permits the EQA/PT sample to be measured in the same manner as patient samples and the numeric result reported to the EQA/PT provider to reflect the actual measured result using the manufacturer's calibration settings.

Interpretation of External Quality Assessment or Proficiency Testing Results

Many countries have regulations requiring EQA/PT and specifying the evaluation criteria for acceptable performance. When criteria are set by regulations, an EQA/PT provider is required to use them. When criteria are not set by regulations, the EQA/PT provider sets evaluation criteria on the basis of clinically acceptable performance, biological

External Quality Assessment (Proficiency Testing) Participant Report
Shipment date: 1 May 2015
Evaluation date: 12 June 2015

Analyte Units Method	Specimen	Reported Result	Mean	SD	Labs *(n)*	SDI	Limits of Acceptability Lower	Upper
Calcium	1	9.6	9.92	0.23	587	–1.4	8.9	11.0
mg/dL	2	8.8	8.86	0.26	592	–0.2	7.8	9.9
Arsenazo dye	3	7.5	7.65	0.23	587	–0.7	6.6	8.7
Manufacturer A	4	8.2	8.43	0.23	590	–1.0	7.4	9.5
	5	10.8	10.87	0.25	589	–0.3	9.8	11.9
Iron	1	190	192.5	7.0	397	–0.4	154	232
µg/dL	2	65	65.0	3.4	394	0.0	51	78
Pyridylazo dye	3	74	69.2	3.2	395	+1.5	55	83
Manufacturer A	4	124	107.9	4.6	395	+3.5	86	130
	5	277	260.9	8.8	396	+1.8	208	314

A

External Quality Assessment (Proficiency Testing) Participant Report
Shipment date: 1 May 2015
Evaluation date: 12 June 2015

Analyte Units Method	Specimen	Reported Result	Mean	SD	Labs *(n)*	SDI	Limits of Acceptability Lower	Upper
Calcium	1	2.40	2.48	0.06	587	–1.4	2.22	2.74
mmol/L	2	2.20	2.21	0.06	592	–0.2	1.95	2.47
Arsenazo dye	3	1.87	1.91	0.06	587	–0.7	1.65	2.17
Manufacturer A	4	2.05	2.10	0.06	590	–1.0	1.85	2.37
	5	2.69	2.71	0.06	589	–0.3	2.45	2.97
Iron	1	34.0	34.5	1.3	397	–0.4	27.6	41.5
µmol/L	2	11.6	11.6	0.6	394	0.0	9.1	14.0
Pyridylazo dye	3	13.3	12.4	0.6	395	+1.5	9.8	14.9
Manufacturer A	4	22.2	19.3	0.8	395	+3.5	15.4	23.3
	5	49.6	46.7	1.6	396	+1.8	37.2	56.2

B

Fig. 7.10 Example of an external proficiency testing evaluation report sent to a participating laboratory. *Part A* uses conventional units and *part B* uses SI units. *SD*, Standard deviation; *SDI*, standard deviation interval. (From Miller, W. G. [2016]. Quality control. In: *Henry's clinical diagnosis and management by laboratory methods* [23rd ed.]. Philadelphia: Elsevier.)

variation, or the analytical capability of the measurement procedures in use. EQA/PT evaluation criteria are usually designed to evaluate the total error of a single measurement. In some programs, measurements are made several times, and it is possible to separately assess the bias and the imprecision. The acceptability limits for EQA/PT include bias and imprecision components considered acceptable for clinical use of a result, plus other error components that are unique to EQA/PT samples such as between-laboratory variation in calibration; variable matrix-related noncommutability bias with different lots of reagent within a peer group; uncertainty in the target value; stability variability in the EQA/PT material, both in storage and shipping, and after reconstitution or opening in the laboratory; and homogeneity of the EQA/PT material vials. Consequently, the acceptability limits for EQA/PT samples are frequently larger than what might be expected for clinically acceptable total error with patient samples.

Fig. 7.10 is an example of a typical evaluation report sent to a participating laboratory. Each reported result is compared with the mean result for the peer group using the same measurement procedure. The report also includes the SD for the distribution of results in the peer group, the number of laboratories in the peer group, and the SDI (also called a *z*-score), which expresses the reported result as the number of SDs it is from the mean value (SDI = [result – mean]/SD). The limits of acceptability are shown. Acceptability criteria may be a number of SDs from the mean value, a fixed percent from the mean value, or a fixed concentration from the mean value. For example, in Fig. 7.10, calcium acceptability criteria are

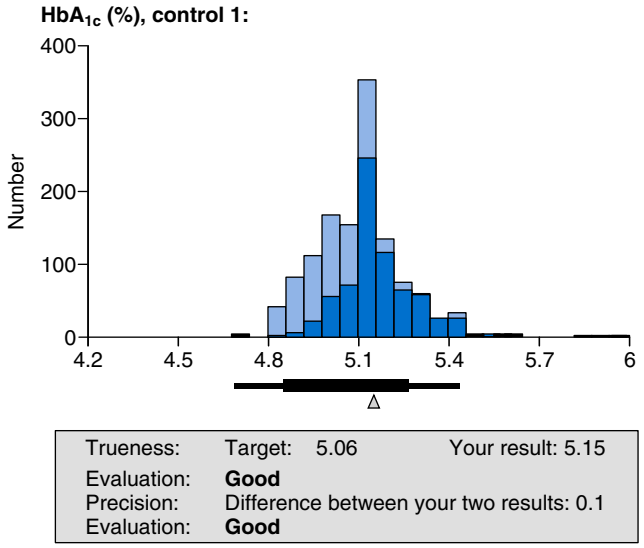


Fig. 7.11 Example of part of a feedback report to hemoglobin A_{1c} *(HbA_{1c})* point-of-care (POC) users in a survey for general practitioners' offices and nursing homes. Commutable external quality assessment/proficiency testing material was circulated in two levels and measured in duplicate. The participant is informed about the bias (mean of the two results) compared with a reference measurement procedure target *(x-axis)* and "precision" as the difference between the two results. The histogram represents the distribution of results among all participants *(light blue)* and for the participant's method group *(dark blue)*. The thick black line represents the interval for "good" results, and the thin black line represents the interval for "acceptable" results. Results outside these limits are characterized as "poor." The triangle points to the result of the participant. (Modified with permission from the Norwegian Quality Improvement of Primary Care Laboratories, the external quality assessment provider in Norway.)

±1 mg/dL (0.25 mmol/L) from the mean value, and iron criteria are ±20% from the mean value.

In Fig. 7.10 the calcium results are in close agreement with the peer group mean (SDI ranges from −0.2 to −1.4), and the laboratory can conclude that its results are consistent with those of others in the peer group using the same measurement procedure and that it is using the measurement procedure according to the manufacturer's specifications. However, the iron results show greater variability, with one result +3.5 SDI. Although all iron results are within the acceptability criteria, it is recommended to investigate the measurement procedure because a +3.5 SDI is more likely to be different from other participants in the peer group than to be in agreement with them.

Fig. 7.11 shows another type of evaluation report sent to a primary care office for hemoglobin A_{1c} (HbA_{1c}) for one of two EQA/PT samples. In this situation the EQA/PT provider is communicating directly with the clinician or the coworker in the general practice office, and the feedback must be easy to understand for nonlaboratory professionals. The EQA/PT result is evaluated as "good," "acceptable," or "poor." The lot numbers of the reagent are registered so that the participant, in case of an aberrant result, can get information if the result was due to the measurement procedure used, the reagent lot used, or the performance of the user.

Fig. 7.12 shows a similar report from the same HbA_{1c} survey provided to hospital laboratories. In addition to the figures about the distribution of results, information is provided on how different measurement procedures performed, as well as a historical overview of performance on consecutive EQA/PT samples and performance related to the concentration of the sample. The EQA/PT material used for the HbA_{1c} is pooled fresh patient blood (commutable), and the target value is set by a reference measurement procedure and is therefore the same for all measurement procedures. Each sample was measured in duplicate (as requested by the EQA/PT provider), and the mean of the duplicate was used to estimate bias versus the reference measurement procedure. In the present example the performance was within the acceptability limits but with a generally high bias during the whole period. Because this observation was true for all the instruments using this measurement procedure, the EQA/PT organizer discussed the results with the manufacturer to solve the problem. Until the problem was solved (the manufacturer had to make a new calibrator), the participants were advised by the EQA/PT provider to use a correction factor when reporting their results for patient samples.

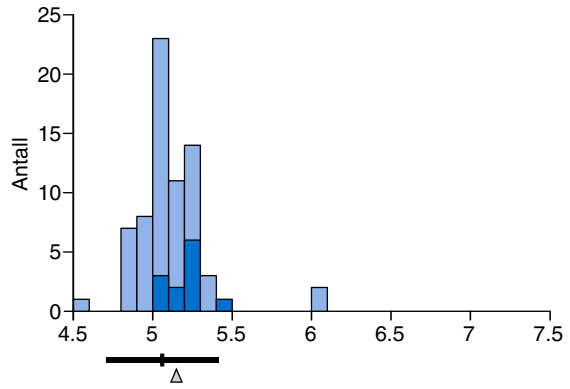
If an unacceptable EQA/PT result is identified, the measurement procedure must be investigated for possible causes and the necessary corrective action taken. Even when an EQA/PT result is within acceptability criteria, it is a good laboratory practice to investigate results that are more than approximately 2.5 SDI from the peer group mean. When the SDI is 2.5, there is only a 0.6% probability that the result will be within the expected distribution for the peer group; consequently, the probability is reasonable that a measurement procedure problem may need to be corrected.

Common causes for EQA/PT failure are listed in Box 7.1. Incorrect handling and reporting are unique to EQA/PT events and may not reflect the process used in the laboratory for patient samples. Because the influence of reagent lots on noncommutability related bias is well documented, a reagent lot-specific bias is a possible explanation when no other root cause can be identified.

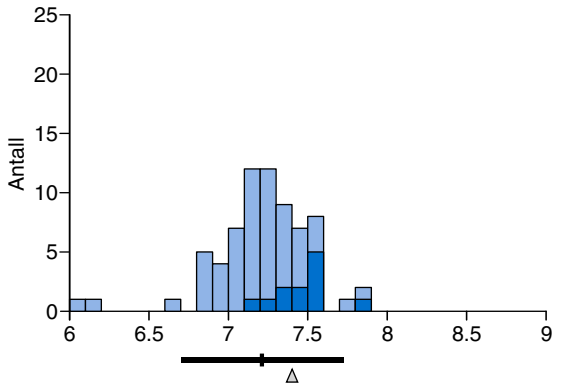
POINTS TO REMEMBER

- An independent external organization circulates EQA/PT samples with unknown target values.
- When commutable "patient-like" material is used, a laboratory can compare its results with results from all other measurement procedures and often with a true value from a reference measurement procedure.
- When noncommutable material is used, a laboratory can compare its results only with results from participants in a "peer group" using a similar measurement procedure.

HbA_{1c} % HbA_{1c} **Tosoh HPLC G8**



Δ Your result: 5.15 (5.2–5.1). Deviation from target value: 1.8%
- Acceptable range (target ± 7%): 4.71–5.41
All
Your method group



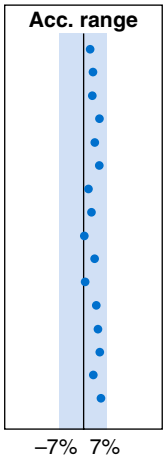
Δ Your result: 7.40 (7.4–7.4). Deviation from target value: 2.6%
- Acceptable range (target ± 7%): 6.71–7.72
All
Your method group

Overview

	1-SHA115					2-SHA115				
Method group	**Median**	**SD**	**CV**	***n***	***n_x***	**Median**	**SD**	**CV**	*n*	*n_x*
Afinion	5.20			1	0	7.40			1	0
Architect ci8200/c8000	5.00			4	1	7.18			4	1
Biorad D10/Variant HPLC/Turbo	5.00	0.107	2.1	8	0	7.23	0.113	1.6	8	0
Cobas. Tina-quant	5.00	0.200	4.0	21	2	7.20	0.268	3.7	21	2
DCA 2000/2000+/Vantage	5.00	0.130	2.6	20	0	7.08	0.216	3.1	20	0
Dimension Vista 1500	5.25			1	0	7.70			1	0
Tosoh HPLC G7/G8	5.20	0.108	2.1	12	0	7.45	0.186	2.5	12	0
Total	**5.05**	**0.157**	**3.1**	**67**	**3**	**7.20**	**0.246**	**3.4**	**67**	**3**

History: Deviation from target value (%)

Survey	Target	Result	Deviation	Instrument
2015.1	5.06	5.15	1.8	TOSG8
2015.1	7.21	7.40	2.6	TOSG8
2014.4	5.71	5.85	2.5	TOSG8
2014.4	7.46	7.80	4.6	TOSG8
2014.3	5.04	5.20	3.2	TOSG8
2014.3	7.46	7.80	4.6	TOSG8
2014.2	5.57	5.65	1.4	TOSG8
2014.2	7.13	7.30	2.4	TOSG8
2014.1	6.34	6.35	0.2	TOSG8
2014.1	6.78	7.00	3.2	TOSG8
2013.4	5.08	5.10	0.4	TOSG8
2013.4	6.84	7.10	3.8	TOSG8
2013.3	5.76	6.00	4.2	TOSG8
2013.3	6.83	7.15	4.7	TOSG8
2013.2	5.06	5.20	2.8	TOSG8
2013.2	7.14	7.50	5.0	TOSG8



Deviation from target value (concentration)

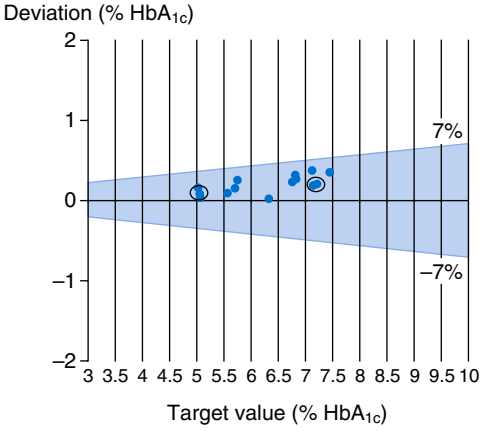


Fig. 7.12 Example of a part of a feedback report to hemoglobin A_{1c} *(HbA_{1c})* users in hospital laboratories. Same survey and same materials as presented in Fig. 7.11. The histogram represents the distribution of results among all participants *(light blue)* and for the participant's method group *(dark blue)*. Only limits for "acceptable" *(Acc.)* results are given (*thin black lines* in figures). Information about performance of measurement procedures is given in addition to a historical overview of percentage deviation from target values dependent on time and concentration of HbA_{1c}. *CV*, Coefficient of variation; *HPLC*, high-pressure liquid chromatography; *SD*, standard deviation. (Modified with permission from the Norwegian Quality Improvement of Primary Care Laboratories, the external quality assessment provider in Norway.)

BOX 7.1 Classification of Potential Problems Identified When Investigating Unacceptable External Quality Assessment or Proficiency Testing Results[a]

1. **Clerical errors**
   - Incorrectly transcribed EQA/PT result from the instrument read-out to the report form
   - The EQA/PT sample was mislabeled in the laboratory
   - Incorrect instrument or measurement procedure was reported on the results submission form
   - Incorrect units were reported
   - Decimal point was misplaced
2. **Measurement procedure problems**
   - Inadequate SOP
   - Problem with manufacture or preparation of reagents or calibrators (e.g., unstable)
   - Lot-to-lot variation in reagents or calibrators
   - Incorrect value assignment of calibrators
   - Measurement procedure lacks adequate specificity for the measurand
   - Measurement procedure lacks adequate sensitivity to measure the concentration
   - Carry-over from a previous sample
   - Inadequate QC procedures used
3. **Equipment problems**
   - Obstruction of instrument tubing or orifice by clot
   - Misalignment of instrument probes
   - Incorrect instrument data processing functions
   - Incorrect instrument setting
   - Automatic pipetter not calibrated to acceptable precision and accuracy
   - Equipment component malfunction (e.g., light source, membrane, fluidics, detector)
   - Incorrect instrument conditions (e.g., water quality, surrounding temperature)
   - Instrument maintenance not performed appropriately
4. **Technical problems caused by personnel errors**
   - Did not operate equipment correctly or did not conform to measurement procedure SOP
   - Incorrect storage, preparation, or handling of reagents or calibrators
   - Delay causing evaporation or deterioration of the EQA/PT sample
   - Failure to follow recommended instrument function checks or maintenance
   - Pipetting or dilution error
   - Calculation error
   - Misinterpretation of test result
5. **A problem with the EQA/PT material such as:**
   - Incorrect storage, preparation, or handling of EQA/PT materials
   - Differences between EQA/PT samples and patient samples (e.g., matrix, additives, stabilizers)
   - Sample deteriorated in transit or during laboratory storage
   - Sample had weak or borderline reaction
   - Sample contained interfering factors (which may be measurement procedure specific)
   - Sample was not homogeneous among vials

[a] This classification scheme assists in developing an appropriate corrective action plan.
EQA, External quality assessment; *PT,* proficiency testing; *QC,* quality control; *SOP,* standard operating procedure.
From Miller, W. G., Jones, G. R. D., Horowitz, G. L., & Weykamp, C. (2011). Proficiency testing/external quality assessment: current challenges and future directions. *Clinical Chemistry, 57*, 1670–1680.

REVIEW QUESTIONS

1. Why should internal quality control be performed?
   - **a.** To be sure that the quality control material is of good quality
   - **b.** To examine if the control material is commutable—behaves like patient samples
   - **c.** To have a high probability that correct patient results are released
   - **d.** To be able to pass the accreditation inspection
   - **e.** To examine if my measurement procedure gives results similar to other laboratories
2. How should I interpret results when an EQA/PT program uses noncommutable materials?
   - **a.** When my result is within the performance specifications, I can be confident that I have no bias compared with a true value.
   - **b.** When my result is outside the performance specifications, I have to recalibrate my measurement procedure.
   - **c.** When my result is close to the target value for the peers in my measurement procedure group, I can be confident my laboratory is performing as well as my peers.
   - **d.** I should compare my results with results from other measurement procedures to be confident my laboratory is not biased.
   - **e.** I should compare my results with the average of all measurement procedure groups.
3. What is the advantage of using commutable EQA/PT materials?
   - **a.** They are similar to patient samples and should therefore not be used for QC.
   - **b.** They are suitable for internal quality control but not for EQA/PT.
   - **c.** They should be avoided since they are contagious.
   - **d.** Their results provide information on the accuracy for patient samples if the target value is set by a reference measurement procedure.
   - **e.** Their results can be compared among different measurement procedures to assess harmonization.
4. When should EQA/PT be used?
   - **a.** To decide if patient results can be released
   - **b.** Should be performed every second day

c. To assess the performance of your measurement procedure compared with other measurement procedures
d. Only when the provider uses commutable material
e. Is not necessary if you have documented traceability of your measurement procedure

5. How is the SD estimated for an internal QC material?
a. From measurements made during the target value assignment of a QC material
b. From the long-term SD that includes most types of variability expected to influence the measurement procedure
c. From the instructions for use provided by the measurement procedure manufacturer
d. From reports of interlaboratory summaries
e. From the range of acceptable results provided by the QC material manufacturer

SUGGESTED READINGS

CLSI. (2011). *Laboratory quality control based on risk management; approved guideline EP23-A*. Wayne, PA: Clinical and Laboratory Standards Institute.

CLSI. (2013). *User evaluation of between-reagent lot variation; approved guideline EP26-A*. Wayne, PA: Clinical and Laboratory Standards Institute.

CLSI. (2016). *Statistical quality control for quantitative measurement procedures: principles and definitions; approved guideline C24–A4*. Wayne, PA: Clinical and Laboratory Standards Institute.

CLSI. (2016). *Using proficiency testing to improve the clinical laboratory; approved guideline GP27–A3*. Wayne, PA: Clinical and Laboratory Standards Institute.

Miller, W. G. (2016). Time to pay attention to reagent and calibrator lots for proficiency testing. *Clinical Chemistry*, *62*, 666–667.

Miller, W. G., Erek, A., Cunningham, T. D., et al. (2011). Commutability limitations influence quality control results with different reagent lots. *Clinical Chemistry*, *57*, 76–83.

Miller, W. G., Jones, G. R. D., Horowitz, G. L., et al. (2011). Proficiency testing/external quality assessment: current challenges and future directions. *Clinical Chemistry*, *57*, 1670–1680.

Miller, W. G., & Sandberg, S. (2017). Quality Control of the Analytical Examination Process. In N. Rifai, A. R. Hovath, & C. T. Wittwer (Eds.), *Tietz textbook of clinical chemistry and molecular diagnostics (Chapter 6)*. Elsevier.

Panteghini, M., & Sandberg, S. (Eds.). (2015). 1st EFLM Strategic Conference "Defining analytical performance goals – 15 years after the Stockholm Conference." *Clinical Chemistry and Laboratory Medicine*, *53*(6, special issue), 829–958.

Stavelin, A., Riksheim, B. O., Christensen, N. G., et al. (2016). The Importance of reagent lot registration in external quality assurance/proficiency testing schemes. *Clinical Chemistry*, *62*, 708–715.

Vesper, H. W., Miller, W. G., & Myers, G. L. (2007). Reference materials and commutability. *Clinical Biochemist Reviews*, *28*, 139–147.

8

Principles of Basic Techniques and Laboratory Safety

*Stanley Lo**

OBJECTIVES

1. Define the following:
   - Analyte
   - Centrifugation
   - Dilution
   - Evaporation
   - Filtration
   - Gravimetry
   - Lyophilization
   - Primary/Secondary reference material
   - Relative centrifugal force/field (RCF)
   - Solution
   - Standard precautions
2. Describe three methods of preparing reagent grade water; identify the criteria for determining when the purification process for Clinical Laboratory Reagent Grade Water is considered acceptable.
3. List and compare the different grades of reagents available, and state which are appropriate for use in a clinical laboratory.
4. Compare the two types of reference materials used in the clinical laboratory, and state the specific uses of each type.
5. List and describe three types of pipettes used in the clinical laboratory; state the proper use and the specific uses of each type.
6. For the process of centrifugation:
   - State the principle.
   - List six uses of centrifugation in the clinical laboratory.
   - Calculate RCF and revolutions per minute (rpm) when given the appropriate information.
   - Determine the time required for centrifugation using an alternate rotor.
   - Outline proper operation and operation practice of a centrifuge.
7. Describe three types of balances used in the clinical laboratory and how they are calibrated.
8. Compare serial dilutions with simple dilutions; state and calculate the formula used to prepare a solution of lesser concentration from one of greater concentration.
9. List and describe the elements of an Occupational Safety and Health Administration (OSHA)-approved chemical hygiene plan, a hazard exposure plan, and a tuberculosis control plan; describe the Standard Precautions document, including source and specific mandates; state the College of American Pathologists (CAP) requirements for a laboratory ergonomics program.
10. For the following hazard types, interpret laboratory hazard signage, and state the appropriate work practice used to control these hazards:
    - Biological
    - Chemical
    - Electrical
    - Fire

KEY WORDS AND DEFINITIONS

Analyte A solute dispersed in a solution that is measured in laboratory practice; also referred to as a *measurand*.

Buffer solution A solution that contains either a weak acid and its salt or a weak base and its salt and is resistant to changes in pH.

Centrifugation The process of using centrifugal force to separate the lighter portions of a solution from the heavier portions; a centrifuge is a device by which centrifugation is affected.

Chemical hygiene plan (CHP) An Occupational Safety and Health Administration (OSHA)-required listing of responsibilities for laboratory employers, employees, and a chemical hygiene officer, and including a complete chemical inventory that is updated annually, along with a copy of the Material Safety Data Sheet (MSDS) that defines each chemical as toxic, carcinogenic, or dangerous, and that must be on file and available to all employees 24 hours a day, 7 days a week.

* The author gratefully acknowledges the original contributions of Drs. Edward W. Bermes, Stephen E. Kahn, Donald S. Young, Edward R. Powsner, and John C. Widman, on which portions of this chapter are based.

Clinical and Laboratory Standards Institute (CLSI) The CLSI (formerly the National Committee for Clinical Laboratory Standards, or NCCLS) that guides the development and implementation of standards and guidelines that help all laboratories fulfill their goals.

Ergonomics The study of capabilities in relationship to work demands completed by defining postures that minimize unnecessary static work and reduce the forces working on the body.

Exposure control plan An OSHA-required plan that ensures the protection of laboratory workers against potential exposure to bloodborne pathogens, while ensuring that medical wastes produced by the clinical laboratory are managed and handled in a safe and effective manner.

Material Safety Data Sheet (MSDS) A technical bulletin that contains information about a hazardous chemical, such as chemical composition, chemical and physical hazard, and precautions for safe handling and use.

Measurand See *Analyte*.

OSHA Occupational Safety and Health Administration, formed by the federal government of the United States to formally regulate the oversight of employee safety.

Pipette Device used for the transfer of a volume of liquid from one container to another.

Reagent Chemical used in many high-purity applications.

Reference material A material or substance, one or more physical or chemical properties of which are sufficiently well established to be used for the calibration of an apparatus, the verification of a measurement method, or the assigning of values to materials. Certified, primary, and secondary are types of reference materials.

Relative centrifugal force or field (RCF) Force required to separate two phases (liquid and solid) in a centrifuge.

Standard precautions Approach to infection control that treats all human blood and certain human body fluids as if they were known to be infectious for bloodborne pathogens.

An understanding of basic principles and techniques of analytical chemistry is necessary to appropriately interpret clinical laboratory tests and adequately validate assays. These are integral to understanding basic laboratory tasks such as making **buffer solutions**, performing dilutions, evaporation, and filtration. Safety is a constant concern for laboratory personnel. A comprehensive safety program consisting of a plan for the handling of chemicals, exposure to bloodborne pathogens, tuberculosis (TB), and other highly infectious agents, as well as training to handle various types of hazards is essential.

CHEMICALS

The quality of the analytical results produced by the laboratory is directly affected by the purity of the chemicals used as analytical **reagents**. The availability and quality of the reference materials used to calibrate assays and to monitor their analytical performance are also important.

Laboratory chemicals are available in a variety of grades. The solutes and solvents used in analytical work are *reagent grade chemicals*, among which water is a solvent of primary importance.

Reagent Grade Water

Preparation of many reagents and solutions used in the clinical laboratory requires "pure" water. Single-distilled water fails to meet the specifications for Clinical Laboratory Reagent Water (CLRW) established by the **Clinical and Laboratory Standards Institute (CLSI)**. Because the terms *deionized water* and *distilled water* describe preparation techniques, they should be replaced by reagent grade water, and if appropriate, followed by the designation of CLRW, which better defines the specifications of the water and is independent of the method of preparation (Table 8.1).

TABLE 8.1 Clinical and Laboratory Standards Institute Specifications for Clinical Laboratory Reagent Water

	CLRW
Microbiological content[a] (cfu/mL) (maximum)	<10
Resistivity[b] (MΩ/cm), 25°C	≥10
Particulate matter[c]	Water passed through 0.22-μm filter
Total organic content (ng/g)	<500

CLRW, Clinical Laboratory Reagent Water.

[a]Microbiological content. The microbiological content of viable organisms, as determined by total colony count after incubation at 36 ± 1°C for 14 h, followed by 48 h at 25 ± 1°C, and reported as colony-forming units per milliliters (cfu/mL).

[b]Specific resistance or resistivity. The electrical resistance in ohms measured between opposite faces of a 1-cm cube of an aqueous solution at a specified temperature. For these specifications, the resistivity will be corrected for 25°C and reported in megaohms per centimeters (MΩ/cm). The higher the quantity of ionizable materials, the lower will be the resistivity and the higher the conductivity.

[c]Particulate matter. When water is passed through a membrane filter with a mean pore size of 0.22 μm, nearly all microorganisms and particulates are removed.

From Clinical and Laboratory Standards Institute. (2006). *Preparation and testing of reagent water in the clinical laboratory* (4th ed.). CLSI Document GP40-A4-AMD. Wayne, PA: CLSI.

Preparation of Reagent Grade Water

Distillation, ion exchange, reverse osmosis, and ultraviolet (UV) oxidation are processes used to prepare reagent grade water. In practice, water is filtered before any of these processes are used.

Distillation. Distillation is the process of vaporizing and condensing a liquid to purify or concentrate a substance or to separate a volatile substance from less volatile substances. It is the oldest method of water purification. Problems with distillation for preparing reagent water include the carryover of volatile impurities and entrapped water droplets that may contain impurities into the purified water. This results in contamination of the distillate with volatiles, sodium, potassium, manganese, carbonates, and sulfates. As a result, water treated by distillation alone does not meet the specific conductivity requirement of CLRW.

Ion exchange. Ion exchange is a process that removes ions to produce mineral-free deionized water. Such water is most conveniently prepared using commercial equipment, which ranges in size from small, disposable cartridges to large, resin-containing tanks. Deionization is accomplished by passing feed water through columns containing insoluble resin polymers that exchange hydrogen (H^+) and hydroxide (OH^-) ions for the impurities present in the ionized form in the water. The columns may contain cation exchangers, anion exchangers, or a *mixed-bed resin exchanger*, which is a mixture of cation- and anion-exchange resins in the same container.

A single-bed deionizer generally is capable of producing water that has a specific resistance in excess of 1 MΩ/cm. When connected in series, mixed-bed deionizers usually produce water with a specific resistance that exceeds 10 MΩ/cm.

Reverse osmosis. Reverse osmosis is a process by which water is forced through a semipermeable membrane that acts as a molecular filter. The membrane removes 95% to 99% of organic compounds, bacteria, other particulate matter, and 90% to 97% of all ionized and dissolved minerals, but has fewer of the gaseous impurities. Although this process is inadequate for producing reagent grade water for the laboratory, it may be used as a preliminary purification method.

Ultraviolet oxidation. UV oxidation is another method that works well as part of a total system. The use of UV radiation at the biocidal wavelength of 254 nm eliminates many bacteria and cleaves many ionizing organics that are then removed by deionization.

Quality, Use, and Storage of Purified Water

Autoclave or wash water (formerly type III water) may be used for glassware washing. However, final rinsing should be done with the water grade suitable for the intended glassware use. It may also be used for certain qualitative procedures, such as those used in general urinalysis. A variety of purification processes can be used to produce this type of water.

The CLRW, formerly described as type I or II water, is used for most routine clinical laboratory testing. Purification processes for CLRW must achieve specific targets for microbial content, resistivity, particulate matter, and total organic content (see Table 8.1). Storage and delivery systems should be constructed to ensure a minimum of chemical or bacterial contamination. The frequency of monitoring of water specifications and the purification system is to be determined by each laboratory.

Special reagent water is used for specific applications. At a minimum, the water should meet CLRW specifications; however, additional parameters may be needed for certain applications. For example, water used for DNA and RNA testing may have specifications for protease and nuclease activity. Water used in trace metal analysis should contain minimally detectable amounts of each metal being measured.

POINTS TO REMEMBER

Water

- There are several types of water used in the clinical laboratory, including CLRW, special reagent water, and autoclave and wash water. CLRW has specifications for its microbiological content, resistivity, particulate matter, and total organic content.
- The monitoring of water depends on the type of water and its intended use.

Testing for Water Purity

At a minimum, water should be tested for microbiological content, particulates, resistivity, and total organic content. Protocols for testing the purity of water can be found in CLSI GP40-A4-AMD:2012. The frequency of monitoring water purity should be established by the laboratory, such that water purity is maintained for its intended purpose. It should be noted that measurements taken at the time of production may differ from those taken at the time and place of use. For example, if the water is piped a long distance, consideration must be given to deterioration en route to the site of use. To meet the specifications for high-performance liquid chromatography (HPLC), in some instances it may be necessary to add a final 0.1-μm membrane filter. The water can be tested by HPLC using a gradient program and monitored with an UV detector. No peaks exceeding the analytical noise of the system should be found.

Reagent-Grade or Analytical Reagent-Grade Chemicals

Chemicals that meet specifications of the American Chemical Society are described as reagent or analytical reagent grade. These specifications have also become the de facto standards for chemicals used in many high-purity applications. These are available in two forms: (1) lot-analyzed reagents, in which each individual lot is analyzed and the actual amount of impurity reported, and (2) maximum impurities reagents, for which maximum impurities are listed. The Committee on Analytical Reagents of the American Chemical Society periodically publishes "Reagent Chemicals," which is a list of specifications (http://pubs.acs.org/reagents/index.html). These reagent-grade chemicals are of very high purity and are recommended for quantitative or qualitative analyses.

Ultrapure Reagents

Many analytical techniques require reagents whose purity exceeds the specifications of those described previously. Manufacturers offer selected chemicals that have been especially purified to meet specific requirements. There is no uniform designation for these chemicals and organic solvents. Terms such as *spectrograde*, *nanograde*, and *HPLC pure* have been used. Data of interest to the user (e.g., absorbance at a specific UV wavelength) are supplied with the reagent.

Other designations of chemical purity include chemically pure (CP), US Pharmacopeia (USP), and National Formulary (NF) grade (chemicals produced to meet specifications set down in the USP or the NF). Chemicals labeled purified, practical, technical, or commercial grade should not be used in clinical chemical analysis without previous purification.

REFERENCE MATERIALS

A **reference material** is a material or substance with one or more physical or chemical properties that is sufficiently well established to be used for (1) calibrating instruments, (2) validating methods, (3) assigning values to materials, and (4) evaluating the comparability of results. Reference materials are of prime importance in establishing *metrologic transferability* (http://www.bipm.org), a term defined as "the property of a measurement result whereby the result can be related to a reference through a documented unbroken chain of calibrations, each contributing to the measurement uncertainty."

Primary, secondary, standard, and certified are types of reference materials.

Primary Reference Materials

Primary reference materials are highly purified chemicals that are directly weighed or measured, and dissolved in a solvent with an exactly measured total volume after the chemical has gone completely into solution, to produce a solution whose concentration is exactly known. The International Union of Pure and Applied Chemistry (IUPAC) has proposed a degree of 99.98% purity for primary reference materials. These highly purified chemicals may be weighed out directly for the preparation of solutions of selected concentration or for the calibration of solutions of unknown strength. They are supplied with a certificate of analysis for each lot.

Secondary Reference Materials

Secondary reference materials are solutions whose concentrations cannot be prepared by weighing the solute and dissolving a known amount into a volume of solution. The concentration of secondary reference materials is usually determined by analysis of an aliquot of the solution by an acceptable reference method, using a primary reference material to calibrate the method.

Standard Reference Materials

Standard reference materials (SRMs) for clinical and molecular laboratories are available from the National Institute of Standards and Technology (NIST; http://www.nist.gov). Cholesterol, the first SRM developed by the NIST, was issued in 1967. It should be noted that not all SRMs have the properties and the degree of purity specified for a primary standard, but each has been well characterized for certain chemical or physical properties and is issued with a certificate that provides results of the characterization. These may then be used to characterize other materials.

Certified Reference Materials

Certified reference materials (CRMs) are available for clinical and molecular laboratories from the Institute for Reference Materials and Measurements (IRMM) in Geel, Belgium (http://www.irmm.jrc.be). The IRMM is one of the seven institutes of the Joint Research Centre (JRC), a Directorate-General of the European Commission (EC). Other acronyms used to label IRMM reference materials include ERM (European Reference Materials) and BCR (Community Bureau of Reference of the Commission of the European Communities). Reference materials are also available from the World Health Organization (WHO; http://www.who.int/biologicals).

POINTS TO REMEMBER

Ultrapure Reagents

There are several types of ultrapure reagents. For the clinical laboratory, they are referred to as reference materials. There are several different types of reference materials:

- Primary reference materials
- Secondary reference materials
- Standard reference materials
- Certified reference materials

BASIC TECHNIQUES AND PROCEDURES

Basic practices used in clinical laboratories include (1) optical, (2) chromatographic, (3) electrochemical, (4) electrophoretic, (5) mass spectrometric, (6) enzymatic, and (7) immunoassay techniques. These techniques are discussed in detail in Chapters 9 through 15. This chapter discusses the basic techniques of volumetric sampling and dispensing, centrifugation, measurement of radioactivity, gravimetry, thermometry, and processing solutions.

Volumetric Sampling and Dispensing

Clinical chemistry procedures require accurate volumetric measurements to ensure accurate results. For accurate work, only class A glassware should be used. Class A glassware is certified to conform to the specifications outlined in NIST circular C-602.

Pipettes

Pipettes are used for the transfer of a volume of liquid from one container to another. They are designed either (1) to contain a specific volume of liquid or (2) to deliver a specified volume. Pipettes used in clinical laboratories include (1) manual transfer and measuring pipettes, (2) micropipettes, and (3) electronic and mechanical pipetting devices. Developments

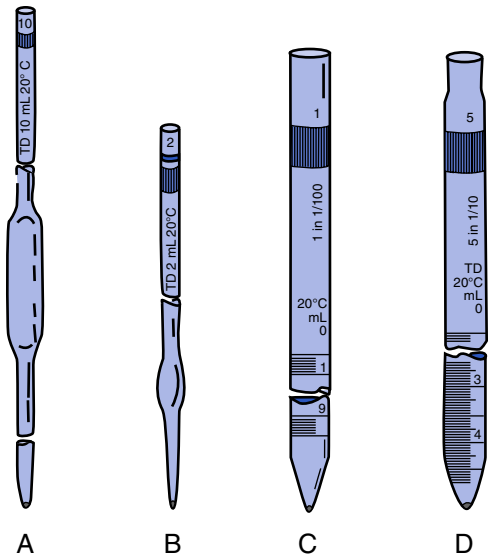


Fig. 8.1 Pipettes. (A) Volumetric (transfer); (B) Ostwald-Folin (transfer); (C) Mohr (measuring); (D) Serologic (graduated to the tip).

in improved design of pipetting systems include robotic automation, the capability to provide electronic and personal computer control of pipetting devices, and careful attention to advanced ergonomic design features. Automatic photometric pipette calibration systems are also available that reduce the time needed to periodically check pipettes and potentially allow more efficient use of personnel.

Transfer and measuring pipettes. A transfer pipette is designed to transfer a single known volume of liquid. Measuring and serologic pipettes are scored in units such that any volume up to a maximum capacity is delivered.

Transfer pipettes. Transfer pipettes include both volumetric and Ostwald-Folin pipettes (Fig. 8.1). They consist of a cylindrical bulb joined at both ends to narrower glass tubing. A calibration mark is etched around the upper suction tube, and the lower delivery tube is drawn out to a gradual taper. The bore of the delivery orifice should be sufficiently narrow to allow rapid outflow of liquid, without incomplete drainage causing measurement errors beyond the specified tolerances.

A volumetric transfer pipette (see Fig. 8.1A) is calibrated to accurately deliver a fixed volume of a dilute aqueous solution. The reliability of the calibration of the volumetric pipette decreases with decreased size, and therefore special micropipettes have been developed.

Ostwald-Folin pipettes (see Fig. 8.1B) are similar to volumetric pipettes, but they have the bulb closer to the delivery tip and are used for accurate measurement of viscous fluids, such as blood or serum. In contrast to a volumetric pipette, an Ostwald-Folin pipette has an etched ring near the mouthpiece, indicating that it is a blow-out pipette. Using a pipetting bulb, the liquid is blown out of the pipette only after the blood or serum has drained to the last drop in the delivery tip. When filled with opaque fluids, such as blood, the top of the meniscus must be read. Controlled slow drainage is required with all viscous solutions so that no residual film is left on the walls of the pipette.

Measuring pipettes. The second principal type of pipette is the graduated or measuring pipette (see Fig. 8.1C). This is a piece of glass tubing that is drawn out to a tip and graduated uniformly along its length. Two types are available. The Mohr pipette is calibrated between two marks on the stem, whereas the serologic pipette has graduated marks down to the tip. The serologic pipette (see Fig. 8.1D) must be blown out to deliver the entire volume of the pipette and has an etched ring (or pair of rings) near the bulb end of the pipette, signifying that it is a blow-out pipette. Mohr pipettes require controlled delivery of the solution between the calibration marks. Serologic pipettes have a larger orifice than do Mohr pipettes and thus drain faster. In practice, measuring pipettes are used principally for measurement of reagents and generally are not considered sufficiently accurate for measuring samples and calibrators.

Pipetting technique. There are general pipetting techniques that apply to the pipettes described previously. For example, pipetting bulbs should always be used, and pipettes must be held in a vertical position when the liquid level is adjusted to the calibration line and during delivery. The lowest part of the meniscus, when it is sighted at eye level, should be level with the calibration line on the pipette. The flow of liquid should be unrestricted when volumetric pipettes are used, and the tips should be touched to the inclined surface of the receiving container for 2 seconds after the liquid has ceased to flow.

With graduated pipettes, the flow of liquid may have to be slowed during delivery. Serologic pipettes are calibrated to the tip, and the etched glass ring on top of the pipette signifies that it is to be blown out. First, the pipette is allowed to drain, and then the remaining liquid is blown out.

Micropipettes. Micropipettes are pipettes used for the measurement of microliter volumes. With such devices, the remaining volume that coats the inner wall of a pipette causes notable error. For this reason, most micropipettes are calibrated to contain the stated volume rather than to deliver it. Proper use requires rinsing the pipette with the final solution after the contents are delivered into the diluent. Volumes are expressed in microliters (μL); the older term *lambda* is no longer recommended. Micropipettes generally are available in small sizes, ranging from 1 to 1000 μL; also, they are available for volumes as low as 0.2 μL.

Semiautomatic and automatic pipettes and dispensers. Fig. 8.2A and B, illustrate two types of adjustable micropipetting devices that demonstrate unique ergonomic design features. These devices are programmable and are used with disposable plastic tips for simultaneously dispensing aliquots of liquid into multiple wells. Each channel is piston driven to allow the user to pipette with as few or as many tips as necessary. Aliquots of liquid as small as 0.2 μL are dispensed at three different aspiration or dispense rates.

Semiautomatic manual and electronic versions of pipettes and dispensers are available in sizes ranging from 0.5 μL to 10 mL or larger. Fig. 8.2C illustrates an electronically operated, positive-displacement multichannel pipettor. This device aspirates and dispenses its predefined volumes (from 0.5 to

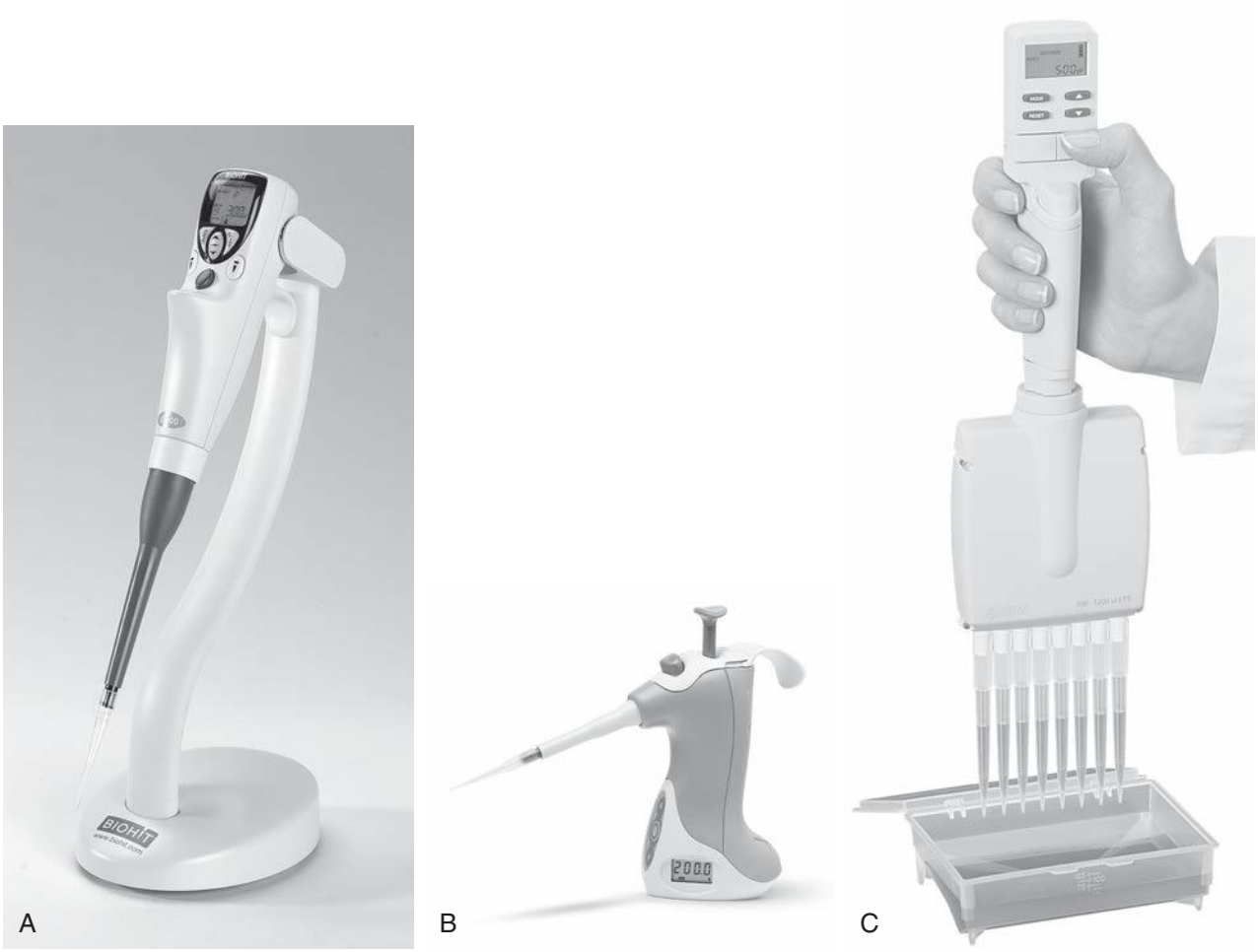


Fig. 8.2 (A) Adjustable volume micropipetting device with ergonomic design. (B) Adjustable volume electronic micropipetting device with ergonomic design. (C) Electronic programmable multichannel pipette. ([A] Courtesy Biohit Plc.; [B] Courtesy Vistalab Technologies, Inc.; [C] Courtesy Rainin Instrument LLC.)

200 μL) when its plunger is moved through a complete cycle. Its disposable, fluid containment tips are made of a plastic material that tends to retain less inner surface film than does glass. Such pipettes (1) avoid the risk of cross-contamination among samples, (2) eliminate the necessity for washing between samples, and (3) improve the precision of measurements. Models that allow for digital adjustment of the volume aspirated and dispensed are available.

Fig. 8.3A shows an automatic dispensing apparatus that aspirates and dispenses preset volumes of two different liquids by two motor-driven syringes, one for metering a volume of the sample and one for metering a volume of the diluent. It is possible to adjust this device to aspirate as little as 1 μL of one liquid and to deliver it with as much as 999 μL of the other. This type of device, available as a dilutor or dispenser, is obtainable as a manual, electronic, or computer-controlled device. The device is microprocessor-controlled and is easily programmed.

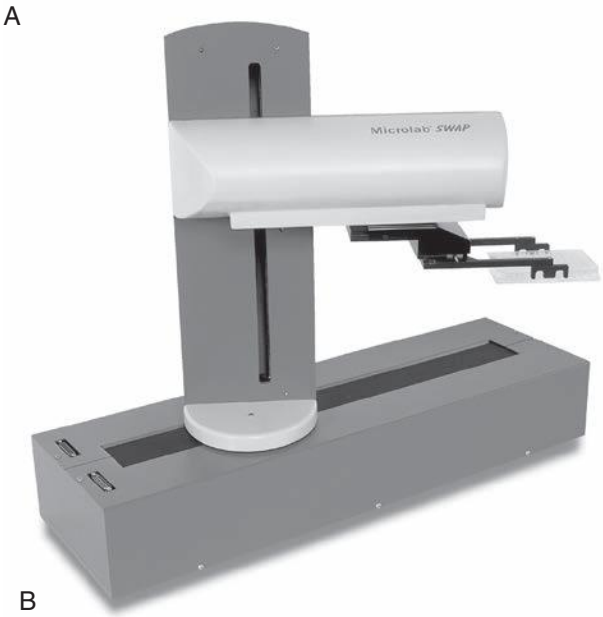
A more versatile piece of equipment is the robotic liquid handling workstation shown in Fig. 8.3B. This automated pipetting station is used with individual reaction tubes and also with 96- and 384-well microtiter plates. Depending on the design of the system, a single probe or multiple probes may be used to rapidly transfer programmed volumes of solution from one container to microtiter plates (e.g., so that the transfer to all 96 wells is complete in 1 minute). In some systems, liquid sensing is incorporated into the sample probes to minimize contact with sample and reagents, although automatic washing of the probes is performed between specimens. Two-dimensional (*X*-*Y*) movement of probes and tubes or microtiter plates is built into the pipetting stations to minimize the necessity for operator intervention. This device dispenses programmed volumes from 0.5 to 1000 μL in serial dilutions from 4 to 16 channels using an autoloaded system with barcodes for positive identification.

Volumetric Flasks

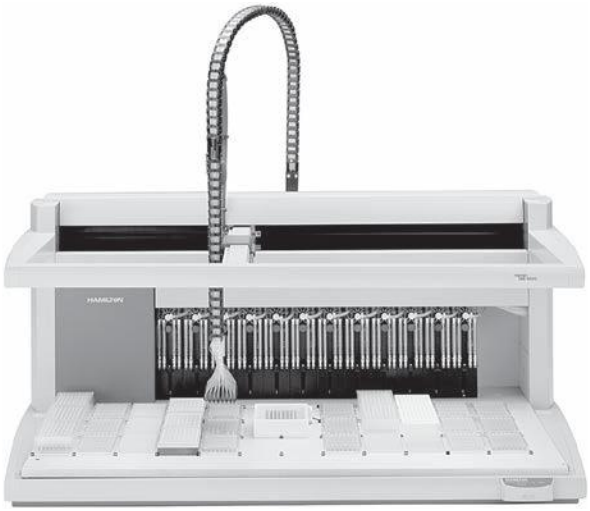
Volumetric flasks (Fig. 8.4) are used to measure exact volumes; they are commonly found in sizes varying from 1 to 4000 mL. In practice, they are used primarily to prepare solutions of known concentration, and they are available in various grades. The most accurate are certified as meeting standards set forth by the NIST.

An important factor in the use of a volumetric apparatus is the requirement for an accurate adjustment of the meniscus.

A



B



C

Fig. 8.3 (A) Personal computer–controlled diluting and/or dispensing apparatus that aspirates and dispenses preset volumes of one or two different liquids, such as a diluent and sample, by motor-driven syringes. (B and C) Robotic liquid handling workstations. ([A and B] Courtesy Hamilton Co.)

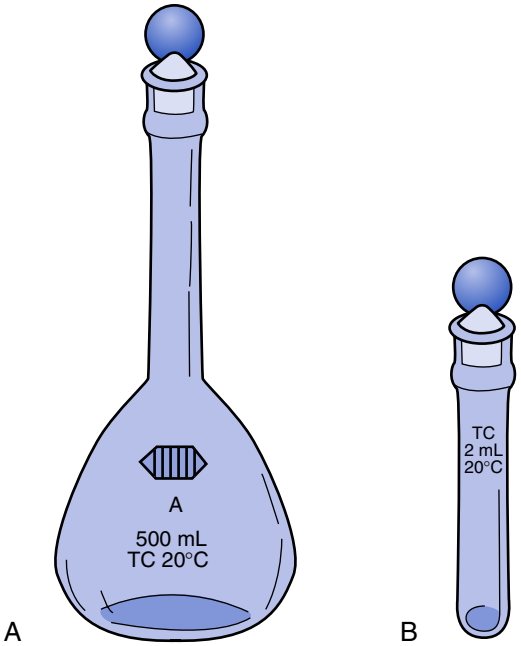


Fig. 8.4 Volumetric Flasks. (A) Macro; (B) Micro.

A small piece of card that is half black and half white is most useful. The card is placed 1 cm behind the apparatus, with the white half uppermost and the top of the black area approximately 1 mm below the meniscus. The meniscus then appears as a clearly defined, thin black line. This device is also useful in reading the meniscus of a burette.

Volumetric equipment should be used with solutions equilibrated to room temperature. Solutions diluted in volumetric flasks should be repeatedly mixed during dilution so that the contents are homogeneous before the solution is made up to final volume. Errors caused by expansion or contraction of liquids upon mixing are thereby minimized.

Volumetric flasks should be thoroughly cleaned and dried before calibration. The flask is then weighed and filled with carbon dioxide–free deionized water until just above the graduation mark. The neck of the flask just above the water level should be kept free of water. The meniscus mark is set at the graduation line by removing excess water, and the flask is reweighed. The final weight is corrected for the equilibrated water and air temperature to obtain the volume of the flask. Flasks also may be calibrated by the spectrophotometric technique described in the next section.

Centrifugation

Centrifugation is the process of using centrifugal force to separate the lighter portions of a solution, mixture, or suspension from the heavier portions. A centrifuge is a device by which centrifugation is performed.

In the clinical laboratory, centrifugation is used to:

1. Remove cellular elements from blood to provide cell-free plasma or serum for analysis (see Chapter 6).
2. Concentrate cellular elements and other components of biological fluids for microscopic examination or chemical analysis.

3. Remove chemically precipitated protein from an analytical specimen.
4. Separate protein- or antibody-bound ligand from free ligand in immunochemical and other assays (see Chapter 15).
5. Extract solutes in biological fluids from aqueous to organic solvents.
6. Separate lipid components (e.g., chylomicrons) from other components of plasma or serum, and lipoproteins from one another (see Chapter 23).

Types of Centrifuges

Several types of centrifuges are used in the clinical laboratory, including (1) horizontal head, (2) swinging bucket, (3) fixed-angle head, (4) angle head, (5) ultracentrifuge, and (6) axial units. In addition, the development of automatic balancing centrifuges has enabled centrifugation to be incorporated as an integral step in the total automation of laboratory testing (see Chapter 16).

Principles of Centrifugation

The correct term to describe the force required to separate two phases in a centrifuge is **relative centrifugal force (RCF)**. Units are expressed as number of times greater than gravity (e.g., 500 × g).

RCF is calculated as follows:

$$\text{RCF} = 1.118 \times 10^{-5} \times r \times \text{rpm}^2$$

where 1.118×10^{-5} = empirical factor; r = radius in centimeters from the center of rotation to the bottom of the tube in the rotor cavity or bucket during centrifugation; and rpm = the speed of rotation of the rotor in revolutions per minute.

The RCF of a centrifuge may also be determined from a nomogram distributed by manufacturers of centrifuges. RCF is derived from the distance from the rotor center to the bottom of the tube, whether the tube is horizontal to, or at an angle to, the rotor center.

The time required to sediment particles depends on the (1) rotor speed, (2) radius of the rotor, and (3) effective path length traveled by the sedimented particles (i.e., the depth of the liquid in the tube). Duplication of conditions of centrifugation is often desirable. The following is a useful formula for calculating the speed required of a rotor whose radius differs from the radius with which a prescribed RCF is originally defined:

$$\text{rpm (alternate rotor)} = 1000 \times \sqrt{\frac{\text{RCF, original rotor}}{11.18 \times [r(\text{cm}), \text{ alternate rotor}]}}$$

The length of time for centrifugation is calculated so that running with an alternate rotor of a different size is equivalent to running with the original rotor:

$$\text{time (alternate rotor)} = \frac{\text{time} \times \text{RCF (original rotor)}}{\text{RCF (alternate rotor)}}$$

Note, however, that it may not be possible to reproduce conditions exactly when a different centrifuge is used. Descriptions of times of centrifugation include the time for the rotor to reach operating speed (which may vary from instrument to instrument) and do not include deceleration time, during which sedimentation is still occurring but is doing so less efficiently. Even with maximal braking, deceleration may take as long as 3 minutes in some centrifuges.

Operation of the Centrifuge

For proper operation of a centrifuge, only those tubes recommended by their manufacturer should be used. The material used for the tube must withstand the RCF to which the tube is likely to be subjected. Polypropylene tubes generally are capable of withstanding RCFs of up to 5000 × g. The tubes should have a tapered bottom, particularly if a supernatant is to be removed, and should be of a size to fit securely into the rack to be centrifuged. The top of the tube should not protrude so far above the bucket that the swing into a horizontal position is impeded by the rotor.

For smooth operation of the centrifuge, the rotor must be properly balanced. The weight of racks, tubes, and their contents on opposite sides of a rotor should not differ by more than 1% or by an acceptable limit established by the manufacturer. Centrifuges that automatically balance their rotors are now available.

Tubes of collected blood should be centrifuged before they are unstoppered to reduce the probability of an aerosol being produced when the tube is opened. The practice of using a wooden applicator to release a clot stuck to the top of the tube or to its stopper should be avoided; it is a potential cause of hemolysis. Centrifugation at an appropriate RCF usually ensures that the clot is released from the tube wall and is drawn to the bottom of the tube.

Specific recommendations for each type of collection tube regarding RCF or times for centrifugation of blood specimens are usually provided by the manufacturer. Standards have not been established for centrifugation of other specimens, such as serum to which a protein precipitant has been added. Additional details pertaining to the handling and processing of blood specimens can be found in Chapter 6 and CLSI document GP44-A4.

Operating Practice

Cleanliness of a centrifuge is important in minimizing the possible spread of infectious agents, such as hepatitis viruses. With proper operation of a centrifuge, fewer tubes will break. In case of breakage, the racks and chamber of the centrifuge must be carefully cleaned. Any spillage should be considered a possible bloodborne pathogen hazard. Gray dust arising from sandblasting of the chamber by fragments of glass indicates tube breakage and possible contamination, necessitating cleaning of the chamber. Broken glass embedded in cushions of tube holders may be a continuing cause of breakage if cushions are not inspected and replaced as part of the cleanup procedure.

The speed of a centrifuge should be checked according to the manufacturer's instructions or according to its intended use. The measured speed should not differ by more than 5% from the rated speed under specified conditions. All the speeds at which the centrifuge is commonly operated should be checked. The centrifuge timer should be checked weekly against a reference timer (such as a stopwatch), and the timer should not be more than 10% in error. Commutators and brushes should be checked depending on the frequency of use. Brushes (where used) should be replaced when they show considerable wear. However, in many modern induction-driven motors, brushes have been eliminated, thus removing a source of dust that causes motor failure.

Because centrifuges generate heat, the temperature in the chamber in many centrifuge models may increase by as much as 5°C after a single run. When the material to be centrifuged is thermolabile, a refrigerated centrifuge should be used. In the simplest form, a refrigerator unit is mounted beside the centrifuge, and cold air is blown into the rotor chamber. This approach is usually inadequate to stabilize the low temperature. In more sophisticated centrifuges, refrigeration coils around the chamber make it possible to maintain a preset temperature within ±1°C. The temperature of a refrigerated centrifuge should be measured monthly under reproducible conditions and should be within 2°C of the expected temperature.

Gravimetry

Mass is an invariant property of matter. Gravimetry is the process used to measure the mass of a substance. Weight is a function of mass under the influence of gravity, a relationship expressed by the relationship

$$\text{Weight} = \text{Mass} \times \text{Gravity}$$

Two substances of equal weight and subject to the same gravitational force have equal mass. Mass is determined using a balance to compare the mass of an unknown with that of a known mass. This comparison is called weighing, and the absolute standards with which masses are compared are called weights. In practice, the terms mass and weight are used synonymously.

The classic form of a balance is a beam poised on an agate knife-edge fulcrum, with a pan hanging from each end of the beam and a rigid pointer hanging from the beam at the poised point. With the object to be weighed on one pan and weights of equal mass on the other pan, the pointer comes to rest at an equilibrium or balance point between the extremes of the path of excursion. The weight required to achieve the equilibrium is therefore equal to the weight of the substance being weighed.

Principles of Weighing

In practice, one of two modes of weighing is used: (1) analytical weights are added to equal the weight of the object being weighed, or (2) the material to be weighed is added to a balance pan to achieve equilibrium with a preset weight. This second mode is used more commonly in clinical laboratories, where the major necessity is to weigh a fixed quantity of chemical so that a calibrator or reagent solution of known concentration may be prepared. Before a sample of the chemical is weighed, the weight of the container must be determined to subsequently allow for deducting the weight of the container from the gross weight of the container plus the sample to obtain the net weight of the sample. Most modern balances allow for the readout to be set to "0" with an empty container on the pan. This is called *taring*. When taring is impractical, the weight of the empty container must be subtracted from the combined weight of the container and the material to obtain the weight of the material alone.

Types of Balances

Double- and single-pan and electronic balances are frequently used in the clinical laboratory.

Double-pan balance. A double-pan balance conforms to the classic design, consisting of a single beam with arms of equal length. Standard weights are usually added by hand to one pan to counterbalance the weight of the object on the other; but in some models, a dial or vernier with a chain is used to make fine adjustments to the mass associated with the right-side pan. Double- and triple-beam balances are forms of the unequal-arm balance. These balances may be coupled directly to a computer or recording device.

Single-pan balance. The single-pan balance is a commonly used balance in the clinical laboratory. It is most often electronically operated and self-balancing. In single-pan balances, the arms are of unequal length. The object to be weighed is placed on the pan attached to the shorter arm. A restoring force is applied mechanically or electronically to the other arm to return the beam to its null position. In the electronic single-pan balance, a load on the pan causes the beam to tilt downward. A null detector senses the position of the beam and indicates when the beam has deviated from the equilibrium point.

Electronic balance. An electronic balance, as mentioned previously, uses an electromagnetic force to return the balance beam to its null position. This force is proportional to the weight on the pan. Most electronic balances have a built-in provision for taring so that the mass of the container is subtracted easily from the total mass measured. In addition, in many modern balances, a built-in computer compensates for changes in temperature and provides both automatic zero tracking and calibration.

Analytical Weights

Analytical weights are used to counterbalance the weight of objects weighed on two-pan balances and to verify the performance of both single- and two-pan balances. The NIST recognizes five classes of analytical weights. Class S weights are used for calibrating balances. In the clinical laboratory, balances should be calibrated at least monthly and before very accurate analytical work is conducted. These weights are typically made from brass or stainless steel, and are lacquered

or plated for protection. The fractional weights of a set of class S standards are usually made of platinum or aluminum. Tolerances of the different weights have been defined by the NIST. For class S weights from 1 to 5 g, the tolerance is ±0.054 mg; from 100 to 500 mg, it is ±0.025 mg; and from 1 to 50 mg, it is ±0.014 mg.

Thermometry

In the clinical chemistry laboratory, measurements of temperature are made primarily to verify that devices measure within their prescribed temperature limits. Water baths and heated cells where reactions take place are examples of such devices, as are refrigerators, whose temperatures must be measured and recorded daily to meet laboratory regulatory requirements.

The two most popular types of thermometers in the chemistry laboratory are liquid-in-glass thermometers and thermistor probes.

All thermometers must be verified against a certified thermometer before they are placed into use. For example, the NIST SRM 934 is a mercury-in-glass thermometer with calibration points at 0°C, 25°C, 30°C, and 37°C. Some manufacturers supply liquid-in-glass thermometers that have ranges greater than the SRM thermometer and are verified to have been calibrated against the NIST thermometers. Details of verification of the calibration of a thermometer have been described. The NIST also supplies several materials that melt at a known temperature, including gallium (SRM 1968), which melts at 29.7723°C, and rubidium (SRM 1969), which melts at 39.3°C.

Procedures for Processing Solutions

Several procedures are used routinely to process solutions in the clinical laboratory, including those used for diluting, concentrating, and filtering solutions.

Dilution

Dilution is the process by which the concentration or activity of a given solution is decreased by the addition of a solvent. In laboratory practice, most dilutions are made by transferring an exact volume of a concentrated solution into an appropriate flask and then adding water or other diluent to the required volume, with appropriate mixing to ensure homogeneity. A serial dilution is a sequential set of dilutions in mathematical sequence. A given dilution is expressed as the amount, either volume or weight, of a solute (**analyte**) in a specified volume. For example, a 1:5 volume-to-volume dilution contains one volume of an initial solution in a total of five volumes of final solution (one volume plus four volumes). The dilution factor for a diluted solution is the reciprocal of the dilution of the initial solution. The dilution ratio is typically the ratio of the parts of volume (V_1) to the parts of another volume (V_2) or 1:4 in the previous example.

To prevent errors that arise when two liquids of different composition are mixed, the technique of diluting to volume is used. Instead of adding 90 mL of water to 10 mL of concentrated solution, the 10 mL of concentrated solution should be pipetted into a 100-mL volumetric flask. Water is added to bring the volume to the 100-mL mark on the neck of the flask.

When a dilution is performed, the following equation is used to determine the final volume (V_2) necessary to dilute a given volume (V_1) of solution of a known concentration (C_1) to the desired lesser concentration (C_2):

$$C_1 \times V_1 = C_2 \times V_2 \tag{8.1}$$

or

$$V_2 = \frac{C_1 \times V_1}{C_2} \tag{8.2}$$

Likewise, Eq. 8.1 is also used to calculate the concentration of the diluted solution when a given volume is added to the starting solution.

Evaporation

Evaporation is a process used to convert a liquid or a volatile solid into vapor. It is used in the clinical laboratory to remove liquid from a sample, thereby increasing the concentrations of analyte(s) left behind. A centrifugal evaporator is used to gently evaporate solvents from single or multiple samples. The components include a centrifuge, a vacuum pump, and a cold trap to collect evaporated solvent. A centrifugal force is applied to create a pressure gradient so samples boil from the top down to prevent "bumping." As the solvent is removed, the samples increase in concentration. The evaporation of solvent can also be accomplished by gently blowing a steady stream of gas, typically with its flow regulated through a needle, over the solution. Dry nitrogen gas is commonly used because it is relatively nonreactive.

Lyophilization

Lyophilization (also known as *freeze drying*) is used in laboratory medicine for the preparation of (1) calibrators, (2) control materials, (3) reagents, and (4) individual specimens for analysis. Lyophilization entails first freezing a material at −40°C or less and then subjecting it to a high vacuum. Very low temperatures cause the ice to sublimate to a vapor state. The solid nonsublimable material remains behind in a dried state.

Filtration

Filtration is defined as the passage of a liquid through a filter and is accomplished by gravity, pressure, or vacuum. Filtrate is the liquid that has passed through the filter, and the fraction that does not pass through the filter is the retentate. The purpose of filtration is to remove particulate matter from the liquid. Many filtrations in the clinical laboratory are carried out with filter paper and with plastic membranes of controlled pore size.

Membrane filters are used (1) under vacuum, (2) with positive pressure, or (3) with gravity. Filters have been incorporated into certain disposable tips for use with semiautomatic pipettes. These filters minimize the exchange of aerosol droplets between the tips and the pipette. This is of particular importance for DNA amplification and microbiological procedures. Other membrane filters are designed for

ultrafiltration and are available with a variety of pore sizes for selective filtration. Ultrafiltration is a technique for removing dissolved particles using an extremely fine filter. It is used to concentrate macromolecules, such as proteins, because smaller dissolved molecules pass through the filter.

SAFETY

In the United States, the Federal Occupational Safety and Health Act of 1970 marked the beginning of formal regulatory oversight of employee safety. Since 1970, the **Occupational Safety and Health Administration (OSHA)** and the Centers for Disease Control and Prevention (CDC) have published numerous safety standards that apply to clinical laboratories. Each year, as The Joint Commission (JC) and CAP revise their guidelines, more attention is devoted to safety. Consideration for the health as well as responsibility for the safety of employees are now accepted as obligations of all employers and laboratory directors. In May 1988, OSHA expanded the Hazard Communication Standard to apply to hospital workers. Part of this standard is frequently referred to as the "Lab Right to Know Standard."

Outside of the United States, the regulations and practices regarding laboratory safety and quality are guided by ISO 15189:2012 Medical laboratories: Requirements for quality and competence. This comprehensive international standard provides guidance on management, staffing, customers, laboratory, quality control, and quality improvement.

There are many aspects to the safe operation of a clinical laboratory. Key elements for safety in the clinical laboratory include:

1. A formal safety program.
2. Documented policies and effective use of mandated plans and/or programs in the areas of chemical hygiene, control of exposure to bloodborne pathogens, TB control, and **ergonomics**.
3. Identification of significant occupational hazards, such as biological, chemical, fire, and electrical hazards, and clear identification and documentation of policies for employees to deal with each type of hazard (e.g., packaging and shipping of diagnostic specimens and infectious substances).
4. Recognition that there are additional important and relevant safety areas of concern. These areas include effective waste management and bioterrorism and chemical terrorism response plans in the event of potential threats or casualties involving these types of agents.

Safety Program

Every clinical laboratory must have and implement a comprehensive formal safety program. Regardless of the size of the clinical laboratory, a specific individual should be designated as the "Safety Officer" or "Chair of the Safety Committee" and given the responsibility to implement and maintain a safety program. Safety is each employee's responsibility, but responsibility for the entire program begins with the laboratory leadership (directors, administrative directors, supervisors, managers, and so on) and is delegated through the leadership to the safety officer or safety committee. This individual or committee then has the duties of providing guidance to laboratory leadership on matters related to the provision of a safe workplace for all employees. Although a small institution may have one individual who deals with all safety-related matters for all departments, including the laboratory, OSHA mandates that the laboratory specifically have a chemical hygiene officer, who is designated on the basis of training or experience to provide technical guidance in the development of the **chemical hygiene plan (CHP)**, which is discussed later.

An integral part of the laboratory safety program is the education and motivation of all laboratory employees in all matters related to safety. All new employees should be given a copy of the general laboratory safety manual as part of their orientation. The continuing education program of the laboratory should include periodic talks on safety. Safety information is available from a variety of sources to support the continuing education element of the safety program.

Another important part of the laboratory safety program relates to ensuring that the laboratory environment meets accepted safety standards. This would include, but would not be limited to, attention to such items as (1) proper labeling of chemicals, (2) types and locations of fire extinguishers, (3) hoods that are in good working order, (4) proper grounding of electrical equipment, (5) ergonomic issues (which include equipment, such as pipetting devices, laboratory furniture, and prevention of musculoskeletal disorders), and (6) providing means for the proper handling and disposal of biohazardous materials, including all patient specimens.

Safety Equipment

OSHA requires that institutions provide employees with all necessary personal protective equipment (PPE). Key important safety items are (1) clothing (such as laboratory coats, gowns, and/or scrubs), (2) gloves, and (3) eye protection. These safety items should be used in areas where they are appropriate. Eye washers or face washers should be available in every chemistry laboratory. Many types are available, and some simply connect to existing plumbing. A handheld eye and/or face safety spray is a requisite safety device and is typically placed in a position next to each sink using only a few inches of space. Safety showers, strategically located in the laboratory, must be available and should be tested on a regular schedule.

Heat-resistant (nonasbestos) gloves should be available for handling hot glassware and dry ice. Safety goggles, glasses, and visors, including some that fit conveniently over regular eyeglasses, are available in many sizes and shapes. Personnel wearing contact lenses should be aware of the danger of irritants getting under a lens, making it difficult to irrigate the eye properly. Shatterproof safety shields should be used in front of systems posing a potential danger because of implosion (vacuum collapse) or pressure explosions. Desiccator guards should be used with vacuum desiccators. Hot beakers should be handled with tongs. Inexpensive polyethylene pumps are available to pump acids

from large bottles. Spill kits for acids, caustic materials, or flammable solvents come in various sizes. Such kits and the other appropriate safety materials should be located in convenient and appropriate sites in the laboratory.

A chemical fume hood is a necessity for every clinical chemistry laboratory. In practice, it is used for (1) opening any container of a material that gives off harmful vapors, (2) preparing reagents that produce fumes, and (3) heating flammable solvents. In the event of an explosion or a fire in the hood, closing its window contains the fire.

Safety Inspections

It is good laboratory practice to organize a safety inspection team from the laboratory staff. This team is responsible for conducting periodic and scheduled safety inspections of the laboratory.

In the United States, several regulatory, private accreditation, state, and federal organizations may conduct a safety inspection of the laboratory. Some of these safety inspections may occur unannounced. From an external perspective, OSHA inspectors have the authority to enter a clinical laboratory unannounced, and on presentation of credentials, inspect it. The inspection may be regular or may occur as the result of a complaint or an accident in the lab. In addition, the Commission on Inspection and Accreditation of CAP inspects clinical laboratories and uses various safety checklists (available to the laboratory before inspection) when evaluating a laboratory for accreditation. Although the JC will accept CAP accreditation of a laboratory, it still may conduct a safety inspection of the laboratory when it inspects the hospital. The CAP and JC conduct their accreditation inspections, which may include a full laboratory or laboratory safety component, unannounced.

Depending on the group designated as responsible for accrediting a particular laboratory, selected laboratories may be subject to inspections for the purposes of accreditation and/or safety only by state agencies or local Centers for Medicare and Medicaid Services (CMS) groups. Inspections may be made on a regular basis by state or local health departments or by local fire departments to determine compliance with their particular safety requirements. Currently, a laboratory that meets federal or state OSHA requirements is likely to satisfy the standards of any other inspecting agency.

Plans for the Clinical Laboratory

In 1991, OSHA mandated that all clinical laboratories in the United States must have a CHP and an **exposure control plan**. OSHA has since updated their requirements for the exposure control plan to provide new examples of engineering controls and to place significantly greater responsibility on employers to minimize and manage employee occupational exposure to bloodborne pathogens. The CAP and other groups require that an accredited laboratory must have a documented TB exposure control plan that conforms to biosafety guidelines published by the CDC. Elements of the plan must include regularly defined intervals for employee exposure, as well as process controls, to minimize exposure to aerosolization of *Mycobacterium tuberculosis*.

In addition, it is now recognized that the workplace setting of a clinical laboratory exposes employees to the occupational risk of developing various musculoskeletal disorders. As a result, the focus of OSHA on laboratories that have an effective ergonomics program has led to federal, state, and private accreditation groups addressing this area of occupational safety. However, considerable controversy on this issue is ongoing, with a final ergonomics rule published and then withdrawn in 2001.

Chemical Hygiene Plan

Major elements of a CHP include listing of responsibilities for employers, employees, and a chemical hygiene officer. Also, every laboratory must have a complete chemical inventory that is updated annually. A copy of the **Material Safety Data Sheet (MSDS)**, which defines each chemical as toxic, carcinogenic, or dangerous, must be on file, readily accessible, and available to all employees 24 hours a day, 7 days a week. The MSDS contains important information for the benefit of laboratory employees. The chemical manufacturer's information, as supplied on the MSDS, is used to ascertain whether a certain chemical is hazardous. Each MSDS must give the product's identity as it appears on the container label and the chemical and common names of its hazardous components. The MSDS also provides physical data on the product, such as boiling point, vapor pressure, and specific gravity. Easily recognized characteristics of the chemical are also listed on the line for "appearance and odor." Information about hazardous properties is given in detail on the MSDS; this includes fire and explosion hazard data and health-related data such as the threshold limit value, exposure limits, and toxicity values. The threshold limit value is the exposure allowable for an employee during one 8-hour day. It also notes effects of overexposure and details first-aid procedures. Each MSDS also provides information on spill cleanup and disposal procedures, and protective personal gear and equipment requirements.

Exposure Control Plan

OSHA regulations require that each laboratory develop, implement, adhere to, and maintain a plan that ensures the protection of laboratory workers against potential exposure to bloodborne pathogens, and that medical wastes produced by the laboratory are managed and handled in a safe and effective manner. OSHA regulations also place responsibility on employers to implement new developments in exposure control technology, to solicit the input of employees directly involved in patient care in the identification, evaluation, and selection of these work practice controls, and in certain instances to maintain a log for employee percutaneous injuries from sharp devices such as syringe needles. Organizationally, the plan should include sections on (1) purpose, (2) scope, (3) applicable references, (4) applicable definitions, (5) definition of responsibilities, and (6) detailed procedural steps.

When implementing the plan, each laboratory employee must be placed into one of the following three groups:

Group I: a job classification in which all employees have occupational exposure to blood or other potentially infectious materials.

Group II: a job classification in which some employees have occupational exposure to blood or other potentially infectious materials.

Group III: a job classification in which employees do not have any occupational exposure to blood or other potentially infectious materials.

Tuberculosis Control Plan

The purpose of the TB control plan is to prevent the transmission of TB, which occurs when an individual inhales a droplet that contains *M. tuberculosis*. *M. tuberculosis* is aerosolized when an infected individual sneezes, speaks, or coughs, or when an infected sputum specimen is handled (http://www.cdc.gov/tb/publications/guidelines/control_elim.htm). Transmission of and exposure to TB are greatly diminished with (1) early identification and isolation of patients at risk, (2) environmental controls, (3) appropriate use of respiratory protection equipment, (4) education of laboratory employees, and (5) early initiation of therapy.

An effective TB control plan will include determination of exposure at regular intervals for all employees who are at occupational risk. Engineering and work practice controls are particularly important in laboratory areas, such as surgical pathology and microbiology. However, there is a clear risk of exposure from specimens of patients with suspected or confirmed TB in every section of the laboratory, including chemistry.

Pandemic Plan

A pandemic is an epidemic of infectious disease that spreads through populations across a large region or even worldwide. An influenza pandemic is a worldwide outbreak or epidemic caused by an influenza virus. For example, the influenza pandemic of 1918 to 1919 killed between 20 and 50 million people with more than 500,000 deaths reported in the United States alone (http://www.cdc.gov/flu/about/qa/1918flupandemic.htm).

Influenza viruses capable of causing a pandemic (1) must cause human disease, (2) have novel surface antigens, and (3) must be able to be spread effectively from person to person. Beginning in 1997, an outbreak of avian influenza (strain H5N1) affected birds in several countries in Asia, Africa, and Europe. This strain has been shown to cause lethal disease among humans and created concern that it might evolve into a strain of virus capable of causing a pandemic. Subsequently, many states in the United States developed pandemic influenza control or response plans.

A novel influenza A (H1N1) of swine origin was first detected in April 2009. This virus was found to be highly infectious, easily spreading from person to person, and led to an outbreak of illness in the United States and internationally (http://www.cdc.gov/h1n1flu). A global pandemic was declared by the WHO in June 2009.

Because of the potential for exposure of individuals working in health care settings, the CDC has published recommendations for establishing infection control. These recommendations provide guidance for all individuals who may be processing or performing diagnostic testing on clinical specimens from patients with suspected novel influenza A (H1N1) virus infection, or performing viral isolation.

The Ebola virus, which first appeared in Central Africa in 1976, reappeared in West Africa in 2014. The Ebola virus causes an acute, serious illness that can lead to death when untreated. The virus is transmitted from animals to humans and can easily spread through human-to-human transmission. This highly contagious virus can easily become a pandemic situation if not controlled properly. Health care workers and laboratories need to be prepared for these types of infectious agents.

Ergonomics Program

Several areas of occupational risk for development of musculoskeletal disorders have been identified in the clinical laboratory. These include (1) routine laboratory activity, (2) functionality of the workspace (including laboratory floor matting, bright lighting, and noise generation), and (3) equipment design (computer keyboards and displays, workstations, and chairs). One particular laboratory function, pipetting and related pipette design, has received considerable attention. As depicted in Fig. 8.2, pipettes are being designed with the goal of reducing an employee's risk of developing cumulative stress disorders caused by awkward posture, repetitive motion, and repeated use of force.

The CAP requires accredited laboratories to have a comprehensive and defined ergonomics program that is designed to prevent work-related musculoskeletal disorders through prevention and engineering controls. The documented ergonomics plan should include elements of employee training regarding the areas of risk, engineering controls to minimize or eliminate risks, and an assessment process to identify problematic issues for documentation and remediation.

Hazards in the Laboratory

Various types of hazards are encountered in the operation of a clinical laboratory. These hazards must be identified and labeled, and work practices developed for dealing with them. The major categories of hazards encountered include (1) biological, (2) chemical, (3) electrical, and (4) fire hazards.

Identification of Hazards

Clinical laboratories deal with each of the nine classes of hazardous materials. These are classified by the United Nations (UN) as (1) explosives, (2) compressed gases, (3) flammable liquids, (4) flammable solids, (5) oxidizer materials, (6) toxic materials, (7) radioactive materials, (8) corrosive materials, and (9) miscellaneous materials not classified elsewhere. Shipping and handling of class 6 toxic materials, specifically

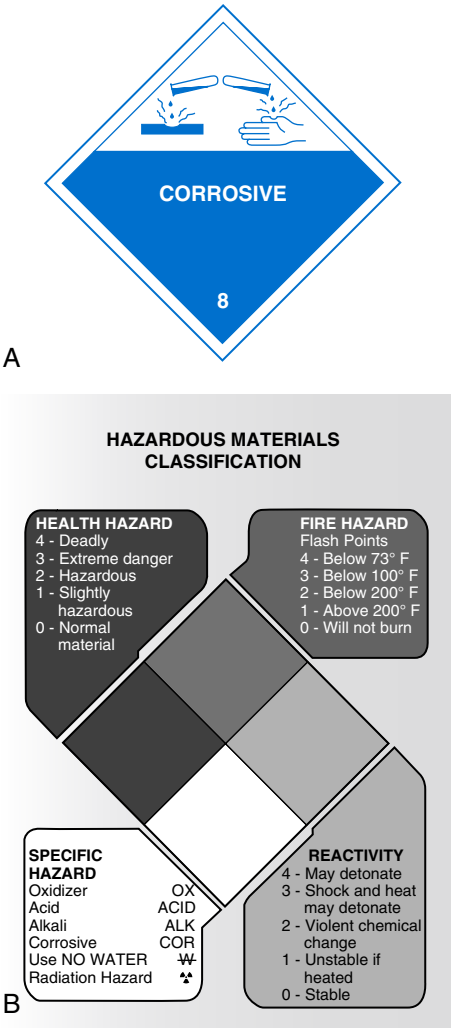


Fig. 8.5 (A) US Department of Transportation label for corrosives. (B) Labeling identification system of the National Fire Protection Association. (Courtesy Lab Safety Supply Inc., Janesville, WI.)

biological and potentially infectious materials, has received considerable attention. In 2002, the US Department of Transportation (DOT) released a revised rule with standards for infectious substance hazardous material handling. Warning labels aid in the identification of chemical hazards during shipment. Under regulations of the DOT, chemicals that are transported in the United States must carry labels based on the UN classification. DOT placards or labels are diamond shaped with a digit imprinted on the bottom corner that identifies the UN hazard class (1 to 9). The hazard is identified more specifically in printed words placed along the horizontal axis of the diamond. Color coding and a pictorial art description of the hazard supplement the identification of hazardous material on the label; the artwork appears in the top corner of the diamond (Fig. 8.5A).

The system is used by the DOT for shipping hazardous materials; however, when the hazardous material reaches its destination and is removed from the shipping container, this identification is lost. The laboratory must then label each individual container. Usually, the information necessary to classify the contents of the container appropriately is on the shipping label and should be noted. Important first aid information is usually provided on this label.

Although OSHA at present prescribes the use of labels or other appropriate warnings, no single uniform labeling system for hazardous chemicals exists for use by clinical laboratories. Appropriate hazard warnings include any words, pictures, symbols, or combinations that convey the health or physical hazards of the container's contents and must be specific as to the effect of the chemical and the specific target organs involved. The National Fire Protection Association (NFPA) has developed the 704-M Identification System, which classifies hazardous material from 0 to 4 (most hazardous) according to flammability and reactivity (instability). This system uses diamond-shaped labels (see Fig. 8.5B), which are available from most companies that sell laboratory safety equipment. The labels are color coded and are divided into quadrants. Three of the quadrants have a characteristic color and represent a type of hazard. A number in the quadrant indicates the degree of hazard. The fourth (lower) quadrant contains information of special interest to firefighters.

Biological Hazards

It is essential to minimize the exposure of laboratory workers to infectious agents, such as the hepatitis viruses, HIV, and flu viruses. Exposure to infectious agents results from (1) accidental puncture with needles, (2) spraying of infectious materials by a syringe or spilling and splattering of these materials on bench tops or floors, (3) centrifuge accidents, and (4) cuts or scratches from contaminated vessels. Any unfixed tissue, including blood slides, must also be treated as potentially infectious material.

OSHA has mandated that all US laboratories have an exposure control plan. In addition, the National Institute for Occupational Safety and Health, a functional unit of the CDC, has prepared and widely distributed a document entitled **Standard Precautions** that specifies how US clinical laboratories should handle infectious agents. In general, it mandates that clinical laboratories treat all human blood and other potentially infectious materials as if they were known to contain infectious agents, such as hepatitis B virus (HBV), HIV, and other bloodborne pathogens. These requirements apply to all specimens of (1) blood, (2) serum, (3) plasma, (4) blood products, (5) vaginal secretions, (6) semen, (7) cerebrospinal fluid, (8) synovial fluid, and (9) concentrated HBV or HIV viruses. In addition, any specimen of any type that contains visible traces of blood should be handled using these Standard Precautions.

Standard Precautions also specify that barrier protection must be used by laboratory workers to prevent skin and mucous membrane contamination from specimens. These barriers, also known as PPE, include (1) gloves, (2) gowns, (3) laboratory coats, (4) face shields or mask and eye protection, (5) mouthpieces, (6) resuscitation bags, (7) pocket masks, and (8) other ventilator devices. A latex allergy is a problem for some individuals when latex gloves

are used for barrier protection. For such individuals, medical grade gloves made of materials such as vinyl, nitrile, neoprene, or thermoplastic elastomer are available. If latex gloves are to be used, they should be made of powder-free, low-allergen latex.

New products for increasing employee protection against needle sticks include an array of novel containers for sharps (e.g., needles, scalpels, glass) and biological safety disposal bags and needle sheaths that may be closed following venipuncture without physically touching the needle or the sheath. Although additional studies on their efficacy and effects on laboratory test results are required, microlaser devices are now available for piercing a patient's skin to collect a capillary blood specimen.

The CLSI has published a similar set of recommendations, several of which are specified as requirements in the OSHA exposure control plan. They include the following:

1. Never perform mouth pipetting and never blow out pipettes that contain potentially infectious material.
2. Do not mix potentially infectious material by bubbling air through the liquid.
3. Barrier protection, such as gloves, masks, and protective eyewear and gowns, must be available and used when drawing blood from a patient and when handling all patient specimens. This includes the removal of stoppers from tubes. Gloves must be disposable, nonsterile latex or of other material to provide adequate barrier protection. Phlebotomists must change gloves and adequately dispose of them between drawing blood from different patients.
4. Wash hands whenever gloves are changed.
5. Facial barrier protection should be used if there is a significant potential for the spattering of blood or body fluids.
6. Avoid using syringes whenever possible and dispose of needles in rigid containers (Fig. 8.6A) without handling them (see Fig. 8.6B).
7. Dispose of all sharps appropriately.
8. Wear protective clothing, which serves as an effective barrier against potentially infective materials. When leaving the laboratory, the protective clothing should be removed.
9. Strive to prevent accidental injuries.
10. Encourage frequent hand washing in the laboratory; employees must wash their hands whenever they leave the laboratory.
11. Make a habit of keeping your hands away from your mouth, nose, eyes, and any other mucous membranes. This reduces the possibility of self-inoculation.
12. Minimize spills and spatters.
13. Decontaminate all surfaces and reusable devices after use with appropriate US Environmental Protection Agency–registered hospital disinfectants. Sterilization, disinfection, and decontamination are discussed in detail in CLSI publication M29-A4.
14. No warning labels are to be used on patient specimens because all should be treated as potentially hazardous.

Fig. 8.6 (A) Convenient needle disposal system for sharps. (B) Needle sheathing devices for prevention of body contact with needle. ([B] Courtesy Marketlab Inc.)

15. Biosafety level 2 procedures should be used whenever appropriate.
16. Before centrifuging tubes, inspect them for cracks. Inspect the inside of the trunnion cup for signs of erosion or adhering matter. Be sure that rubber cushions are free from all bits of glass.
17. Use biohazard disposal techniques (e.g., "Red Bag").
18. Never leave a discarded tube or infected material unattended or unlabeled.
19. Periodically, clean out freezer and dry-ice chests to remove broken ampoules and tubes of biological specimens. Use rubber gloves and respiratory protection during this cleaning.
20. OSHA requires that hepatitis B vaccine be offered to all employees at risk of potential exposure as a regular or

occasional part of their duties. The Advisory Committee on Immunization Practices of the CDC recommends that medical technologists, phlebotomists, and pathologists be vaccinated with hepatitis B vaccine. It is a regulatory mandate that all of the previously mentioned laboratory employees at a minimum be given the option to receive free hepatitis B vaccine.

Investigation of tragic air accidents in the late 1990s by the US National Transportation Safety Board led to the development of revised and strict requirements for the shipping and handling of hazardous materials by DOT, in cooperation with the International Air Transport Association and the International Civil Aviation Organization.

These regulations place particular emphasis on the hazardous material training that must be given to laboratory employees regarding shipping and handling of infectious substances. Elements include general awareness and familiarization, function-specific information, and safety training. Proper training, particularly in the areas of package labeling and documentation (including a shipper's declaration of contents for dangerous goods), is mandatory, and documented certification from employers that the relevant employees have had appropriate training programs is required. Although the adverse impact of improper training can be reflected most by potential human morbidity and mortality, identified violations of these regulations also carry large financial fines and penalties for both the infringing individual and the employer or institution.

Chemical Hazards

Proper storage and use of chemicals is necessary to prevent dangers, such as burns, explosions, fires, and toxic fumes. Thus knowledge of the properties of the chemicals in use and of proper handling procedures greatly reduces dangerous situations. Bottles of chemicals and solutions should be handled carefully, and a cart should be used to transport heavy or multiple numbers of containers from one area to another. Glass containers with chemicals should be transported in rubber or plastic containers that protect them from breakage, and in the event of breakage, contain the spill. Appropriate spill kits should be available in strategic locations.

Spattering from acids, caustic materials, and strong oxidizing agents is a hazard to clothing and eyes and is a potential source of chemical burns. A bottle should never be held by its neck but instead firmly around its body with one or both hands, depending on the size of the bottle. Acids must be diluted by slowly adding them to water while mixing; water should never be added to concentrated acid. When working with acid or alkali solutions, safety glasses should be worn. Acids, caustic materials, and strong oxidizing agents should be mixed in the sink. This provides water for cooling and for confinement of the reagent in the event that the flask or bottle breaks.

All bottles containing reagents must be properly labeled. It is good practice to label the container before adding the reagent, thus preventing the possibility of having an unlabeled reagent. The label should bear the (1) name and concentration of the reagent, (2) initials of the person who made up the reagent, and (3) the date on which the reagent was prepared. When appropriate, the expiration date should also be included. The labels should be color-coded or an additional label added to designate specific storage instructions, such as the requirement for refrigeration or special storage related to a potential hazard. All reagents found in unlabeled bottles should be disposed of using appropriate procedures and precautions.

Strong acids, caustic materials, and strong oxidizing agents should be dispensed by a commercially available automatic dispensing device. Under no circumstances is mouth pipetting permitted.

In some instances, all waste materials are not collected in the same container. With certain pieces of equipment, strong acids or other hazardous materials are pumped directly into the drain. This should always be accompanied by a steady flow of water from the faucet. Safety glasses should be used by instrument operators when acids are pumped under pressure.

Perchloric acid, because it is potentially explosive in contact with organic materials, requires careful handling. Perchloric acid should not be used on wooden bench tops, and bottles of this acid should be stored on a glass tray. Disposal may be accomplished by adding the acid dropwise (using a splatter shield) to at least 100 volumes of cold water and pouring the diluted acid down the drain with large amounts of additional cold water. Special perchloric acid hoods, with special washdown facilities, should be installed if large amounts of this acid are used.

Special care is necessary when dealing with mercury. Even small drops of mercury on bench tops and floors may poison the atmosphere in a poorly ventilated room. The element's ability to amalgamate with a number of metals is well known. After an accidental spillage of mercury, the spill area should be cleaned carefully until no droplets are remaining. All containers of mercury should be kept well stoppered. Because it is highly hazardous, most recommend that no mercury be used in the laboratory.

The Environmental Protection Agency controls the disposal of nonradioactive hazardous wastes. The Resource Conservation and Recovery Act of 1976 states that disposal of materials classifiable within any of the nine UN hazardous materials classes is enforced in such a way that health and safety professionals involved in the disposal of such materials are personally liable for each individual violation.

A CLSI publication covers hazardous waste disposal; however, many municipalities and states have their own regulations. These agencies should be contacted by the laboratory for specifics. Volatile chemicals and compressed gases pose specific hazards.

Hazards from volatiles. The use of organic solvents in a clinical laboratory represents a potential fire hazard and hazards to health from inhalation of toxic vapors or skin contact. These solvents should be used in a fume hood. Storage of organic solvents is regulated by rules set down by OSHA. However, some local fire department rules are more stringent. Solvents should be stored in an OSHA-approved metal storage

cabinet that is properly vented. The maximum working volume of flammable solvents allowed outside storage cabinets is 5 gallons per room. No more than 60 gallons of type I and II solvents may be stored in a single cabinet. No more than three cabinets may be located in each 5000 sq ft of laboratory space.

Vaporization is a major problem in the ignition and spread of fires. Vapors from flammable and combustible liquids and solids form a flammable mixture with air. They are characterized by their flash point, where the flash point is defined as the lowest temperature at which a solvent gives off flammable vapors in the close vicinity of its surface. The mixture at its flash point ignites when exposed to a source of ignition. At temperatures below the flash point, the vapor given off is considered too lean for ignition.

Disposal of flammable solvents in storm sewers or sanitary sewers generally is not allowed. Exceptions are small amounts of those materials that are miscible with water, but even disposal of these should be followed by large amounts of cold water. Other solvents should be collected in safety cans. Separate cans should be used for ether and for chlorinated solvents; all other solvents may be combined in a third can. The cans should be stored, in keeping with storage quantity rules, in a safety cabinet until pickup by a waste disposal firm. A more economical approach is to transfer the solvents to larger cans or drums in an outside storage facility, so that pickup can be less frequent. Some large institutions have their own in-house disposal facilities.

Hazards from compressed gases. DOT regulations cover the labeling of cylinders of compressed gases that are transported by interstate carriers. The diamond-shaped labels described previously are used on all large cylinders and on any boxes containing small cylinders. Some general rules for handling large cylinders of compressed gas are as follows:

1. Always transport cylinders using a hand truck to which the cylinder is secured.
2. Leave the valve cap on a cylinder until the cylinder is ready for use, at which time the cylinder should have been secured by a support around the upper one third of its body. Disconnect the hose or regulator, shut off the valve, and replace the cap before the cylinder is completely empty to prevent the possibility of development of a negative pressure. Place an "empty" sign or label on the cylinder.
3. Chain or secure cylinders at all times even when empty.
4. Always check cylinders for the composition of their contents before connection.
5. Never force threads; if a regulator does not thread readily, something is wrong.

The precautions cited for large, refillable gas cylinders also apply to small cylinders that are not refillable. Propane cylinders and cylinders of calibrating gases for blood gas equipment are examples of disposable cylinders. Cylinders in floor-standing base supports require the additional security of a chain or strap attached to a wall or fixed piece of furniture. Local fire department regulations (which vary considerably from place to place) govern the disposal of exhausted cylinders.

Electrical Hazards

Electrical wires or connections are potential shock or fire hazards. Worn wires on all electrical equipment should be replaced immediately, and all equipment should be grounded using three-prong plugs. OSHA regulations stipulate that the requirements for grounding of electrical equipment of the National Electrical Code (published by NFPA) be met. If grounded receptacles are not available, a licensed electrician should be consulted for proper alternative grounding techniques. Some local codes are more stringent than OSHA requirements and do not allow for two-pole mating receptacles with adapters for a three-pole plug.

Use of extension cords is prohibited. This standard is more stringent than any other existing regulation. In some instances, an extension cord may have to be used temporarily. In such cases, the cord should (1) be less than 12 ft in length, (2) include at least 16 American Wire Gauge wire, (3) be approved by the Underwriters Laboratory, and (4) have only one outlet at the end. If several outlets are necessary in a single area, a power strip with its own fuse or circuit breaker may be installed at least 3 in above bench top level. Several manufacturers now sell devices to check for high resistance in neutral or ground wiring or excess voltage in neutral wiring.

Electrical equipment and connections should not be handled with wet hands, nor should electrical equipment be used after liquid has been spilled on it. The equipment must be turned off immediately and dried thoroughly. In case of a wet or malfunctioning electrical instrument that is used by several people, the plug should be pulled and a note cautioning coworkers against use should be left on the instrument.

Fire Hazards

NFPA and OSHA publish standards covering subjects from emergency exits (including means of egress) to safety and firefighting equipment. NFPA also publishes the National Fire Codes. Many state and local agencies have adopted these codes (some of which are more stringent than OSHA requirements), thus making them legally enforceable.

Every laboratory should have the necessary equipment to extinguish or confine a fire in the laboratory and to extinguish a fire on the clothing of an individual. Easy access to safety showers is essential. A safety shower should have a pull chain attached to the wall at a convenient height or hanging down from the shower head; the chain should have a large ring attached so that the shower may be easily activated, even with eyes closed. Fire blankets for smothering fire on clothing should be available in an easily accessible wall-mounted case. The blanket is unrolled from the case and is rolled around the body by taking hold of the rope that is attached to the blanket and turning the body around. The location of this equipment and the locations of fire alarms and maps of evacuation routes are dictated by the local fire marshal.

Various types of fire extinguishers are available. The type to use depends on the type of fire. Because it is impractical to

TABLE 8.2 Classification of Fires and Fire Extinguisher Requirements

Type of Hazard	Class of Fire	Recommended Extinguisher Agents
Ordinary combustibles: wood, cloth, paper	A	Water, dry chemical foam, loaded steam
Flammable liquids and gases: solvents and greases, natural or manufactured gases	B	Dry chemical, carbon dioxide, loaded steam, Halon 1211 or 1301 foam
Electrical equipment: any energized electrical equipment; if electricity is turned off at source, this reverts to a class A or B	C	Dry chemical, carbon dioxide, Halon 1211 or 1301 foam
Combinations of ordinary combustibles and flammable liquids and gases	A and B	Dry chemical, loaded steam, foam
Combinations of ordinary combustibles and electrical equipment	A and C	Dry chemical
Combinations of flammable liquids and gases and electrical equipment	B and C	Dry chemical, carbon dioxide, Halon 1211 or 1301 foam
Combinations of ordinary combustibles, flammable liquids and gases, and electrical equipment	A, B, and C	Triplex dry chemical

have several types of fire extinguishers present in every area, dry chemical fire extinguishers are among the best all-purpose extinguishers for laboratory areas. An extinguisher should be provided near every laboratory door, and in a large laboratory, at the end of the room opposite the door. Everyone in the laboratory should be instructed in the use of these extinguishers and any other available firefighting equipment. All fire extinguishers should be tested by qualified personnel at intervals specified by the manufacturer. The three classes of fires and the type of fire extinguisher to be used for each are listed in Table 8.2. Every fire extinguisher is labeled as to the type of fire it should be used to extinguish.

Two additional types of fires, designated "D" and "E," should be handled only by trained personnel. Type D fires include those involving powdered metal materials (e.g., magnesium). A special powder is used to fight this hazard. A type E fire is one that cannot be put out or is liable to result in a detonation (such as an arsenal fire). A type E fire is usually allowed to burn out while nearby materials are appropriately protected.

Many clinical laboratories now have a computer that is housed in a temperature- and humidity-controlled room. The most popular automatic fire control system used for these rooms is Halon 1301 (bromotrifluoromethane). Although this is the least toxic of the halons, NFPA regulations require a warning sign at the entrance to the room and availability of self-contained breathing equipment.

POINTS TO REMEMBER

Safety Elements

Safety in the clinical laboratory contains many elements, including:

- A formal safety program
- A CHP
- A plan for exposure to bloodborne pathogens
- A plan for TB control
- An ergonomics plan
- Identification of significant occupational hazards
- A waste management plan
- A biochemical and chemical terrorism response plan

REVIEW QUESTIONS

1. One aspect of Standard Precautions, and an Occupational Safety and Health Administration (OSHA) requirement, involves the provision of personal protective equipment (PPE) to laboratory employees. PPE includes (but is not limited to) which of the following?
   a. Gloves only
   b. Gloves and a lab coat only
   c. Gloves, eye protection, and a lab coat
   d. Gloves, eye protection, a lab coat, and special footwear
2. Which of the following uses for analytical weights is false?
   a. Analytical weights are used for verifying performance of single pan balances.
   b. Class S weights are used for calibrating balances.
   c. Balances must be calibrated each day of use for accurate analytical work.
   d. Weights can be made of materials such as brass and platinum.
3. The organization that develops and implements standards and guidelines for laboratories in addition to setting certain safety guidelines is:
   a. CAP.
   b. CLSI.
   c. EPA.
   d. DOT.
4. The process by which the concentration or activity of a given solution is decreased by the addition of solvent is referred to as:
   a. solution.
   b. evaporation.
   c. lyophilization.
   d. dilution.
5. The formula used to determine the molarity of a solution is which of the following?
   a. $C_1V_1 = C_2V_2$
   b. number of moles of solute/number of liters of solution

c. mg/L divided by mg molecular mass
d. mass in grams/gram molecular weight in grams

6. Which type of highly purified material needs to minimally be 99.98% pure?
   a. Certified Reference Material
   b. Standard Reference Material
   c. Secondary Reference Material
   d. Primary Reference Material
7. Which of the following statements regarding water purification is *incorrect?*
   a. Water treated by distillation alone does not meet the specific conductivity requirement for Clinical Laboratory Reagent Water (CLRW).
   b. Reverse osmosis forces water through a semipermeable membrane.
   c. Only one purification process is necessary for obtaining CLRW.
   d. UV oxidation can eliminate many bacteria and cleave ionizing organics.
8. The two chemical grades that are suitable for use in a clinical chemistry laboratory are:
   a. technical and analytical reagent grades.
   b. ultrapure and technical grades.
   c. technical and National Formulary grades.
   d. ultrapure and analytical reagent grades.
9. Relative centrifugal force or field (RCF) is measured in which of the following units?
   a. Gravities (g)
   b. Centimeters (cm)
   c. rpm
   d. Forces (f)
10. Protection of laboratory workers against potential exposure to bloodborne pathogens and assurance that medical waste is handled safely are OSHA requirements that are part of which of the following plans?
   a. Chemical hygiene plan
   b. Exposure control plan
   c. Tuberculosis control plan
   d. Fire safety plan

SUGGESTED READINGS

Armbruster, D., & Miller, R. R. (2007). The Joint Committee for Traceability in Laboratory Medicine (JCTLM): a global approach to promote the standardisation of clinical laboratory test results. *The Clinical Biochemist Reviews*, *28*, 105–114.

Centers for Disease Control and Prevention. Interim biosafety guidance for all individuals handling clinical specimens or isolates containing 2009-H1N1 influenza A virus (Novel H1N1), including vaccine strains <http://www.cdc.gov/h1n1flu/guidelines_labworkers.htm>(Accessed April 2017)

Clinical and Laboratory Standards Institute. (2009). *Laboratory instrument implementation, verification, and maintenance. CLSI Document GP31-A*. Wayne, PA: Clinical and Laboratory Standards Institute.

Clinical and Laboratory Standards Institute. (2010). *Procedures for the handling and processing of blood specimens: approved guideline* (4th ed.). CLSI Document GP44–A4. Wayne, PA: Clinical and Laboratory Standards Institute.

Clinical and Laboratory Standards Institute. (2011). *Clinical laboratory waste management: approved guideline* (3rd ed). CLSI Document GP5-A3. Wayne, PA: Clinical and Laboratory Standards Institute.

Clinical and Laboratory Standards Institute. (2012). *Clinical laboratory safety* (3rd ed.). CLSI Document GP17–A3. Wayne, PA: Clinical and Laboratory Standards Institute.

Clinical and Laboratory Standards Institute. (2012). *Preparation and testing of reagent water in the clinical laboratory: approved guideline* (4th ed.). CLSI Document GP40-A4-AMD. Wayne, PA: Clinical and Laboratory Standards Institute.

Clinical and Laboratory Standards Institute. (2014). *Protection of laboratory workers from occupationally acquired infections: approved guideline* (4th ed.). CLSI Document M29–A4. Wayne, PA: Clinical and Laboratory Standards Institute.

Davis, D. L. (2008). *Laboratory safety*. Washington, DC: AACC Press.

De Bièvre, P. (2009). The 2007 International Vocabulary of Metrology (VIM), JCGM 200:2008 (ISO/IEC Guide 99): meeting the need for intercontinentally understood concepts and their associated intercontinentally agreed terms. *Clinical Biochemistry*, *42*, 246–248.

Department of Health and Human Services, & National Institutes of Health. (1999). *Centers for Disease Control and Prevention. Biosafety in microbiological and biomedical laboratories* (4th ed.). Washington, DC: U.S. Government Printing Office.

Department of Transportation. (2002). 49 CFR, Part 171. Hazardous materials: revisions to standards for infectious substances. Final rule. *Federal Register*, *67*(157).

National Institute for Occupational Safety and Health (NIOSH). (1997). *Musculoskeletal disorders and workplace factors: a critical review of epidemiologic evidence for work-related musculoskeletal disorders of the neck, upper extremities, and low back. Centers for Disease Control (NIOSH) Publication No. 97-141*. Atlanta, GA: Centers for Disease Control.

National Institute for Occupational Safety and Health. (2001). Updated U.S. Public Health Service guidelines for the management of occupational exposures to HBV, HCV, and HIV and recommendations for postexposure prophylaxis. *MMWR Recommendations and Reports*, *50*, 1–50.

Poon, L. L., Chan, K. H., Smith, G. J., et al. (2009). Molecular detection of a novel human influenza (H1N1) of pandemic potential by conventional and real-time quantitative RT-PCR assays. *Clinical Chemistry*, *55*, 1555–1558.

US Department of Labor. (2001). Occupational exposure to bloodborne pathogens: needlesticks and other sharps injuries. Final rule. Occupational Safety and Health Administration (OSHA). *Federal Register*, *66*, 5318–5325.

US Department of Labor, & Ergonomics program. Final rule: removal. (2001). Occupational Safety and Health Administration (OSHA). *Federal Register*, *666*, 20403.

Vesper, H. W., & Thienpont, L. M. (2009). Traceability in laboratory medicine. *Clinical Chemistry*, *55*, 1067–1075.

9

Optical Techniques

*Larry J. Kricka, Jason Y. Park**

OBJECTIVES

1. Define the following terms:
   - Absorbance
   - Atomic absorption (AA)
   - Beer's law
   - Electromagnetic radiation
   - Flame emission spectrophotometry
   - Fluorescence
   - Fluorescence polarization
   - Light
   - Nephelometry
   - Percent transmittance
   - Photometry
   - Spectral/natural bandwidth
   - Spectrophotometry
   - Stokes shift
   - Transmittance
   - Turbidimetry
   - Visible/ultraviolet/infrared spectrum
   - Wavelength
2. Describe the relationship between the light transmitted and the light absorbed by a solution of a compound; state this relationship as a mathematical formula.
3. Express Beer's law mathematically, and define each of the components of the formula; calculate the concentration of a substance in solution using Beer's law.
4. Explain how Beer's law is used to create a calibration curve, and list five conditions that must be met before Beer's law can be applied to a measurement.
5. List the basic components of generic single-beam and double-beam spectrophotometers and state the purpose of each of the components; provide and describe two examples, if appropriate, of each component.
6. State the principle of reflectance photometry and the clinical laboratory applications of this technique.
7. State the principle of AA spectrophotometry; list the clinical laboratory applications of this technique.
8. List the components of an AA spectrophotometer and the role each component plays in measurement.
9. Describe and give examples of spectral and nonspectral interferences that might limit measurements made using AA.
10. State the principle of fluorescence; list the clinical laboratory applications of this technique.
11. Express as an equation the derivation of the Beer-Lambert law in relation to fluorescence intensity and concentration, and define each of the components of this equation.
12. Compare the sensitivity of fluorescence intensity measurements with that of absorbance measurements, and list three reasons why fluorescence measurements are more sensitive.
13. Describe fluorescence polarization and its use in the clinical laboratory.
14. List the components of a generic fluorometer; list and describe four types of fluorometers, including their use in the clinical laboratory.
15. List and describe six possible interferences that influence fluorescence measurements.
16. Explain the basic concepts of chemiluminescence and electrochemiluminescence; list the clinical laboratory applications of these techniques.
17. Explain the basic concepts of light scattering, including six factors that affect light scattering.
18. Compare nephelometry and turbidimetry, and list the clinical laboratory applications of these techniques; describe a turbidimeter and a nephelometer.
19. List and describe two possible interferences that might limit light-scattering measurements.

* The authors gratefully acknowledge the original contributions by Dr. Merle A. Evenson and Dr. Thomas O. Tiffany, upon which portions of this chapter are based.

KEY WORDS AND DEFINITIONS

Absorbance (A) The amount of light absorbed as incident light passes through a sample, which is equivalent to log (1/T), where *T* is transmittance. Absorbance and transmittance are inverse properties. For example, when absorbance is low (0.0), the transmittance is high (100%) [A = log (1/1.00)]. Conversely, when transmittance is low (10%), the absorbance is 1.0 [A = 1.0 = log (1/.10)].

Absorptivity (a) A proportionality constant for a compound that is the measure of the absorption of radiant energy at a given wavelength as it passes through a solution of that compound at a concentration of 1 g/L; expressed mathematically as absorbance divided by the product of the concentration of a substance in g/L and the sample path length in centimeters (a = A/bc).

Atomic absorption spectrophotometry A technique in which an element in a sample is dissociated from its chemical bonds (atomized) and placed in an unexcited or ground state (neutral atom); the atom in the ground state is able to absorb radiation and the decrease in radiant energy from the light source transmitted through the sample is measured.

Bandpass The range of wavelengths passed by a filter or a monochromator at one-half the peak transmittance of that filter.

Beer's law A mathematical equation stating that the concentration of a substance is directly proportional to the amount of light absorbed, mathematically expressed as A = abc or A = εbc.

Blank A solution used in photometry/spectrophotometry that is identical to the unknown solution except for the substance to be measured.

Chemiluminescence The emission of light when a molecule returns from an excited or higher energy level to a lower energy level, in which the excitation event is caused by a chemical reaction and not by photo illumination, for example, by the oxidation of an organic compound.

Electrochemiluminescence The emission of light when a molecule returns from an excited or higher energy level to a lower energy level, in which the excitation event is a reaction generated electrochemically on the surface of an electrode.

Flow cytometry The measurement of optical properties of cells or particles made while they pass singly through a measuring apparatus in a flowing fluid stream.

Fluorescence The emission of electromagnetic radiation that occurs when a molecule absorbs light at one wavelength and reemits light at a longer wavelength.

Fluorometry The measurement of emitted fluorescence light that occurs when a molecule absorbs light at one wavelength and reemits light at a longer wavelength.

Light The energy transmitted via electromagnetic waves that are characterized by frequency and wavelength; light is composed of photons whose energy is inversely proportional to the wavelength.

Light scattering The redirection of light that results from the interaction of light with molecules or particles in solution without absorption taking place.

Molar absorptivity (ε) A proportionality constant for a compound that is the measure of the absorption of radiant energy at a given wavelength as it passes through a solution of that compound at a concentration of 1 mol/L; expressed mathematically as absorbance divided by the product of the concentration of a substance in mol/L and the sample path length in centimeters (ε = A/bc).

Nephelometry The detection and measurement of light energy scattered or reflected toward a detector that is not in the direct path of the transmitted light; nephelometers may measure scattered light at right angles to the incident light, or at some angle other than 90 degrees.

Photometry/spectrophotometry The measurement of the luminous intensity of light or the amount of luminous light falling on a surface of a detector; spectrophotometry is the measurement of the intensity of light at varied wavelengths.

Reflectance photometry A spectrophotometric technique in which diffused light illuminates a reaction mixture in a carrier containing a substance of interest, and the intensity of the reflected light is measured and compared with a reference.

Spectral bandwidth The width in nanometers of the spectral transmittance curve at a point equal to one-half of the peak transmittance; used to describe the spectral purity of a filter or other monochromator.

Transmittance The intensity of a light beam that passes through a square cell containing a solution of a compound that absorbs light at a specific wavelength, divided by the intensity of an incident (incoming) light beam, stated as $T = I/I_0$.

Turbidimetry The detection and measurement of a decrease in intensity of an incident beam of light after it passes through a solution of molecules or particles.

Wavelength A characteristic of electromagnetic radiation; the distance between two adjacent maxima in a wave that is measured in nanometers.

Many determinations made in the clinical laboratory are based on measurements of radiant energy (1) emitted, (2) transmitted, (3) absorbed, (4) scattered, or (5) reflected under controlled conditions (Table 9.1). The basic principles involved in such measurements are described in this chapter.

PHOTOMETRY AND SPECTROPHOTOMETRY

Photometry is the measurement of the luminous intensity of a light source or the amount of luminous light falling on a surface from such a source. **Spectrophotometry** is the measurement of the intensity of light at selected wavelengths. Photometric

TABLE 9.1 Scope of Optical Methods

Type	Example
Absorption	Atomic absorption, densitometry, Fourier transform infrared spectroscopy, photometry, spectrophotometry, reflectance photometry, x-ray spectroscopy
Emission	Fluorescence correlation spectroscopy, fluorescence energy transfer spectroscopy, fluorometry, luminometry (light emission from a bioluminescent, chemiluminescent, or electrochemiluminescent reaction), phosphorimetry, time-resolved fluorometry
Polarization	Fluorescence polarization spectroscopy, polarimetry
Scattering	Nephelometry, turbidimetry

measurement was defined originally as the process used to measure light intensity independent of wavelength. However, modern instruments isolate a narrow wavelength range of the spectrum for measurements. Those that use filters for this purpose are referred to as filter photometers, whereas those that use prisms or gratings are called *spectrophotometers*. The primary analytical utility of filter photometry or spectrophotometry is the isolation and use of discrete portions of the spectrum for purposes of measurement.

Basic Concepts

Energy is transmitted via electromagnetic waves that are characterized by their frequency and **wavelength**. Analytically, the term wavelength describes a position within a spectrum. Electromagnetic radiation includes radiant energy that extends from cosmic rays with wavelengths as short as 10 nm up to radio waves longer than 1000 km. However, in this chapter, the term **light** is used to describe radiant energy from the ultraviolet (UV) to the visible light portions of the spectrum (290 to 750 nm).

In addition to possessing wavelength characteristics, light behaves as if it is composed of discrete energy packets called *photons*, whose energy is inversely proportional to the wavelength. For example, UV radiation at 200 nm possesses greater energy than infrared (IR) radiation at 750 nm.

Table 9.2 shows the approximate relationship between wavelengths and color characteristics for the UV, visible, and IR portions of the spectrum.

POINTS TO REMEMBER

Radiant Energy

- Wide distribution in wavelength size, ranging from as small as 10 nm to greater than 1000 km
- The visible portion of the spectrum ranges from ~380 to ~750 nm
- The energy of light is inversely proportional to the wavelength; shorter wavelengths have greater energy
- Light possesses both wavelength and energy packet characteristics; the packets are described as photons

TABLE 9.2 Ultraviolet, Visible, and Short Infrared Spectrum Characteristics Colors

Wavelength, nm	Region Name	Observed[a]
<380	Ultraviolet[b]	Invisible
380–440	Visible	Violet
440–500	Visible	Blue
500–580	Visible	Green
580–600	Visible	Yellow
600–620	Visible	Orange
620–750	Visible	Red
750–2,500	Near-infrared	Not visible
2,500–15,000	Mid-infrared	Not visible
15,000–1,000,000	Far-infrared	Not visible

[a]Owing to the subjective nature of color, the wavelength intervals shown are only approximations.
[b]The ultraviolet (UV) portion of the spectrum is sometimes further divided into "near" UV (220 to 380 nm) and "far" UV (<220 nm). This arbitrary distinction has a practical basis because cuvets made from silica transmit light effectively at wavelengths ≥220 nm.

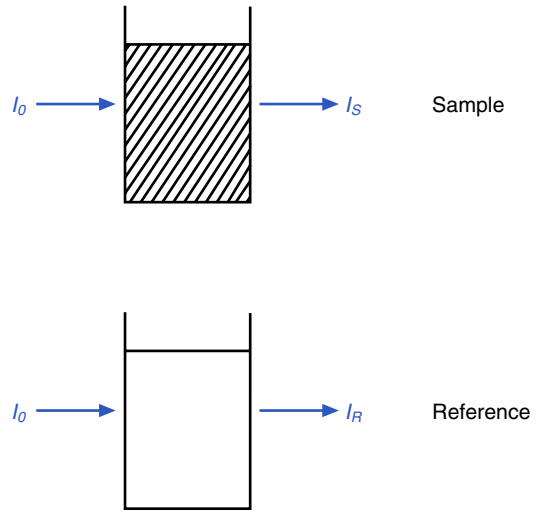


Fig. 9.1 Transmittance of light through sample and reference cells. The transmittance of the sample vs. the reference cell is then I_S/I_R. I_0, Intensity of incident light; I_R, intensity of transmitted light through a reagent blank in the reference cell; I_S, intensity of transmitted light for the compound in solution.

Relationship Between Transmittance and Absorbance

When an incident light beam with intensity I_0 passes through a square cell containing a solution of a compound that absorbs light of a specific wavelength, λ (Fig. 9.1), given that the intensity of the transmitted light beam is I_S, the **transmittance** (T) of light is defined as:

$$T = \frac{I_S}{I_0} \tag{9.1}$$

However, some of the incident light may be reflected by the surface of the cell or absorbed by the cell wall or solvent. These factors are eliminated by using a reference cell identical to the sample cell, except that the compound of interest is omitted from the solvent in the reference cell. The transmittance through this reference cell is I_R divided by I_0; the transmittance for the compound in solution then is defined

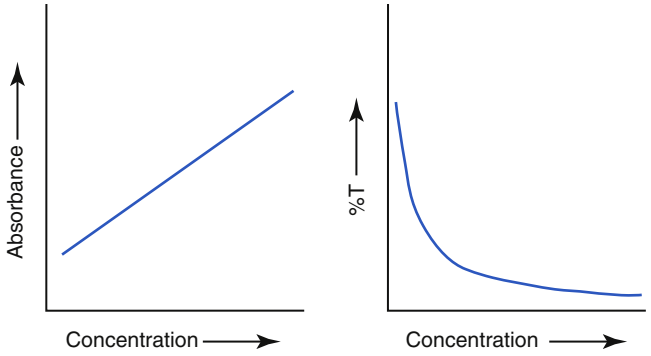


Fig. 9.2 Absorbance and %T relationship.

as I_S divided by I_R. In practice the reference cell is inserted and the instrument adjusted to an arbitrary scale reading of 100 (corresponding to 100% transmittance), after which the percent transmittance reading is made on the sample. The amount of light absorbed (A) as the incident light passes through the sample is equivalent to:

$$A = -\log\left(\frac{I_S}{I_R}\right) = -\log T \quad \textbf{(9.2)}$$

Beer's Law

Beer's law states that the concentration of a substance is directly proportional to the amount of light absorbed:

$$A = abc \quad \textbf{(9.3)}$$

where A = absorbance; a = proportionality constant defined as **absorptivity**; b = light path in centimeters; and c = concentration of the absorbing compound.

Because of the relationship between absorbance and transmittance (Eq. 9.2), concentration is also inversely proportional to the logarithm of the transmitted light (Fig. 9.2). Beer's law forms the basis of quantitative analysis by absorption photometry. **Absorbance (A)** values have no units; hence, the units for a are the reciprocal of those for b and c. When b is 1 cm and c is expressed in moles per liter, the symbol ε (epsilon) is substituted for the constant a. The value for ε is a constant for a given compound at a given wavelength under prescribed conditions of solvent, temperature, pH, etc., and is called the **molar absorptivity (ε)**. The nomenclature of spectrophotometry is summarized in Table 9.3.

Application of Beer's Law

In practice, the direct proportionality between absorbance and concentration must be established experimentally for a given instrument under specified conditions. Frequently, a linear relationship exists up to a certain concentration or absorbance. When this relationship occurs, the solution is said to obey Beer's law up to this point. Within this limitation, a calibration constant (K) may be obtained through measurement of the absorbance of a calibrator of known concentration, and used to calculate the concentration of an unknown solution (c_u) from the measured absorbance (A_u):

$$c_u = A_u \times K \quad \textbf{(9.4)}$$

TABLE 9.3 Spectrophotometry Nomenclature

Name	Symbol	Definition
Absorbance	A	log T or log I_0/I
Absorptivity	a	A/bc (c in g/L)
Molar absorptivity	ε	A/bc (c in mol/L)
Path length	b	Internal cell or sample length, in cm
Transmittance	T	I/I_0[a]
Wavelength unit	nm	10^{-9} m
Absorption maximum	λmax	Wavelength at which maximum absorption occurs

[a] I/I_0 is the ratio of the intensity of transmitted light to incident light.

Certain precautions must be observed with the use of such calibration constants. Under no circumstances should the constant be used when the calibrator or unknown readings exceed the linear portion of the calibration curve (i.e., when the curve no longer obeys Beer's law). At least two or preferably more calibrators should be included in the generation of a calibration curve. A nonlinear calibration curve may be used if a sufficient number of calibrators of varying concentrations are included to cover the entire range encountered for readings on unknowns.

In addition, Beer's law is followed only if the following conditions are met:

- Incident radiation on the substance of interest is monochromatic.
- The solvent absorption is insignificant, compared with the solute absorbance.
- The solute concentration is within given limits.
- An optical interferent is not present.
- A chemical reaction does not occur between the molecule of interest and another solute or solvent molecule.

Measurement Errors

With most photometers, the response of the detector to a signal of transmitted light is such that any uncertainty in %T is constant over the entire %T scale. The uncertainty derives from electrical and mechanical imperfections in the instrument and individual variations in the use of the instrument.

A fixed distance on the linear scale (e.g., 1% T) represents a greater change in absorbance for low values of %T than for high values of %T. For this reason, the absolute concentration error or uncertainty is greater when readings are taken at high absorbance. However, the relative concentration error is greater for readings at both low and high absorbances. The relative error is minimal at an absorbance of 0.434 (36.8% T). Consequently, methods should be designed within an absorbance interval of approximately 0.1 and 0.7 (20% and 80% T).

INSTRUMENTATION

Modern instruments isolate a narrow wavelength range of the spectrum for measurements. Those that use filters for this purpose are referred to as *filter photometers;* those that use prisms

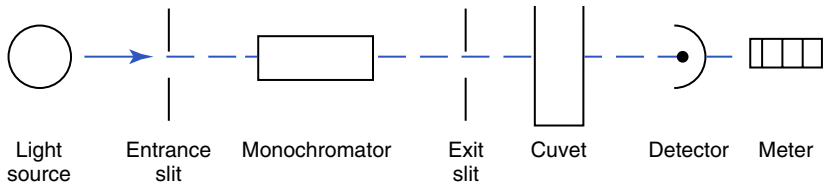


Fig. 9.3 Single-beam spectrophotometer.

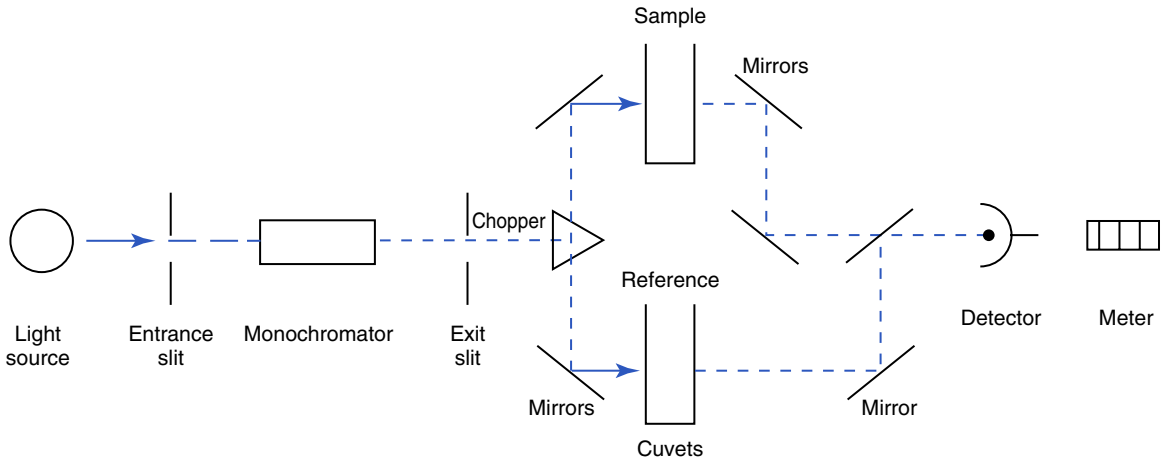


Fig. 9.4 Double-beam spectrophotometer.

or gratings are called *spectrophotometers*. Spectrophotometers are classified as single- or double-beam.

The major components of a single-beam spectrophotometer are shown schematically in Fig. 9.3. With such an instrument, a beam of light is passed through a monochromator that isolates the desired region of the spectrum to be used for measurements. Slits are used to isolate a narrow beam of the light and improve its chromatic purity. The light next passes through an absorption cell (cuvet) where a portion of the radiant energy is absorbed, depending on the nature and concentration of the substance in the solution. Any light not absorbed is transmitted to a detector (photocell or phototube), which converts light energy to electrical energy that is registered on a meter or recorder or digitally displayed.

In manual operation, an opaque block is substituted for the cuvet, so that no light reaches the photocell, and the meter is adjusted to read 0% T. Next, a cuvet containing a reagent **blank** is inserted, and the meter is adjusted to read 100% T (zero absorbance). The composition of the reagent blank should be identical to that of calibrating or unknown solutions except for the absorbing substance to be measured. Calibrating solutions containing various known concentrations of the substance are inserted, and readings are recorded. Finally, a reading is made of the unknown solution, and its concentration is determined by comparison with readings obtained on the calibrators. With most spectrophotometers, digital hardware and software are integral components that automate blanking and calibrating procedures.

Fig. 9.4 illustrates a typical double-beam instrument that uses a light-beam chopper (a rotating wheel with alternate silvered sections and cutout sections) inserted after the exit slit. A system of mirrors passes portions of the light reflected off the chopper alternately through the sample and a reference cuvet onto a common detector. The chopped-beam approach, using one detector, compensates for light source variation and for sensitivity changes in the detector.

Components

The basic components of a spectrophotometer include (1) a light source, (2) a means to isolate light of a desired wavelength, (3) fiber optics, (4) cuvets, (5) a photodetector, (6) a readout device, (7) a computer, and (8) a recorder.

Light Sources

Types of light sources used in spectrophotometers include incandescent lamps, light emitting diodes (LEDs), and lasers.

Incandescent lamps for measuring the visible portion of the spectrum include a tungsten light bulb, a quartz-halogen lamp, hydrogen and deuterium lamps, high-pressure mercury and xenon arc lamps, and zinc hollow cathode lamps.

An LED is a p-n junction diode that emits light when activated. Advantages of this type of light source include a wide range of wavelengths (infrared to UV), low power consumption, and long lifetimes.

A laser (*l*ight *a*mplification by *s*timulated *e*mission of *r*adiation) is a device that may also be used as a light source in spectrophotometers. These devices transform light of various frequencies into an extremely intense, focused, and nearly nondivergent beam of monochromatic light. Through the selection of different materials, different wavelengths of light emitted by the laser are obtained. Examples of laser sources include argon, fluoride, nitrogen, neodymium-YAG (yttrium aluminum garnet), and carbon dioxide lasers.

Spectral Isolation

A system for isolating radiant energy of a desired wavelength and excluding that of other wavelengths is called a *monochromator.* Devices used for spectral isolation include (1) filters, (2) prisms, and (3) diffraction gratings. In addition, combinations of lenses and slits are often inserted before or after the monochromatic device to render light rays parallel, or to isolate narrow portions of the light beam. Variable slits also are used to permit adjustments in total radiant energy reaching the photocell.

Filters. The simplest type of filter is a thin layer of colored glass. Strictly speaking, a glass filter is not a true monochromator because it transmits light over a relatively wide range of wavelengths. The spectral purity of a filter or other monochromator is usually described in terms of its **spectral bandwidth**. This is defined as the width, in nanometers, of the spectral transmittance curve at one-half the peak transmittance. Commonly used glass filters have spectral bandwidths of approximately 50 nm and are referred to as wide **bandpass** filters.

Other glass filters include narrow bandpass and sharp-cutoff types. Operationally, a cutoff filter typically shows a sharp rise in transmittance over a narrow portion of the spectrum and is used to eliminate light below or above a given wavelength. Narrow bandpass filters are constructed by combining two or more sharp-cutoff filters or regular filters.

Interference filters are also used as monochromators. These filters have narrow spectral bandwidths, usually from 5 to 15 nm. Because they also transmit harmonics, or multiples, of the desired wavelength, accessory glass filters are required to eliminate these undesired wavelengths. Thus an interference filter designed for 620 nm will also transmit some radiation at 310 and 1240 nm unless accessory cutoff filters are provided to absorb this undesired stray light.

Prisms and gratings. Prisms and diffraction gratings are also widely used as monochromators. A prism separates white light into a continuous spectrum by refraction with shorter wavelengths bent, or refracted, to a greater extent than longer wavelengths as they pass through the prism. A diffraction grating is prepared by depositing a thin layer of aluminum-copper alloy on the surface of a flat glass plate; then, many small parallel grooves are made in the metal coating. Diffraction gratings are extremely accurate, have low light scatter, and are widely used in the spectrophotometers used in analytical instruments. For example, most UV-visible spectrophotometers and virtually all IR spectrophotometers use reflective gratings. In addition, high performance liquid chromatography (HPLC) detectors frequently use a concave holographic reflective grating in their optical system.

Each groove of the grating, when illuminated, gives rise to a tiny spectrum. Wave fronts are formed that reinforce those wavelengths in phase and cancel those not in phase. The net result is a uniform linear spectrum. Some instruments contain diffraction gratings that produce spectral bandwidths of 20 nm or more; higher-priced instruments may have a resolution of 0.5 nm or less.

A grating may also be ruled at a specified angle, so that a maximum fraction of the radiant energy is directed into wavelengths diffracted at a selected angle. This type of grating is said to have been given a *blaze* at a particular angle or to have been blazed at a certain wavelength (e.g., 250 nm).

Selection of a wavelength isolation device. The type of monochromator chosen depends on the analytical purpose for which it is to be used. For example, narrow spectral bandwidths are required in spectrophotometers for resolving and identifying sharp absorption peaks that are closely adjacent. Lack of agreement with Beer's law will occur when a part of the spectral energy transmitted by the monochromator is not absorbed by the substance being measured. This is more commonly observed with wide bandpass instruments. In practice, an increase in absorbance and improved linearity with concentration are usually observed with instruments that operate at narrower bandwidths of light. This is especially true for substances that exhibit a sharp peak of absorption. The natural bandwidth of an absorbing substance is defined as "the bandwidth of the spectral absorbance curve at a point equal to one-half of the maximum absorbance." As a general rule, for peak absorbance readings to be within 99.5% of true values, the spectral bandwidth should not exceed 10% of the natural bandwidth. For example, many chemistry procedures used in the clinical laboratory produce an absorbing species for which the natural bandwidth ranges from 40 to greater than 200 nm. The natural bandwidth of NADH is 58 nm (λmax = 339 nm). Therefore, for accurate measurements of this compound, an instrument should be used that has a spectral bandwidth of 6 nm or less.

In practice, the wavelength selected is usually at the peak of maximum absorbance to achieve maximum sensitivity; however, it may be desirable to choose another wavelength to minimize interfering substances. For example, turbidity readings on a spectrophotometer are greater in the blue region than in the red region of the spectrum, but the latter region is chosen for turbidity measurements to avoid absorption of light by bilirubin (460 nm) or hemoglobin (417 and 575 nm). In addition, measurements should not be taken on the steep slope of an absorption curve because a slight error in wavelength adjustment will introduce a significant error in absorbance readings.

Fiber Optics

In the single- and double-beam spectrophotometers shown diagrammatically in Figs. 9.3 and 9.4, the positioning of the individual components dictates the path that the light beam must follow as it travels from the source to the detector. This approach places certain restrictions on the (1) design, (2) size, and (3) cost of such instruments. To overcome these restrictions, fiber optics are now integrated into the optical design of spectrophotometers. Fiber optics, also known as *light pipes,* are bundles of thin, transparent fibers of glass, quartz, or plastic that are enclosed in material of a lower index of refraction and that transmit light throughout their lengths by internal reflections. The use of fiber optics in spectrophotometers offers the advantage of better directional control of the beam

of light within the geometrical confines of an instrument. This allows for the design and manufacture of miniature and inexpensive optical subsystems for use in automated instruments. For example, a single light source is multiplexed with multiple detectors by fiber optics for optimal positioning of the source and detectors in an automated system. Disadvantages of fiber optics include (1) greater amounts of stray light; (2) refractive index changes in the glass, quartz, or plastic rods; and (3) loss of transmitted energy after continued use in the UV region of the spectrum. This loss of energy is known as *solarization* and results in a decrease in the optical performance of an instrument.

Cuvets

A cuvet (or cuvette) is a small vessel used to hold a liquid sample to be analyzed in the light path of a spectrometer. Cuvets may be (1) round, (2) square, or (3) rectangular and are constructed from (1) glass, (2) silica (quartz), or (3) plastic. Square or rectangular cuvets have plane-parallel optical surfaces and a constant light path. Most have a 1.0 cm light path, held to close tolerances. Ordinary borosilicate glass cuvets are suitable for measurements in the visible portion of the spectrum. However, for readings below 340 nm quartz cells are usually required. Some plastic cells have good clarity in both visible and UV ranges but often present problems related to (1) tolerance, (2) cleaning, (3) etching by solvents, and (4) temperature deformation. Many plastic cuvets are designed for disposable, single-use applications.

Photodetectors

Photodetectors are devices that convert light into an electrical signal that is proportional to the number of photons striking its photosensitive surface. The photomultiplier tube (PMT) is a photodetector that is commonly used to measure light intensity in UV and visible regions of the spectrum. PMTs (1) have extremely rapid response times, (2) are very sensitive, and (3) are slow to fatigue.

Photodiodes also are used as photodetectors. They are solid-state devices that are fabricated from photosensitive semiconductor materials, such as (1) silicon, (2) gallium arsenide, (3) indium antimonide, (4) indium arsenide, (5) lead selenide, and (6) lead sulfide. These materials absorb light over a characteristic wavelength range (e.g., 250 to 1100 nm for silicon). Their development and use as detectors in spectrophotometers have resulted in instruments capable of measuring light at a multitude of wavelengths. When a photodetector consists of multiple diodes in an array, each of which responds to a specific wavelength, it is known as a *photodiode array.* For example, photodiode arrays have been designed to have a 2 nm resolution per diode from 200 to 340 nm, and a 1 nm resolution per diode from 340 to 800 nm.

Readout Devices

Electrical energy from a detector is displayed on a digital readout device (e.g., a bank of light-emitting diodes) that provides a visual numerical display of absorbance or converted values of concentrations.

Digital Hardware and Software Computers

Various types of digital hardware, with resident software and processors, are incorporated and integrated into both photometers and spectrophotometers. For example, with a resident computer and software (1) output from a calibrator is digitally stored, (2) digital signals from a blank is subtracted from the calibrator and the unknown, and (3) the concentration of unknown is automatically calculated. Data from multiple calibrators often are used to (1) store a complete calibration curve, (2) display or print out the curve for visible inspection, and (3) calculate the results of unknowns based on the curve or some mathematical transformation of the data. Computers and their resident software also are used to convert kinetic data into concentration or enzyme activity.

Recorders

Historically, spectrophotometers were equipped with recorders in addition to or instead of a digital display. These were synchronized to provide line traces of transmittance or absorbance as a function of time or wavelength. When a continuous tracing of absorbance versus wavelength is recorded, the resultant figure is called an *absorption spectrum.* If a substance absorbs light, distinct peaks of absorbance will be observed. Measuring the absorption spectra of an unknown sample and comparing them with spectra from known compounds is very useful for qualitative purposes.

Performance Parameters

In most spectrophotometric analytical procedures, the absorbance of an unknown sample is compared directly with that of a calibrator or a series of calibrators. Under these circumstances,(1) minor errors in wavelength calibration, (2) variations in spectral bandwidths, and (3) the presence of stray light are compensated for and do not usually contribute to serious errors. The use of a series of calibrators covering a wide range of concentrations also provides a measure of linearity and validation of agreement with Beer's law for a given procedure and instrument. However, when calculations are based on published or previously determined values for molar absorptivities or absorption coefficients, the spectrophotometer must be checked more rigorously.

The National Institute of Standards and Technology (NIST) provides several standard reference materials (SRMs) for spectrophotometry that are useful in the calibration or verification of the performance of photometers or spectrophotometers (e.g., SRM 930*e* is used for the verification and calibration of the transmittance and absorbance scales of visible absorption spectrometers) (http://www.nist.gov/srm). The Institute for Reference Materials and Measurements (IRMM), a metrology institute that belongs to the European Commission, also provides reference materials for verification of the performance of photometers or spectrophotometers (https://ec.europa.eu/jrc/en/news/welcome-irmm-reference-materials-catalogue-9438).

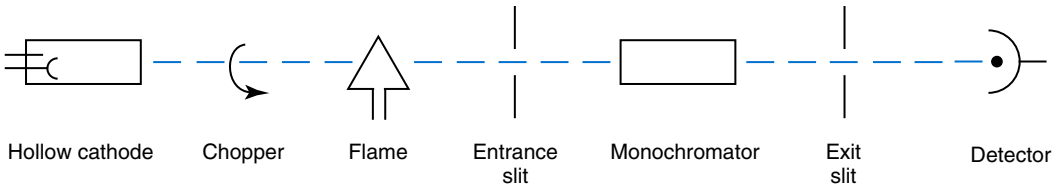


Fig. 9.5 Atomic absorption spectrophotometer.

REFLECTANCE PHOTOMETRY

In **reflectance photometry**, diffused light illuminates a reaction mixture in a carrier, and the reflected light is measured. Alternatively, the carrier is illuminated and the reaction mixture generates a diffuse reflected light, which is measured. The intensity of the reflected light from the reagent carrier is compared with the intensity of light reflected from a reference surface. Because the intensity of reflected light is nonlinear in relation to the concentration of the analyte, transformations are commonly used to convert the data into a linear format. The electro-optical components used in reflectance photometry are essentially the same as those required for absorbance photometry. Reflectance photometry is used as the measurement method with dry-film chemistry systems.

ATOMIC ABSORPTION SPECTROPHOTOMETRY

Atomic absorption (AA) spectrophotometry is used in clinical laboratories to measure elements such as aluminum, calcium, copper, lead, lithium, magnesium, zinc, and other metals. A method for metal analysis that has begun to replace AA is ICP-MS (inductively coupled plasma mass spectrometry). Atomic absorption retains advantages over ICP-MS in terms of overall cost and ease of use. However, a single ICP-MS instrument has greater flexibility in measuring multiple metals simultaneously and with greater sensitivity (see Chapter 13).

Basic Concepts

AA is an optical technique in which an ion in the sample is converted to a ground state atom of that element by a flame or graphite furnace, and introduced into the light path of a spectrophotometer. The ground state atom is capable of absorbing radiation at a very narrow bandwidth corresponding to its own line spectrum. A hollow-cathode lamp with the cathode made of the material to be analyzed is used to produce a wavelength of light specific for the material. Thus, if the cathode was made of calcium, light at predominantly 423 nm would be emitted by the lamp. When the light from the hollow-cathode lamp enters the flame, some of it is absorbed by the ground-state atoms in the flame, resulting in a net decrease in the intensity of the beam from the lamp. This process is referred to as *atomic absorption.* Owing to the unique specificity of the wavelength from the hollow-cathode lamp, these methods are highly specific for the element being measured.

Instrumentation

The components of an AA spectrophotometer are shown in Fig. 9.5. A hollow-cathode lamp serves as the light source for an AA spectrophotometer. Such lamps are made of the metal of the substance to be analyzed; this is different for each metal analysis. When an alloy is used to make the cathode, the result is a multielement lamp.

With flameless AA techniques (carbon rod or "graphite furnace"), the sample is placed in a depression on a carbon rod in an enclosed chamber. Strips of tantalum or platinum metal are used as sample cups. In successive steps, the temperature of the rod is raised to dry, char, and finally atomize the sample in the chamber. The atomized element then absorbs energy from the corresponding hollow-cathode lamp. This approach is very sensitive and permits determination of trace metals in small samples of blood or tissue. However, background absorbance by molecules and particulate matter from the sample may limit the sensitivity of the method.

With flameless AA, background absorption can be corrected by placing the analyte in a strong magnetic field. The intense magnetic field splits the degenerate (i.e., of equal energy) atomic energy levels into two components that are polarized parallel and perpendicular to the magnetic field, respectively. The parallel component is at the resonance line of the source, whereas the two perpendicular components are shifted to different wavelengths. The two components interact differently with polarized light. A polarizer is placed between the source and the atomizer, and two absorption measurements are taken at different polarizer settings. One measures both analyte and background absorptions (A_t), the other only background absorption (A_{bc}). The difference between the two absorption readings is the corrected absorbance.

The major advantage of magnetic correction is that the same light source at the same wavelength is used to measure total and background absorption. Implementation is complex and expensive, and the strength of the magnetic field needs to be optimized for every element, but the method gives more accurate results at higher background levels than are associated with other correction techniques.

Limitations of Atomic Absorption Spectrophotometry

Spectral and nonspectral interferences are limitations of AA spectroscopy.

Spectral Interferences

Spectral interferences include (1) absorption by other closely absorbing atomic species, (2) absorption by molecular species, (3) scattering by nonvolatile salt particles or oxides, and

(4) background emission (which is corrected for by electronic filtering). Absorption by other atomic species usually is not a problem because of the extremely narrow bandwidth (0.01 nm) of the light generated by a hollow cathode lamp that is used in AA instruments. Absorption and scattering by molecular species are particularly problematic at lower atomizing temperatures.

Nonspectral Interferences

Nonspectral interferences may be nonspecific or specific. Nonspecific interferences affect nebulization by altering the (1) viscosity, (2) surface tension, or (3) density of the analyte solution, and consequently the sample flow rate. Specific interferences (chemical interferences) are analyte dependent. *Solute volatilization interference* refers to the situation in which the contaminant forms nonvolatile species with the analyte. An example is seen in phosphate interference in the determination of calcium that is caused by the formation of calcium-phosphate complexes. The phosphate interference is eliminated by adding a cation, usually lanthanum or strontium that competes with calcium for the phosphate. Enhancement effects are also observed, in which the addition of contaminants increases the volatilization efficiency. Such is the case with aluminum, which normally forms nonvolatile oxides, but in the presence of hydrofluoric acid forms more volatile aluminum fluoride. Dissociation interferences affect the degree of dissociation of the analyte. Analytes that form oxides or hydroxides are especially susceptible to dissociation interferences. Ionization interference occurs when the presence of an easily ionized element, such as K, affects the degree of ionization of the analyte, which leads to changes in the analyte signal. In cases of excitation interference, the analyte atoms are excited in the atomizer, with subsequent emission at the absorption wavelength. This type of interference is more pronounced at higher temperatures.

FLUOROMETRY

Fluorescence occurs when a molecule absorbs light at one wavelength and reemits light at a longer wavelength. An atom or molecule that fluoresces is termed a *fluorophore*. **Fluorometry** is defined as the measurement of emitted fluorescence light. Fluorometric analysis is a very sensitive and widely used method of quantitative analysis in the chemical and biological sciences.

Basic Concepts

The relationship between (1) absorption, (2) fluorescence, and (3) phosphorescence is shown in Fig. 9.6. As indicated, each molecule contains a series of closely spaced energy levels. Absorption of a quantum of light energy by a molecule causes the transition of an electron from the singlet ground state to one of a number of possible vibrational levels of its first excited singlet state. The actual number of molecules in the excited state under typical reaction conditions and excited with a typical 150-W light source is very small and is estimated to be about 10^{-13} mole per mole of fluorophore.

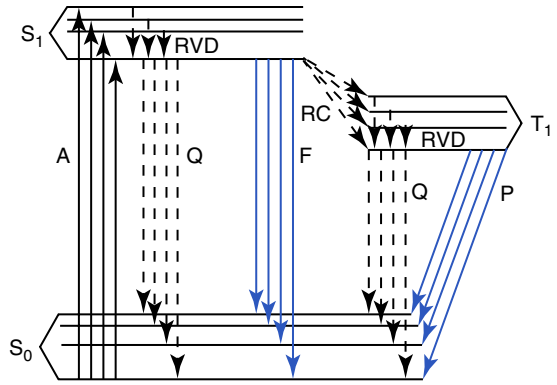


Fig. 9.6 Luminescence energy-level diagram of typical organic molecule. *A,* Absorption process; *F,* fluorescence process from the first excited singlet state; *P,* phosphorescence process from the first excited triplet state; *Q,* quenching of the excited singlet or triplet state; *RC,* radiation-less crossover from the first excited singlet state to the first excited triplet state; *RVD,* radiation-less vibrational deactivation; S_0, ground level singlet state; S_1, first excited singlet state; T_1 is the first excited triplet state.

Once the molecule is in an excited state, it returns to its original energy state by different mechanisms. These include (1) radiation-less vibrational equilibration, (2) the fluorescence process, (3) quenching of the excited singlet state, (4) radiation-less crossover to a triplet state, (5) quenching of the first triplet state, and (6) the phosphorescence process.

As shown in Fig. 9.6, vibrational equilibration before fluorescence results in some loss of excitation energy. The emitted fluorescence light is therefore of less energy and has a longer wavelength than the excitation light. The difference between the maximum wavelength of the excitation light and the maximum wavelength of the emitted fluorescence light is a constant for a given molecule, referred to as the *Stokes shift.* This constant is a measure of energy lost during the lifetime of the excited state (radiation-less vibrational deactivation) before returning to the ground singlet level (fluorescence emission).

Time Relationships of Fluorescence Emission

The time required for a molecule to absorb radiant energy and to be promoted to an excited state is approximately 10^{-15} s. The length of time needed for vibrational equilibration to occur to the lowest excited state is of the order of 10^{-14} to 10^{-12} s. The length of time required for fluorescence emission to occur is of the order of 10^{-8} to 10^{-7} s. Relatively speaking, there is a considerable time delay between (1) absorption of light energy, (2) return to the lowest excited state, and (3) emission of fluorescence light. This time relationship is shown in Fig. 9.7. Phase I represents the time period between the absorbance of light energy and the radiation-less loss of energy during vibrational rearrangement to the lowest excited energy state. This time period is represented by the up and down arrows in the diagram. Phase II shows the emission and decay of a short-lived (b) and a longer-lived (a) fluorophore. If the fluorescence emission is measured over time following a pulse of light from an excitation source, such as a xenon lamp or laser, the intensity of the emitted

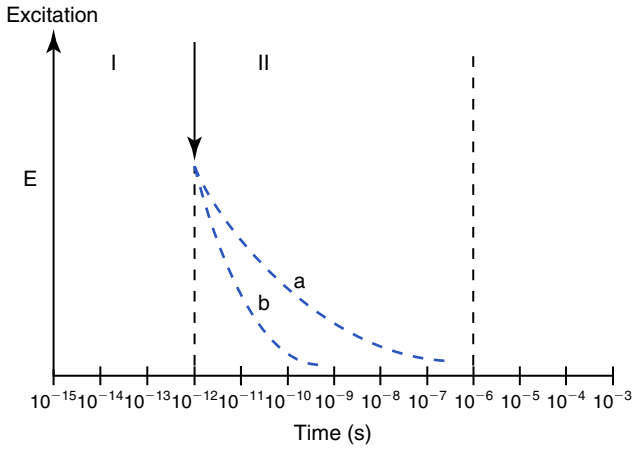


Fig. 9.7 Fluorescence decay process. *a,* Long fluorescence decay time; *b,* short fluorescence decay time; *E,* absorption of energy; *I,* vibrational deactivation time phase; *II,* fluorescence emission time phase.

light decays as a first-order process similar to radioactive decay. The time required for the emitted light to reach 1/e of its initial intensity, where *e* is the Naperian base 2.718, is the average lifetime of the excited state of the molecule, or the fluorescence decay time.

The time delay between absorption of energy and fluorescence is used in a fluorescence instrument called a *time-resolved fluorometer.* Advantages of a time-resolved fluorometer include the elimination of background light scattering and background from short-lived fluorescence. This results in a dramatic increase in the signal-to-noise ratio and detection sensitivity.

Depending on how the fluorescence emission response is measured, time-resolved fluorometry is categorized as pulse or phase fluorometry. In pulse fluorometry, the sample is illuminated with an intense brief pulse of light, and the intensity of the resulting fluorescence emission is measured as a function of time with a fast detector system. In phase fluorometry, a continuous-wave laser illuminates the sample, and the fluorescence emission response is monitored for impulse and frequency response.

Relationship of Concentration and Fluorescence Intensity

The relationship of concentration to the intensity of fluorescence emission is derived from Beer's law and is expressed as:

$$F = \Phi I_0 \varepsilon bc \qquad \textbf{(9.5)}$$

where F = relative intensity; Φ = fluorescence efficiency (i.e., the ratio between quanta of light emitted and quanta of light absorbed); I_0 = initial excitation intensity; ε = molar absorptivity; b = volume element defined by geometry of the excitation and emission slits; and c = the concentration in mol/L.

Eq. 9.5 indicates that fluorescence intensity is directly proportionate to the concentration of the fluorophore and the intensity of the excitation. This relationship holds only for dilute solutions in which absorbance is less than 2% of the exciting radiation. Above 2%, the fluorescence intensity becomes nonlinear. This phenomenon is called the *inner filter effect,* and it is discussed in greater detail in a later section. Other factors influencing the measurement of fluorescence intensity are the sensitivity of the detector and the degree of background light scatter seen by the detector.

Fluorescence intensity measurements are more sensitive than absorbance measurements. The magnitude of absorbance of a chromophore in solution is determined by its concentration and the path length of the cuvet. The magnitude of the fluorescence intensity of a fluorophore is determined by (1) its concentration, (2) the path length, and (3) the intensity of the light source. Comparatively, fluorescence measurements are 100 to 1000 times more sensitive than absorbance measurements. This is due to the use of (1) more intense light sources, (2) digital signal filtering techniques, and (3) sensitive emission photometers.

Frequently, fluorescence measurements are expressed in relative intensity units. The word *relative* is used because the intensity measured is not an absolute quantity. It is a small part of the total fluorescence emission, and its magnitude is defined by (1) instrument slit width, (2) detector sensitivity, (3) monochromator efficiency, and (4) excitation intensity. Because these are instrument-related variables, establishing an absolute intensity unit for a given concentration of a fluorophore that is valid from instrument to instrument is difficult, if not impossible.

Fluorescence Polarization

Light is composed of electrical and magnetic waves at right angles to each other. Light waves produced by standard excitation sources have their electrical vectors oriented randomly. Light waves passed through certain crystalline materials (polarizers) have their electrical vectors oriented in a single plane and are said to be plane-polarized. Fluorophores absorb light most efficiently in the plane of their electronic energy levels. If their rotational relaxation (Brownian movement) is slower than their fluorescence decay time—as is the case for large fluorescent-labeled molecules—the emitted fluorescence light will remain polarized. Because small molecules have rotational relaxation times that are much shorter than their fluorescence decay time, their emitted fluorescence light is depolarized. However, if the small fluorescent molecule is attached to a macromolecule, or if it is placed in a viscous solution, the small molecule will emit polarized light. Fluorescence polarization (P) is defined by the following equation:

$$P = \frac{I_v - I_h}{I_v + I_h} \qquad \textbf{(9.6)}$$

where I_v = intensity of emitted fluorescence light in the vertical plane; and I_h = intensity of emitted fluorescence light in the horizontal plane.

As indicated, P is the difference between the two observed intensities divided by their sum. Fluorescence polarization is measured by placing a mechanically or electrically driven polarizer between the sample cuvet and the detector. In the normal instrumentation mode, the sample is excited with polarized light to obtain maximum sensitivity. The polarization analyzer is positioned to measure the intensity of emitted

fluorescence light in the vertical plane (I_v), and then the polarization analyzer is rotated 90 degrees to measure emitted fluorescence light intensity in the horizontal plane (I_h). P is then calculated manually or automatically by using Eq. 9.6.

Fluorescence polarization is used to quantify analytes by using the change in fluorescence depolarization following immunologic reactions (see Chapter 15). Quantification is accomplished by adding a known quantity of fluorescent-labeled analyte molecules to a reaction solution containing an antibody specific to the analyte. The labeled analyte binds to the antibody, and the slowed rotation and longer relaxation time results in an increased degree of fluorescence polarization. The addition of a nonlabeled analyte, such as an unknown quantity of a therapeutic drug in a serum specimen, will result in competition for binding to the antibody with the fluorescent-labeled analyte. This change in binding of the fluorophore-labeled analyte causes a change in fluorescence polarization that is inversely proportional to the amount of analyte contained in a given sample. Because the change in fluorescence polarization is a direct response to the reaction mixture, the bound fluorophore need not be separated from free fluorophore. Thus fluorescence polarization is applicable to homogeneous assays of low-molecular-weight analytes, such as therapeutic drugs.

Instrumentation

Fluorometers and spectrofluorometers are used to measure fluorescence. Operationally, a fluorometer uses interference filters or glass filters to produce monochromatic light for sample excitation and for isolation of fluorescence emission, whereas a spectrofluorometer uses a grating or prism monochromator.

Components

Basic components of fluorometers and spectrofluorometers include (1) an excitation source, (2) an excitation monochromator, (3) a cuvet, (4) an emission monochromator, and (5) a detector. In Fig. 9.8, these components are shown as they would be configured in a 90-degree optical system.

With fluorometers and spectrofluorometers, the placement of the cuvet and the excitation beam relative to the photodetector is critical in establishing the optical geometry for fluorescence measurements. As fluorescence light is emitted in all directions from a molecule, several excitation/emission geometries are used to measure fluorescence (Fig. 9.9). Most commercial spectrofluorometers and fluorometers use the right angle detector approach because it minimizes the background signal that limits analytical detection. The end-on approach allows the adaptation of a fluorescence detector to existing 180 degrees absorption instruments. Its limit of detection is restricted by (1) the quality of the excitation and/or emission interference filter pair, (2) the excitation and/or emission spectral band overlap, and (3) the inner filter effect (which is discussed later). The front surface approach provides the greatest linearity over a broad range of concentration because it minimizes the inner filter effect. The front surface approach has a comparable limit of detection to right angle detectors but is more susceptible to background light scatter. Front surface fluorometry has been widely applied to heterogeneous solid phase fluorescence immunoassay systems.

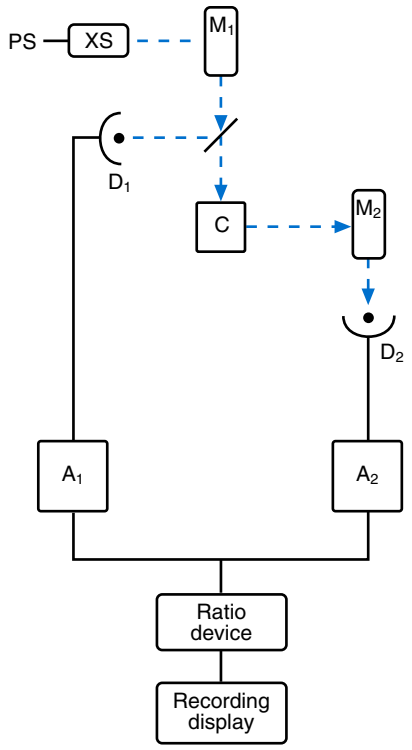


Fig. 9.8 Block diagram of a typical spectrofluorometer. A_1 and A_2, Excitation signal and emission signal amplifiers, respectively; *C*, sample cell; D_1 and D_2, detectors (D_1 monitors the variation in excitation intensity and D_2 measures fluorescence emission intensity); M_1, excitation monochromator; M_2, emission monochromator; *PS*, power supply; *XS*, xenon source.

To accommodate these different geometries, the sample cell is oriented at different angles in relation to the excitation source and the detector. Major concerns related to the geometry of the sample cell are (1) light scattering, (2) the inner filter effect, and (3) the sample volume element seen by the detector. Fig. 9.10 shows the sample cell and slit arrangement for a conventional fluorescence spectrophotometer with excitation and emission slits oriented at a right angle. $Slit_1$ and $Slit_2$ designate the excitation and emission slits, respectively. The position of the emission slit and the width of the slit are important. If the emission slit is located near the front edge of the sample cell, as shown in Fig. 9.10B, the inner filter effect is minimized. If the emission slit width is increased, the detector will be more sensitive, but specificity may decrease.

Performance Verification

As with spectrophotometers, NIST provides a number of SRMs for use in calibration or verification of the performance of fluorometers or fluorospectrophotometers. These include SRM 936a (quinine sulfate dihydrate) for calibration and SRM 1932 (fluorescein) for establishing a reference scale for fluorescence measurements (see http://www.nist.gov/srm).

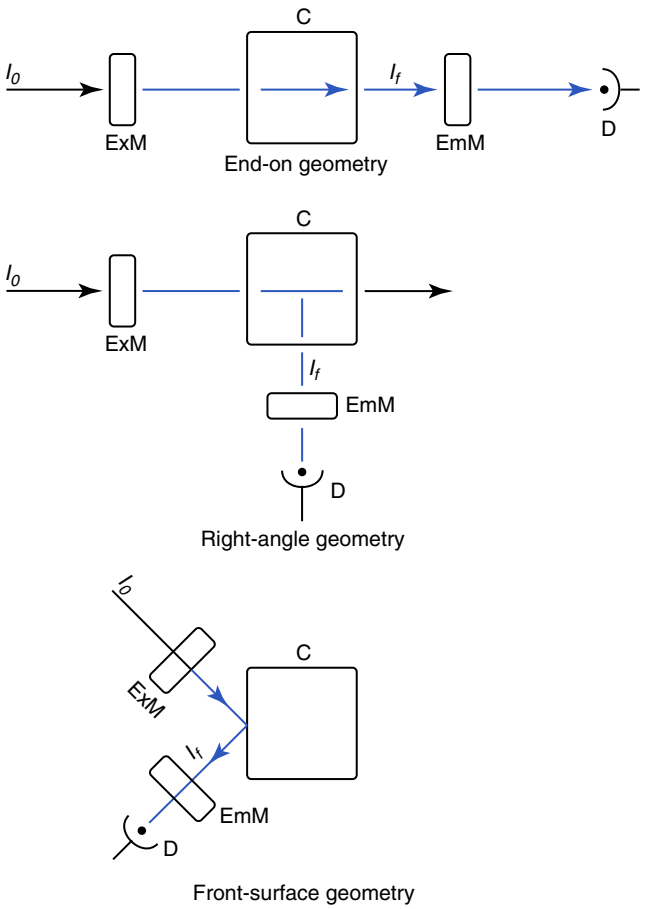


Fig. 9.9 Fluorescence excitation/emission geometries. *C,* Sample cuvet; *D,* detector; *EmM,* emission monochromator; *ExM,* excitation monochromator; I_0, initial excitation energy; I_f, fluorescence intensity.

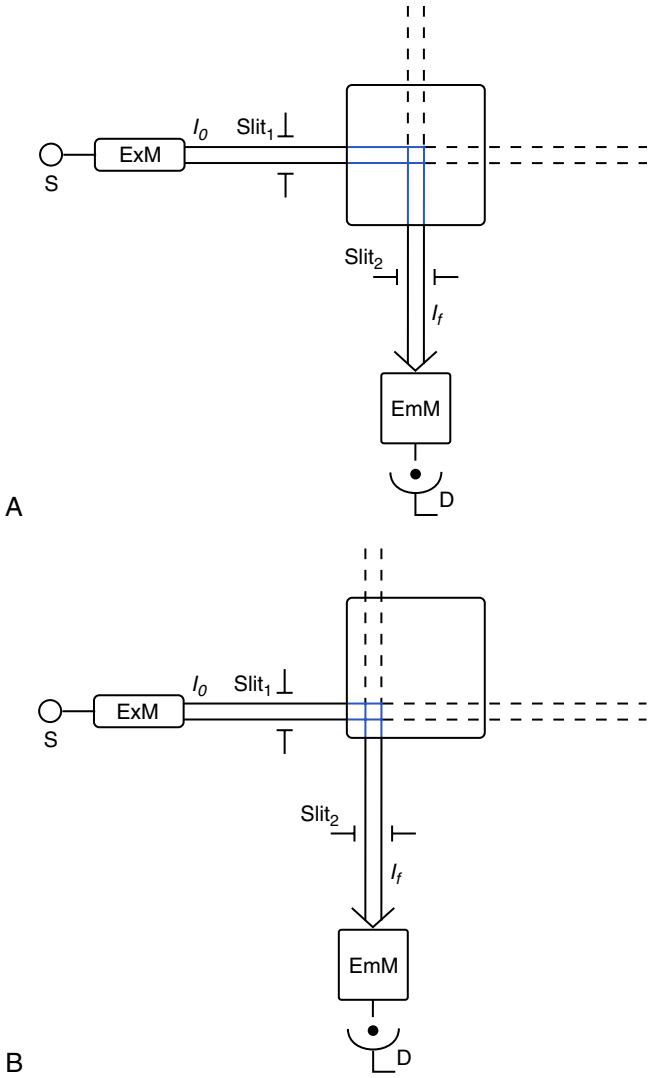


Fig. 9.10 Two right-angle fluorescence sample cuvet positions. (A) The standard 90 degrees configuration. (B) The offset positioning of the cuvet to minimize the inner filter effect. *D,* Detector; *EmM,* emission monochromator; *ExM,* excitation monochromator; I_0, initial excitation energy; I_f, fluorescence intensity; *S,* source.

Types of Fluorometers and Spectrofluorometers

Fluorometers and fluorescence spectrophotometers are available that offer a variety of features. These features include (1) ratio referencing, (2) computer-controlled excitation and emission monochromators, (3) pulsed xenon light sources, (4) photon counting, (5) a rhodamine cell for corrected spectra, (6) polarizers, (7) flow cells, (8) front surface viewing adapters, (9) multiple cell holders, and (10) computer-based data reduction systems.

In addition to the basic spectrofluorometer discussed earlier (see Fig. 9.8), other types of fluorometric instruments include (1) a ratio-referencing spectrofluorometer, (2) a time-resolved fluorometer, (3) a flow cytometer, and (4) a hematofluorometer.

Ratio-referencing spectrofluorometer. A typical ratio-referencing spectrofluorometer is illustrated in Fig. 9.11. Basically, this is a simple right-angle instrument that uses two monochromators (M1 and M2), two PMT detectors (D1 and D2, the reference and sample PMTs), and a xenon lamp source. The light from the exciter monochromator (M1) is split, and a small portion (10%) is directed to the reference PMT (D1) for ratio-referencing purposes. The remaining excitation light is focused into the sample cuvet (C). Emission optics are positioned at a right angle to the excitation optics. An emission monochromator (M2) is used to select or scan the desired portion of the emission spectra, which is directed to the sample PMT (D2) for measurement of emission intensity. The output signals from the reference and from the sample PMTs are amplified (A1 and A2), and a ratio of the sample to the reference signal is provided by a digital display or a chart recorder. The operational mode of a ratio fluorometer is similar to that of a spectrofluorometer; however, only discrete excitation and emission wavelengths are available, and the use of this type of instrument is precluded from scanning fluorophores to obtain emission and excitation spectra. The ratio filter fluorometer is most useful for obtaining concentration measurements at defined excitation and emission wavelengths.

The ratio-referencing spectrofluorometer is operated at fixed excitation and emission wavelength settings for

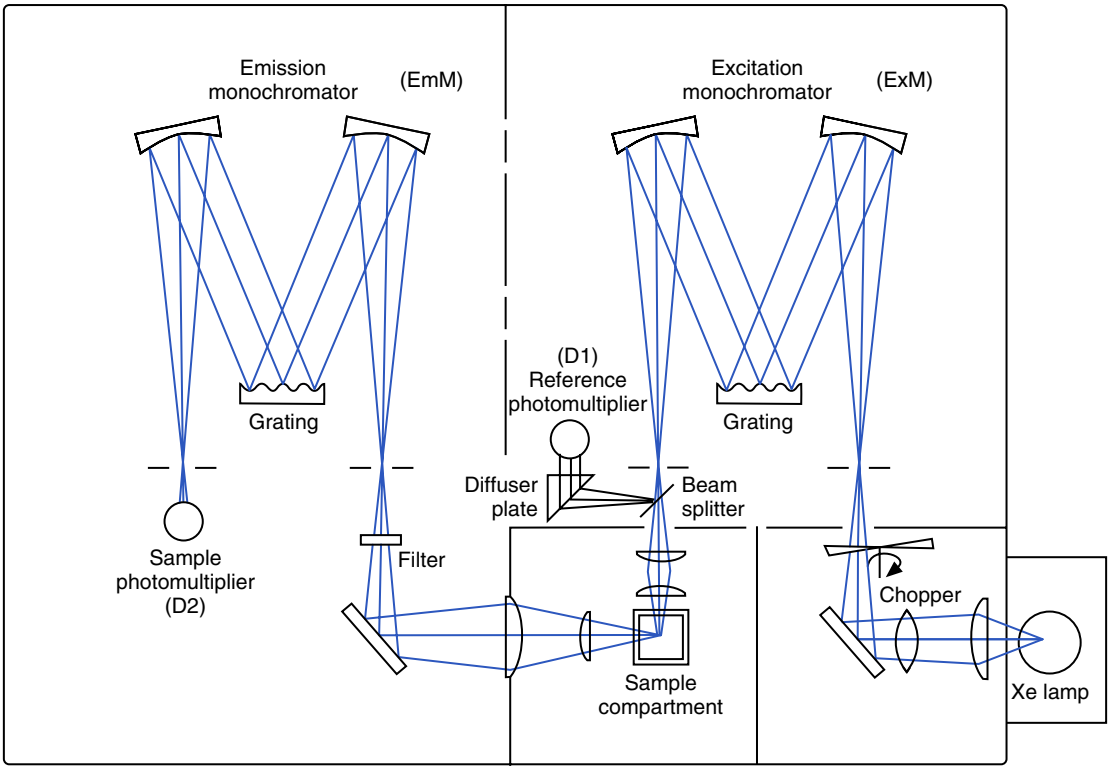


Fig. 9.11 Diagram of a typical ratio-referencing spectrofluorometer.

concentration measurements, or it is used to measure the excitation or emission spectrum of a given compound. Measurement of the concentration of unknowns is accomplished in a similar manner as with a single-beam fluorometer. A blank solution and a calibrating solution are first measured, and then the unknown samples are measured. The ratio-referencing spectrofluorometer in Fig. 9.11 provides two advantages over single-beam spectrofluorometers. First, it eliminates short- and long-term xenon lamp energy fluctuations (i.e., arc flicker and lamp decay), thus minimizing the need for frequent calibration of the instrument during analysis. Second, it provides excitation spectra, which are corrected for wavelength-dependent energy fluctuations.

Time-resolved fluorometers. A time-resolved fluorometer is similar to the ratio-referencing fluorometer, with the exception that the light source is pulsed and the detector monitors, in a fast photon-counting mode, the exponential decay of the fluorescence signal after excitation. Time-resolved fluorometry requires the use of long-lived fluorophores, such as the lanthanide (rare earth) metal ions europium (Eu^{3+}) and samarium (Sm^{3+}). Whereas most fluorescence compounds have decay times of 5 to 100 ns, europium chelates decay in 0.6 to 100 s. Thus time-resolved fluorescence assays take advantage of the difference in the lifetimes of fluorophore and background fluorescence by measuring the decaying fluorescence signal. This eliminates background interferences and at the same time averages the signal to improve the precision of measurement. Detection limits of approximately 10^{-13} mol/L have been achieved with time-resolved fluorometry, for an improvement of about four orders of magnitude compared with conventional fluorometric measurements. For example, Eu^{3+}-labeled nanoparticles in combination with time-resolved fluorometry have been used to develop a highly sensitive immunoassay for free and a total prostate-specific antigen with a functional sensitivity of 0.5 ng/L.

Flow cytometry. *Cytometry* refers to the measurement of physical and/or chemical characteristics of cells or particles. **Flow cytometry** is a process in which such measurements are made while cells or particles pass, preferably in single file, through the measuring apparatus in a fluid stream. Flow sorting extends flow cytometry by using electrical or mechanical means to divert and collect cells with one or more measured characteristics falling within a range of values set by the user.

Operationally, flow cytometry combines laser-induced fluorometry and particle light-scattering analysis that allow different populations of (1) molecules, (2) cells, or (3) particles to be differentiated by size and shape using low-light and right-angle light scattering. The use of a laser is ideally suited for low-angle light scattering. These cells, molecules, or particles are labeled with different specific fluorescent labels, such as (1) β-phycoerythrin, (2) fluorescein isothiocyanate, (3) rhodamine-6G, and (4) dye-labeled antibodies. As the particles flow through the flow cell, simultaneous fluorescence and light-scattering measurements are automatically performed by the flow cytometer. Most flow cytometers incorporate two or more fluorescence emission detection systems so that multiple fluorescent labels are able to be used. In this manner, molecules, cells, and particles are classified by (1) size, (2) shape, and (3) type according to their light-scattering and fluorescent properties. A schematic diagram of a flow cytometer is shown in Fig. 9.12.

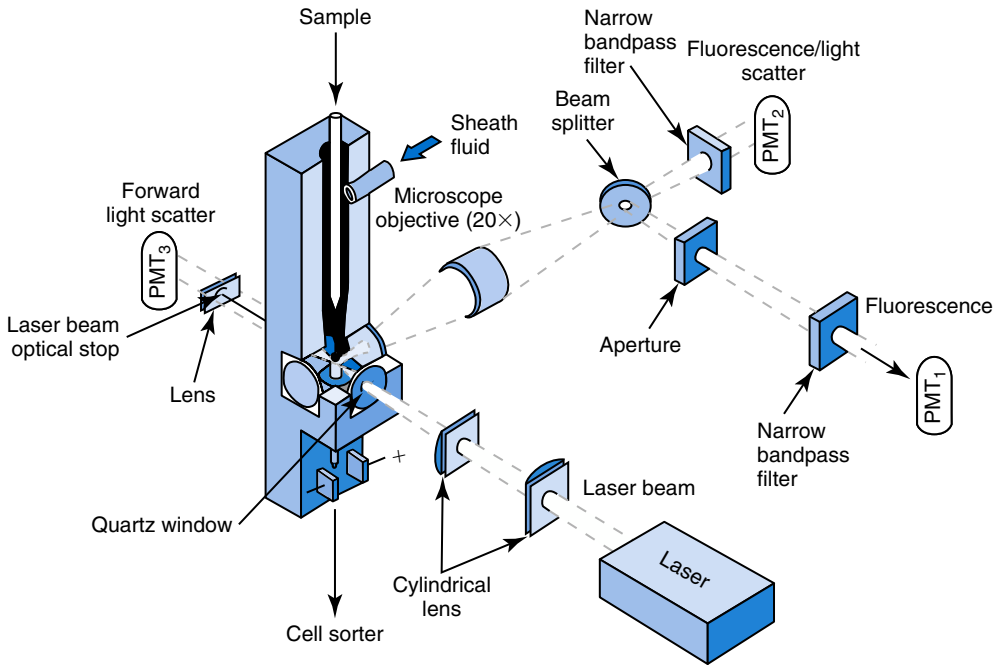


Fig. 9.12 Schematic diagram of a flow cytometer. *PMT,* Photomultiplier tube.

Flow cytometers are able to measure multiple parameters, including (1) cell size (forward scatter), (2) granularity (90-degree scatter), (3) DNA and RNA content, (4) DNA (AT)/(GC) nucleotide ratios, (5) chromatin structure, (6) antigens, (7) total protein content, (8) cell receptors, (9) membrane potential, and (10) calcium ion concentration as a function of pH. In particle-based flow cytometry, a flow cytometer is combined with microspheres that are used as the solid support for conventional immunoassay, affinity assay, or DNA hybridization assays. The microspheres contain different proportions of two fluorescent dyes to generate 100 or more different signatures. Each of these signatures can be assigned to a specific analyte to multiplex many analytes in the same sample. The resulting system is very flexible for small volume assays that simultaneously assess a wide variety of analytes in the same sample.

Hematofluorometer. The *hematofluorometer* is a single-channel front surface photofluorometer dedicated to the analysis of zinc protoporphyrin in whole blood (see Chapter 29). A typical hematofluorometer uses a quartz-tungsten lamp, a narrow bandpass excitation filter (420 nm), front surface optics, a narrow bandpass emission filter (594 nm), and a PMT. A drop of whole blood is placed on a small rectangular glass slide that serves as a cuvet.

Limitations of Fluorescence Measurements

Factors that influence fluorescence measurements include (1) concentration effects (e.g., inner filter effect, concentration quenching), (2) background effects (due to Rayleigh and Raman scattering), (3) solvent effects (e.g., interfering nonspecific fluorescence, quenching from the solvent), (4) sample effects (e.g., light scattering, interfering fluorescence, sample adsorption), (5) temperature effects, and (6) photodecomposition (bleaching) of the sample.

Inner Filter Effect

The linear relationship between concentration and fluorescence emission (Eq. 9.5) is valid when solutions absorb less than 2% of the exciting light. As the absorbance of the solution increases above this amount, the relationship becomes nonlinear—a phenomenon known as the *inner filter effect*. It is caused by loss of excitation intensity across the cuvet path length as excitation light is absorbed by the fluorophore. Thus, as the fluorophore becomes more concentrated, absorbance of excitation intensity increases, and loss of excitation light as it travels through the cuvet increases. This effect is most often encountered with a right-angle fluorescence instrument, in which the emission slits are set to monitor the center of the sample cell, where absorbance of excitation light is greater than at the front surface of the cuvet. Therefore it is less problematic if a front surface fluorescence instrument is used. However, most fluorescence measurements are made on very dilute solutions, and the inner filter effect therefore is not a problem.

Concentration Quenching

Another related phenomenon that results in a lower quantum yield than expected is called *concentration quenching.* This occurs when a macromolecule, such as an antibody, is heavily labeled with a fluorophore, such as fluorescein isothiocyanate. When this compound is excited, the fluorescence labels are in such close proximity that a radiation-less energy transfer occurs. Thus the resulting fluorescence is much lower than expected for the concentration of the label. This is a common problem in flow cytometry and in laser-induced fluorescence when attempts are made to enhance detection sensitivity by increasing the density of the fluorescing label.

Light Scattering

Light scattering—Rayleigh and Raman—limits the use of fluorescence measurements. Rayleigh scattering occurs with no change in wavelength. For fluorophores with small Stokes shifts, the excitation and emission spectra overlap and are particularly susceptible to loss of detection caused by background light scatter. Rayleigh scatter is controlled by the use of well-defined emission and excitation interference filters or by appropriate monochromator settings and the use of polarizers.

Raman scattering occurs with the lengthening of the wavelength. This type of light scattering is independent of excitation wavelength and is a property of the solvent. Because Raman light scattering appears at longer wavelengths than the exciting radiation, it is a difficult interference to eliminate when working at very low fluorophore concentrations.

Cuvet Material and Solvent Effects

Certain quartz glass and plastic materials that contain UV absorbers will fluoresce. Some solvents, such as ethanol, are also known to cause appreciable fluorescence. It is therefore important when developing a fluorescence assay to assess the background fluorescence of all components of the reaction mixture.

Sample Matrix Effects

A serum or urine sample contains many compounds that fluoresce. Thus the sample matrix is a potential source of unwanted background fluorescence and must be examined when new methods are developed. The most serious contributors to unwanted fluorescence are proteins and bilirubin. However, because protein excitation maxima are in the spectral region of 260 to 290 nm, their contribution to overall background fluorescence is minor when excitation occurs above 300 nm.

Light scattering of proteins and other macromolecules in the sample matrix has been known to cause unwanted background fluorescence. Lipemic samples, for example, are noted for their intense light scattering, and the relative contribution of lipids to the background signal of a fluorescence measurement should be investigated when a new method is developed.

In addition to background interferences, dilute solutions of some fluorophores in the concentration range of 10^{-9} mol/L and below will absorb to the walls of glass cuvets and other reaction vessels. Also, dilute solutions of fluorophores, when excited over long periods, are susceptible to photodecomposition by intense excitation light. Operationally, these problems are prevented by (1) selecting proper reaction vessels, (2) adding wetting agents, and (3) minimizing the length of time a sample is exposed to the excitation light.

Temperature Effects

The fluorescence quantum efficiency of many compounds is sensitive to temperature fluctuations. Therefore the temperature of the reaction must be tightly controlled. In general, fluorescence intensity decreases with increasing temperature by approximately 1% to 5% per degree Celsius. Increased temperatures result in more frequent molecular collisions and quenching. Collisional quenching can be decreased by lowering the temperature or by increasing the viscosity.

Photodecomposition

In conventional fluorometry, excitation of weakly fluorescing or dilute solutions with intense light sources will cause photochemical decomposition of the analyte (photobleaching).

The following steps help to minimize photodecomposition effects:

1. Always use the longest feasible wavelength for excitation that does not introduce light-scattering effects.
2. Decrease the duration of excitation of the sample.
3. Protect unstable solutions from ambient light by storing them in dark bottles.
4. Remove dissolved oxygen from the solution.

Highly intense laser light sources are useful for applications that require higher sensitivity to measure low analyte concentrations (e.g., flow cytometry, fluorescence microscopy). Laser light sources can have energy outputs of 5 to 10 mW, which enable high sensitivity detection; however, these intense light sources rapidly photodecompose some fluorescent analytes. This photodecomposition introduces nonlinear response curves and loss of most of the sample fluorescence. Thus, the optimization of laser intensity balances higher sensitivity with increased photodecomposition.

LUMINOMETRY

Chemiluminescence and **electrochemiluminescence** are types of luminescence in which the excitation event is caused by a chemical, biochemical, or electrochemical reaction, and not by photo illumination. Instruments for measuring this type of light emission are known generically as luminometers.

Basic Concepts

The physical event of light emission in chemiluminescence and electrochemiluminescence is similar to fluorescence in that it occurs from an excited singlet state, and the light is emitted when the electron returns to the ground state.

Chemiluminescence

Chemiluminescence is the emission of light when a molecule returns from an excited or higher energy level to a lower energy level. The excitation event is caused by a chemical reaction with compounds such as luminol and acridinium esters. Light is emitted from the excited product formed in the oxidation reaction. These reactions occur in the presence of catalysts such as (1) enzymes (e.g., alkaline phosphatase, horseradish peroxidase, microperoxidase) and (2) metal ions or metal complexes (e.g., Cu^{2+} and Fe^{3+} phthalocyanine complexes).

Chemiluminescence assays are ultrasensitive (attomole-to-zeptomole detection limits) and have wide dynamic ranges. They are now widely used in automated immunoassay and DNA probe assay systems, such as (1) acridinium ester and acridinium sulfonamide labels, (2) 1,2-dioxetane substrates

for alkaline phosphatase labels, and (3) the enhanced luminol reaction for horseradish peroxidase labels.

Electrochemiluminescence

Electrochemiluminescence differs from chemiluminescence in that the reactive species that produce the chemiluminescent reaction are electrochemically generated from stable precursors at the surface of an electrode. A ruthenium (Ru^{2+}), tris(bipyridyl) chelate is the most commonly used electrochemiluminescence label, and electrochemiluminescence is generated at an electrode via an oxidation reduction type of reaction with tripropylamine. This chelate is very stable and relatively small, and has been used to label haptens or large molecules (e.g., proteins, oligonucleotides). The electrochemiluminescence process has been used in both immunoassays and nucleic acid assays. Advantages of this process include (1) improved reagent stability, (2) simple reagent preparation, and (3) enhanced sensitivity. With its use, detection limits of 200 fmol/L and a dynamic range extending over six orders of magnitude have been obtained.

Instrumentation

The basic components of a luminometer include (1) the sample cell housed in a light-tight chamber, (2) the injection system for adding reagents to the sample cell, and (3) the detector. The detector is usually a PMT. However, a charged coupled detector (a multichannel device), an x-ray film, or photographic film has been used to image chemiluminescence reactions on a membrane or in the wells of a microplate. For electrochemiluminescence, the reaction vessel incorporates an electrode at which the electrochemiluminescence is generated.

Limitations of Chemiluminescence and Electrochemiluminescence Measurements

Factors that commonly degrade the analytical performance of luminescence measurements include (1) light leaks, (2) light piping, (3) and high background luminescence from assay reagents and reaction vessels (e.g., plastic tubes exposed to light). The ultrasensitive nature of chemiluminescence assays requires stringent controls on the purity of reagents and the solvents (e.g., water) used to prepare reagent solutions. Efficient capture of the light emission from reactions that produce a flash of light requires an efficient injector that provides adequate mixing when the triggering reagent is added to the reaction vessel. Chemiluminescence, and electrochemiluminescence assays have a wide linear range, usually including several orders of magnitude, but very high-intensity light emission has led to pulse pileup in PMT tubes, and this may lead to a serious underestimation of the true light emission intensity.

POINTS TO REMEMBER

Radiant Energy in Analytical Measurements

- Fluorescence: light is absorbed by a molecule and subsequently emitted as light at a longer wavelength
- Chemiluminescence: light is emitted as a result of chemical energy; no light is absorbed
- Electrochemiluminescence: light is generated from precursor molecules at the surface of an electrode

NEPHELOMETRY AND TURBIDIMETRY

Light scattering is a physical phenomenon that results from the interaction of light with particles in solution. **Nephelometry** and **turbidimetry** are analytical techniques that are used to measure scattered light. Light-scattering measurements have been applied to immunoassays of specific proteins and haptens.

Basic Concepts

Light scattering occurs when radiant energy passing through a solution encounters a molecule in an elastic collision, which results in scattering of light in all directions. Unlike fluorescence emission, the wavelength of the scattered light is the same as that of the incident light. The principal factors that influence light scattering include the size, concentration and molecular weight of the particles, wavelength, and distance of observation.

Light scattering from small particles is directly proportional to their concentration and molecular weight. Light scatter is inversely proportional to the forth power of wavelength (blue light scatters more than red light) and inversely proportional to the square of the distance from the light scatter source to the detector. Thus the detector should be located close to the analytical cell by combining the cell and the detector or by using good collection optics.

As the particles become larger than the incident light wave, the radiated light waves are no longer all in phase. Reinforcement of radiation occurs in some directions, and destructive interference occurs in others. The scattering patterns from these large particles are characteristic of the size and shape of the particle.

Measurement of Scattered Light

Turbidimetry and nephelometry are methods used to measure scattered light. These techniques have proved useful for the quantification of serum proteins.

Turbidimetry

Because of light scattering, the intensity of light reaching the detector at 180 degrees is reduced. The measurement of this decrease in intensity is called *turbidimetry.* Analogous to absorption spectroscopy, turbidity is defined as:

$$I = I_0 e^{-bt} \tag{9.7}$$

$$\text{or } t = \left(\frac{1}{b}\right) \ln\left(\frac{I_0}{I}\right) \tag{9.8}$$

where t = turbidity; b = path length of incident light through the solution of light-scattering particles; I = intensity of transmitted light; and I_0 = intensity of incident light.

A turbidimeter is used to measure the intensity of light scattering. Photometers or spectrophotometers are often used as turbidimeters, as turbidimetric measurements are easily performed with them and require little optimization. The principal concern of turbidimetric measurements is signal-to-noise ratio. Photometric systems with electro-optical noise in the range of ±0.0002 absorbance unit or less are useful for turbidity measurements.

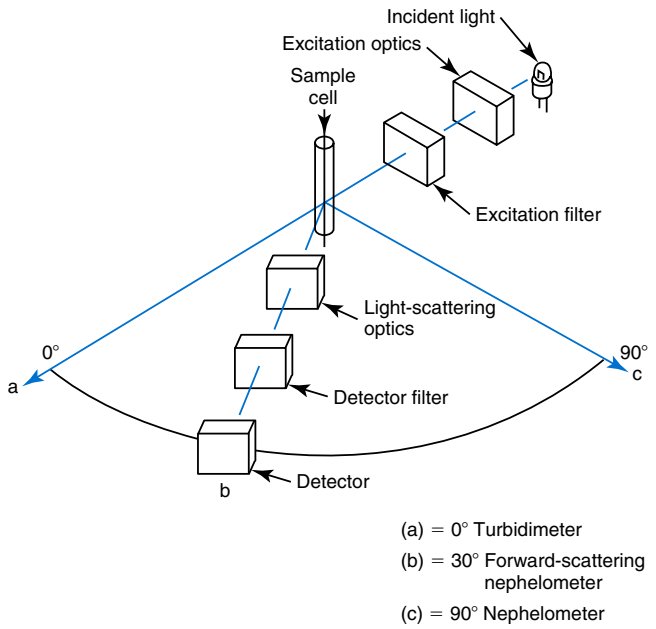


Fig. 9.13 Schematic diagram of light-scattering instrumentation shows (a) the optics position for a turbidimeter, (b) the optics position for a forward-scattering nephelometer, and (c) the optics position for a right-angle nephelometer.

Nephelometry

Nephelometry is defined as the detection of light energy scattered or reflected toward a detector that is not in the direct path of the transmitted light. Common nephelometers measure scattered light at right angles to the incident light. Some nephelometers are designed to measure scattered light at an angle other than 90 degrees to take advantage of the increased forward-scatter intensity caused by light scattering from larger particles (e.g., immune complexes).

Fluorometers are often used to perform nephelometric measurements. However, the angular dependence of light-scattering intensity has resulted in the design of special nephelometers. These devices place the PMT detector at appropriate angles to the excitation light beam. The design principle of a nephelometer is similar to the design principle applied in fluorescence measurements. The major operational difference between the fluorometer and the nephelometer is that the excitation and detection wavelengths of a nephelometer will be set to the same value. The principal parameters of light scatter instrumentation are (1) excitation intensity, (2) wavelength, (3) distance of the detector from the sample cuvet, and (4) minimization of external stray light. As shown in Fig. 9.13, the basic components of a nephelometer include (1) a light source, (2) collimating optics, (3) a sample cell, and (4) collection optics, which include light-scattering optics, a detector optical filter, and a detector. The schematic diagram also shows the different angles from the incident light beam where the detector, filter, and optics are placed to measure light scattering. Fig. 9.13(a) shows the straight-through arrangement for turbidimetry, whereas Fig. 9.13(b and c) shows arrangements frequently found in nephelometers. The detector arrangement shown in Fig. 9.13(b) is used for measurement of forward scatter at 30 degrees, which is the optical arrangement used in some commercial nephelometers.

Operationally, the optical components used in turbidimeters and nephelometers are similar to those used in fluorometers or photometers. For example, the light sources commonly used are (1) quartz-halogen lamps, (2) xenon lamps, and (3) lasers. Helium-neon (He-Ne) lasers, which operate at 633 nm, have typically been used for light-scattering applications, such as (1) nephelometry (2) immunoassays, (3) particle size determination, and (4) particle shape determination. The laser beam is used specifically in some nephelometers because of its high intensity; in addition, the coherent nature of laser light makes it ideally suited for nephelometric applications. Ratio-referencing fluorometers are also well suited for nephelometric measurements.

Limitations of Light-Scattering Measurements

Antigen excess and matrix effects are limitations encountered in the use of turbidimeters and nephelometers for measurement of analytes of clinical interest.

Antigen Excess

As turbidity increases during the addition of antigen to antibodies, the signal increases to a maximum value and then decreases. The point at which the decrease begins marks the beginning of the phase of antigen excess; this phenomenon is explained in Chapter 15. Consequently, light-scattering methods for quantification of antigen-antibody reactions must provide a method for detecting antigen excess. The kinetics of immune complex formation measured by nephelometry or turbidimetry is sufficiently different in each of the three phases—(1) antibody excess, (2) equivalence, and (3) antigen excess—that computer algorithms can automatically flag antigen excess.

Matrix Effects

Particles, solvents, and all serum macromolecules scatter light. Lipoproteins and chylomicrons in lipemic serum provide the highest background turbidity or nephelometric intensity. When measuring an antigen by light scattering, lipemia within the sample will cause background interference since the lipid molecules will also scatter light. Lipemic samples may continue to have interference upon dilution since the antigen concentration is also decreased. Instead of dilution, rate measurements are a more effective method for minimizing lipemic interference.

POINTS TO REMEMBER

Light Scattering Assays

- Turbidity assays measure light scattering to determine the formation of antigen-antibody complexes
- Interferences in light scattering include:
  - An excess of antigen decreases the turbidity
  - Macromolecules, such as lipoproteins, chylomicrons, and dust in the sample matrix, will increase the background turbidity
- Macromolecule interference in light scattering can be decreased by using rate measurements rather than directly assessing light scattering

REVIEW QUESTIONS

1. What wavelength of light in the electromagnetic spectrum has the highest energy?
   a. 400 nm
   b. 500 nm
   c. 550 nm
   d. 660 nm
2. The emission of light caused by a chemical reaction is:
   a. chemiluminescence.
   b. nephelometry.
   c. turbidimetry.
   d. fluorescence.
3. The expression of the relationship between the concentration of a substance in solution and absorbance of light by that substance is referred to as Beer's law. This relationship is expressed in a formula as:
   a. log (1/T).
   b. C = abc.
   c. $I_0/I_s \times 100$.
   d. A = abc.
4. Calculate the concentration of NADH, given that the absorbance at 340 nm is 0.5 and the molar absorptivity constant is 6220 M^{-1} cm^{-1}. The cuvet has a path length of 1 cm.
   a. 1.0 μM
   b. 10.0 μM
   c. 0.8 μM
   d. 8.0 μM
   e. 0.05 μM
5. Which component of a single-beam spectrophotometer isolates radiant energy of a specific wavelength and excludes that of other wavelengths?
   a. Monochromator
   b. Entrance slit
   c. Cuvet
   d. Light source
6. Measurement of the concentration of a substance that is performed by measuring scattered light at right angles to the incident light beam is called:
   a. fluorometry.
   b. atomic absorption.
   c. turbidimetry.
   d. nephelometry.
7. A solution that is identical to that of calibrating or unknown solutions and is used to set 100% T (zero absorbance) at the beginning of a photometric analysis is referred to as a:
   a. standard solution.
   b. calibrating solution.
   c. reference blank.
   d. reagent blank.
8. Calculate the transmittance of light when the absorbance in a sample is 0.5.
   a. 25.6%
   b. 40.1%
   c. 31.6%
   d. 10.9%
9. In a flameless atomic absorption technique, the sample holder is typically a:
   a. carbon rod in an enclosed chamber.
   b. flame.
   c. borosilicate glass cuvet.
   d. reflective surface.
10. The bandwidth of the spectral absorbance curve of an absorbing substance at a point equal to one-half of the maximum absorbance is referred to as the:
    a. standard curve.
    b. bandpass filter.
    c. natural bandwidth.
    d. monochromatic bandwidth.

SUGGESTED READINGS

Blackburn, G. F., Shah, H. P., Kenten, J. H., et al. (1991). Electrochemiluminescence: development of immunoassays and DNA probe assays for clinical diagnostics. *Clinical Chemistry*, *37*, 1534–1539.

Diamandis, E., & Christopoulos, T. K. (1990). Europium chelate labels in time-resolved fluorescence immunoassays and DNA hybridization assays. *Analytical Chemistry*, *62*, 1149A–1157A.

Gore, M. G. (2000). *Spectrophotometry and spectrofluorimetry: a practical approach* (2nd ed.). London: Oxford University Press, 1–368.

Heiftje, G. M., & Vogelstein, E. E. (1981). A linear response theory approach to time-resolved fluorometry. In E. L. Wehry (Ed.), *Modern fluorescence spectroscopy Vol. 4.* (pp. 25–50). New York: Plenum Press.

Jolley, M. E., Stroupe, S. D., Schwenzer, K. S., et al. Fluorescence polarization immunoassay. III. An automated system for therapeutic drug determination. Clinical Chemistry, 27, 1575–1579.

Kricka, L. J., & Park, J. Y. (2018). Optical techniques. In N. Rifai, A. R. Horvath, & C. T. Wittwer (Eds.), *Tietz textbook of clinical chemistry and molecular diagnostics* (6th ed) (pp. 200–223). St. Louis, Missouri: Elsevier.

Patrick, C. W. (2002). Clinical flow cytometry: milestones along the pathway of progress. *MLO: Medical Laboratory Observer*, *34*, 10–16.

Slavin, W. (1982). Atomic absorption spectroscopy: the present and future. *Analytical Chemistry*, *54*, 685A–694A.

Sternberg, J. A rate nephelometer for measuring specific proteins by immunoprecipitin reactions. Clinical Chemistry, 25:1456–1464.

Vignali, D. A. (2000). Multiplexed particle-based flow cytometric assays. *Journal of Immunological Methods*, *243*, 244–255.

10

Electrochemistry and Chemical Sensors

*Paul D'Orazio**

OBJECTIVES

1. Define the following: (a) amperometry, (b) biosensor, (c) conductometry, (d) coulometry, (e) electromotive force, (f) indicator electrode, (g) ion activity, (h) molality, (i) optode, (j) potentiometry, (k) reference electrode, and (l) voltammetry.
2. State the major advantages of electrochemical sensors for whole blood measurements.
3. Describe the components of a basic potentiometric cell and their functions, including electrodes, filling solutions, and voltmeter.
4. State the principle, a clinical application, and a potential interfering substance for each of the following electrochemical measurement techniques: (a) potentiometry, (b) voltammetry/amperometry, (c) conductometry, and (d) coulometry.
5. Describe the fundamental differences between direct potentiometry for measurement of electrolytes in serum, plasma, and whole blood and analytical methods which measure ion concentration.
6. Describe a PCO_2 electrode, including components and internal reactions.
7. Describe a PO_2 electrode, including components and internal reactions.
8. Define a biosensor; describe the two major categories of biosensors and a practical example of each type in the clinical laboratory.
9. Describe four transduction mechanisms used for enzyme-based biosensors and a practical example of each in the clinical laboratory.
10. State the application of the Nernst equation to electrochemical measurements.

KEY WORDS AND DEFINITIONS

Activity (of an ion) The concentration of free, unbound ion in solution.

Amperometry An electrolytic electrochemical process in which current is monitored at a fixed (controlled) voltage between working and reference electrodes in an electrochemical cell.

Biosensor A type of chemical sensor consisting of a biologic recognition element and a physicochemical transducer, often an electrochemical or an optical device.

Conductometry An electrochemical technique used to determine the quantity of an analyte present in a mixture by measuring its effect on the electrical conductivity of the mixture.

Coulometry An electrochemical technique that measures the electrical charge passing between two electrodes in an electrochemical cell, with the amount of charge passing between the electrodes being directly proportional to oxidation or reduction of an electroactive substance at one of the electrodes.

Electrochemical cell A device that consists of two electrodes (electron or metallic conductors) connected by an electrolyte solution that conducts ions (galvanic cell), or a device in which an external voltage is applied to a polarizable working electrode versus a reference electrode, with the resulting cathodic or anodic current of the cell being monitored (electrolytic cell).

Electrode A half-cell that consists of a single metallic conductor in contact with an electrolyte solution; the indicator (measuring) electrode is one half-cell and the reference electrode is the second half-cell.

Electrode potential The electromotive force (EMF) of a single half-cell measured with respect to the standard hydrogen electrode, set at zero by convention.

Ion-selective electrode An electrode that selectively interacts with a single ionic species; the potential produced at the membrane/sample solution interface is proportional to the logarithm of the ionic activity or the concentration of the ion in question.

* The author gratefully acknowledges the contributions of Drs. Richard A. Durst, Ole Siggaard-Andersen, and Mark E. Meyerhoff to earlier versions of this chapter.

Nernst equation The equation used to relate the potential of an electrochemical cell to the activity of a chemical species in solution.

Optode An optical sensor that measures specific substances such as pH, blood gases, and electrolytes, using dye immobilization, fluorescence quenching, or phosphorescence.

Potential difference The work required to move an electrical charge and measured in volts.

Potentiometry An electrochemical technique that measures an electrical potential difference between two electrodes (half-cells) in an electrochemical cell.

Redox couple A conjugate pair of substances that consists of any substance that accepts electrons (the oxidant) and any substance that donates electrons (the reductant); redox processes take place only between two redox couples, with electrons transferred from a reductant (Red_1) to an oxidant (Ox_2).

Voltammetry An electrolytic process in which a specific oxidation or reduction reaction occurs at the surface of the working electrode in response to an applied potential. The charge transfer at this interface (current flow) provides analytical information.

Electrochemical measurements are key analytical methods used especially in critical care clinical laboratories. In this chapter the fundamental electrochemical sensing principles of (1) potentiometry, (2) voltammetry/amperometry, (3) conductometry, and (4) coulometry will be summarized and clinical applications presented. Chemical sensors based on optical measurements and biosensors, both biocatalytic and affinity based, will be introduced and practical examples will be presented.

POTENTIOMETRY

Potentiometry is used clinically for measurement of pH, PCO_2, and electrolytes (Na^+, K^+, Cl^-, Ca^{2+}, Mg^{2+}, Li^+) in whole blood, serum, plasma, and urine and as the basis for some biosensors for metabolites of clinical interest.

Basic Concepts

Potentiometry is the measurement of an electrical **potential difference** between two **electrodes** (half-cells) in an **electrochemical cell** when the cell current is zero (Fig. 10.1). Such a cell consists of two electrodes, known as the indicator (measuring) and the reference electrodes, shown in Fig. 10.1 on the left and right, respectively. The two electrodes are connected by an electrolyte solution (sample). Each half-cell consists of an internal silver–silver chloride (Ag/AgCl) electrode, which is a specific type of potentiometric electrode known as a redox electrode. Each Ag/AgCl electrode is in contact with an internal electrolyte. The indicator electrode, also known as an **ion-selective electrode** (ISE), contains an inner electrolyte with a constant **activity** of the ion of interest in the form of a chloride salt. The reference electrode contains a high concentration of electrolyte, usually potassium chloride.

The electromotive force (E or EMF) is defined as the maximum difference in potential between the two electrodes when the cell current is zero. The cell potential is measured using a voltmeter, of which the common pH meter is a special type. To obtain an accurate potential measurement, it is necessary that no current flows through the cell. This is accomplished by incorporating a high input impedance of greater than 10^{12} Ω within the voltmeter.

The overall potential of a potentiometric cell is the sum of all potential gradients that exist between different phases of the cell. The potential of a single electrode with respect to the surrounding electrolyte and the absolute magnitude of the individual potential gradients between phases are unknown and cannot be measured. Only potential differences between two electrodes (half-cells) can be measured. Potential gradients can be classified as (1) redox potentials, (2) membrane potentials, or (3) diffusion potentials. In general, it is possible to devise a cell in such a manner that all potential gradients except one are constant. This potential then can be related to the activity of a specific ion of interest (e.g., hydrogen [H^+] or Na^+).

Types of Electrodes

Different types of electrodes are used for potentiometric applications. They include (1) redox, (2) ion-selective membrane (glass and polymer), and (3) PCO_2 gas-sensing electrodes.

Redox Electrodes

Redox potentials are the result of chemical equilibria involving electron transfer reactions:

$$\text{Oxidized form (Ox)} + ne^- \leftrightarrow \text{Reduced form (Red)} \quad (10.1)$$

where n = the number of electrons involved in the reaction. Any substance that accepts electrons is an oxidant (Ox), and

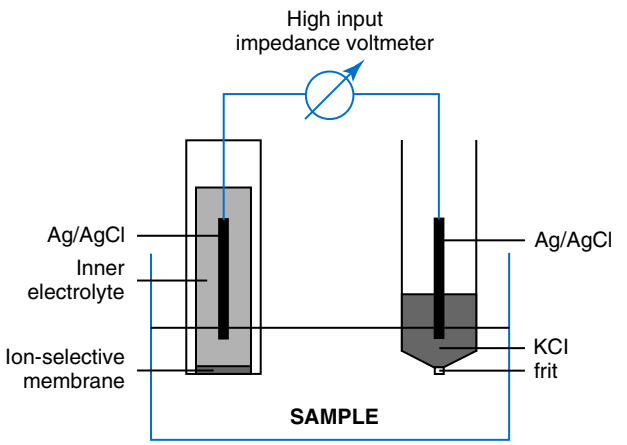


Fig. 10.1 Schematic of ion-selective membrane electrode–based potentiometric cell. *Ag/AgCl*, Silver–silver chloride; *KCl*, potassium chloride.

any substance that donates electrons is a reductant (Red). The two forms, Ox and Red, represent a **redox couple** (conjugate redox pair). Usually, homogeneous redox processes take place only between two redox couples. In such cases, electrons are transferred from Red_1 to an Ox_2. In this process, Red_1 is oxidized to its conjugate Ox_1, whereas Ox_2 is reduced to Red_2:

$$Red_1 + Ox_2 \leftrightarrow Ox_1 + Red_2 \quad (10.2)$$

The **electrode potential** for a redox couple is defined as the couple's potential measured with respect to the standard hydrogen (H_2) electrode, which by convention is set equal to zero. In such a cell, the standard H_2 electrode is the reference electrode and the given half-cell is the indicator electrode. The potential for the redox couple is shown by the **Nernst equation**:

$$E = E^0 - \frac{N}{n} \times \log \frac{a_{Red}}{a_{Ox}} = E^0 - \frac{0.0592\ V}{n} \times \log \frac{a_{Red}}{a_{Ox}} \quad (10.3)$$

where E = electrode potential of the cell; E^0 = standard electrode potential when $a_{Red}/a_{Ox} = 1$; n = number of electrons involved in the redox reaction; $N = (R \times T \times \ln 10)/F$ (the Nernst factor if $n = 1$); $N = 0.0592$ V if $T = 298.15$ K (25°C); $N = 0.0615$ V if $T = 310.15$ K (37°C); R = gas constant (8.31431 Joule × K^{-1} × mol^{-1}); T = absolute temperature (in kelvins); F = faraday constant (96,487 coulombs × mol^{-1}); ln 10 = natural logarithm of 10 = 2.303; a = activity; and a_{Red}/a_{Ox} = product of mass action for the redox reaction.

Redox electrodes currently in use include (1) inert metal electrodes immersed in solutions containing redox couples and (2) metal electrodes whose metal functions as a member of the redox couple.

Inert Metal Electrodes

Platinum and gold (Au) are examples of inert metals used to monitor the potential of a redox couple in a background electrolyte. The H_2 electrode is a special redox electrode for pH measurement. It consists of a Pt or Au electrode that is electrolytically coated (platinized) with highly porous Pt (Pt black) to catalyze the electrode reaction:

$$H^+ + e^- \leftrightarrow \frac{1}{2} H_2 \quad (10.4)$$

The electrode potential is given by:

$$E = E^0 - N \times \log \frac{(f_{H_2})^{1/2}}{f_{H_2} a_{H^+}} \quad (10.5)$$

or

$$E = E^0 - N \times \left[\log (f_{H_2})^{1/2} - \log a_{H^+}\right] \quad (10.6)$$

where $E^0 = 0$ at all temperatures (by convention); f_{H_2} = fugacity of H_2 gas (the "effective" partial pressure of a gas under ideal conditions); a_{H^+} = activity of H^+ ions; and $-\log a_{H^+}$ = negative log of H^+ activity (pa_{H^+} or pH).

When fugacity H_2 (f_{H_2}) in the solution is maintained constant by bubbling H_2, the potential is a linear function of $\log a_{H^+}$ ($= -$pH). In the standard H_2 electrode, the electrolyte consists of an aqueous solution of HCl with activity of HCl equal to 1.000 (concentration of HCl = 1.2 mol/L) in equilibrium with a gas phase, and with f_{H_2} equal to 1.000 (PH_2 = 101.3 kPa = 1 atm). The standard H_2 electrode is also used as a reference electrode.

Metal Electrodes Participating in Redox Reactions

The Ag/AgCl electrode is an example of a metal electrode that participates as a member of a redox couple. The Ag/AgCl electrode consists of an Ag wire or rod coated with $AgCl_{(solid)}$ in contact with a Cl^- solution of constant activity; this sets the half-cell potential. The Ag/AgCl electrode is itself considered a potentiometric electrode because its phase boundary potential is governed by an oxidation-reduction electron transfer equilibrium reaction that occurs at the surface of Ag:

$$AgCl_{(solid)} + e^- \leftrightarrow Ag^0_{(solid)} + Cl^- \quad (10.7)$$

The Nernst equation for the reference half-cell potential of an Ag/AgCl reference electrode is written as follows:

$$E_{Ag/AgCl} = E^0_{Ag/AgCl} + \frac{RT}{nF} \times \ln \frac{a_{AgCl}}{a_{Ag} a_{Cl^-}} \quad (10.8)$$

Because AgCl and Ag are both solids, their activities are equal to unity $\left(a_{AgCl} = a^0_{Ag} = 1\right)$. Therefore, from Eq. 10.8, the half-cell potential is controlled by the activity of the Cl^- ion in solution (a_{Cl^-}) contacting the electrode.

The Ag/AgCl electrode is used both as an internal reference element in potentiometric ISEs and as an external reference electrode half-cell of constant potential, which is required to complete a potentiometric cell (see Fig. 10.1). In both cases the Ag/AgCl electrode must be in equilibrium with a solution of constant Cl^- ion activity.

The Ag/AgCl element of the reference electrode half-cell is in contact with a high-concentration solution of a soluble Cl^- salt. Saturated KCl is commonly used. A porous membrane or frit is frequently used to separate the concentrated KCl from the sample solution. The frit serves both as a mechanical barrier to hold the concentrated electrolyte within the electrode and as a diffusional barrier to prevent proteins and other species in the sample from coming into contact with the internal Ag/AgCl element, which could poison and alter its potential. The interface between two dissimilar electrolytes (concentrated KCl/calibrator or sample) occurs within the frit and develops the liquid-liquid junction potential (E_j), a source of error in potentiometric measurements. The difference in E_j between calibrator and sample (residual liquid junction potential) is responsible for this error and can be minimized and usually neglected in practice if the compositions of the calibrating solutions are matched as closely as possible to the sample with respect to ionic content and ionic strength. A reference electrolyte at high concentration, with equal cation and anion mobility, further helps to minimize the residual liquid junction potential. KCl at a concentration 2 mol/L or more is preferred. Differences of approximately −2% in the measurement of sodium by ISEs have been demonstrated

when the KCl concentration in the reference electrolyte is lowered from 3 to 0.5 mol/L.

The presence of erythrocytes in the sample may affect the magnitude of the residual liquid junction potential in a less predictable manner. For example, erythrocytes in a blood sample with a normal hematocrit are estimated to produce approximately 1.8 mmol/L positive error in the measurement of Na^+ by ISEs when an open, unrestrictive liquid-liquid junction is used. This bias may be minimized if a restrictive membrane or frit is used to modify the liquid-liquid junction.

ION-SELECTIVE MEMBRANE ELECTRODES

Membrane potentials are produced by permeability of certain types of membranes to selected anions or cations. Such membranes are used to fabricate ISEs that selectively interact with a single ionic species. The potential produced at the membrane–sample solution interface is proportional to the logarithm of the ionic activity or concentration of the ion in question. Measurements with ISEs are simple, rapid, nondestructive, capable of being used directly in a turbid matrix such as whole blood and applicable to a wide range of concentrations.

The ion-selective membrane controls the selectivity of the electrode. Ion-selective membranes are typically composed of glass, crystalline, or polymeric materials. The chemical composition of the membrane is designed to achieve an optimal permselectivity toward the ion of interest. In practice, other ions exhibit finite interaction with membrane sites and display some degree of interference for determination of an analyte ion. In clinical practice, if the interference exceeds an acceptable quantity, a correction is required.

The selectivity of an ISE for the ion of interest over interfering ions is:

$$E = E^0 + \left[\frac{2.303\ RT}{z_i F}\right] \log\left(a_i + \sum K_{i/j} a_j^{z_i/z_j}\right) \qquad \textbf{(10.9)}$$

where a_i = activity of the ion of interest; a_j = activity of the interfering ion; and $K_{i/j}$ = selectivity coefficient for the primary ion over the interfering ion. Low values indicate good selectivity for the analyte i over the interfering ion j; z_i = charge of the primary ion; and z_j = charge of the interfering ion. All other terms are identical to those in the Nernst equation (Eq. 10.3).

Various approaches may be used to determine the selectivity of an ISE for a primary ion over an interfering ion. A straightforward approach is the separate solution method, in which the potential of an ISE is determined in solutions of the primary and interfering ions separately, at equal ionic activities. The selectivity coefficient is then calculated as follows:

$$\log K_{ij} = \frac{E_j - E_i}{\frac{2.303\ RT}{nF}} + \left(1 - \frac{z_i}{z_j}\right) \log a_i \qquad \textbf{(10.10)}$$

An alternate approach to determine the selectivity coefficient is the fixed interference method, in which the potential response of an ISE to the primary ion is determined in solutions of constant activity of the interfering ion. The potential values obtained are plotted versus the logarithm of the activity of the primary ion a_i. The intersection of the extrapolation of the linear portions of this plot indicates the value of a_i to be used to calculate $K_{i/j}$:

$$K_{i/j} = a_i / a_j^{z_i/z_j} \qquad \textbf{(10.11)}$$

where all terms have the same definition as in Eq. 10.9. Traditionally the fixed interference method has been preferred because it more closely resembles a practical application of the sensor, in that primary and interfering ions are present simultaneously in solution and must compete for complexation sites in the ISE membrane.

Most ISEs used in clinical practice have sufficient selectivity and do not require correction for interfering ions. Table 10.1 shows required selectivity coefficients for the measurement of cations of interest in clinical chemistry over potentially interfering cations, assuming an acceptable maximum interference of 1% for the ion of interest.

Glass membrane and polymer membrane electrodes are two types of ISEs that are commonly used in clinical chemistry.

TABLE 10.1 Required Selectivities for Cation-Selective Ion-Selective Electrodes for Whole Blood, Plasma, and Serum Measurements

Primary Ion (I)	Required Selectivity Coefficient (log $K_{i/j}$) Interfering Cation (J)					
	H^+	Li^+	Na^+	K^+	Mg^{2+}	Ca^{2+}
H^+	—	−6.5	−8.5	−7.0	−7.7	−7.7
Li^+[a]	2.1	—	−4.3	−2.8	−3.5	−3.6
Na^+	4.4	−0.1	—	−0.6	−1.2	−1.3
K^+	2.8	−1.7	−3.6	—	−2.8	−2.9
Mg^{2+}	8.9	0.1	−3.9	−0.9	—	−2.4
Ca^{2+}	9.3	0.4	−3.6	−0.6	−1.9	—

[a]Assumes a therapeutic range for Li^+ between 0.7 and 1.5 mmol/L.
Ca^{2+}, Ionized calcium; *H^+*, hydrogen; *K^+*, potassium; *Li^+*, lithium; *Mg^{2+}*, magnesium; *Na^+*, sodium.

Glass Electrode

Glass membrane electrodes are used to measure pH and Na^+ and as an internal H^+ selective transducer for PCO_2 sensors. Glass electrode membranes are formulated from melts of silicon and/or aluminum oxide (Al_2O_3) mixed with oxides of alkaline earth or alkali metal cations. By varying the glass composition, electrodes with selectivity for H^+, Na^+, K^+, Li^+, rubidium (Rb^+), cesium (Cs^+), Ag^+, thallium (Tl^+), and ammonium (NH_4^+) have been demonstrated. However, glass electrodes for H^+ and Na^+ are currently the only types with sufficient selectivity over interfering ions to allow practical application in clinical chemistry. A typical formulation for H^+ selective glass is 72% silicon dioxide (SiO_2), 22% Na_2O, and 6% calcium oxide (CaO), which has a selectivity order of $H^+ >>> Na^+ > K^+$. This glass membrane has sufficient selectivity for H^+ over Na^+ to allow error-free measurements of

pH in the range of 7.0 to 8.0 ($[H^+] = 10^{-7}$ to 10^{-8} mol/L) in the presence of greater than 0.1 mol/L Na^+. Glass pH electrodes with selectivity coefficients ($K_{H/Na}$) over Na^+ of 10^{-7} and better have been realized. By altering the formulation of the glass membrane slightly to 71% SiO_2, 11% Na_2O, and 18% Al_2O_3, its selectivity order becomes $H^+ > Na^+ > K^+$. Thus the preference of the glass membrane for H^+ over Na^+ is greatly reduced, resulting in a practical sensor for Na^+ at pH values typically found in blood.

Polymer Membrane Electrodes

Polymer membrane ISEs are used for monitoring pH and for measuring electrolytes, including K^+, Na^+, Cl^-, Ca^{2+}, Li^+, Mg^{2+}, and carbonate (CO_3^{2-}) (for total CO_2 measurements). They are the predominant class of potentiometric electrodes used in clinical analysis instruments.

Response mechanisms of polymer membrane ISEs fall into three categories: (1) charged, dissociated ion exchanger; (2) charged associated carrier; and (3) neutral ion carrier (ionophore).

1. For charged, dissociated ion exchangers, selectivity is not controlled by any specific interaction between a ligand in the polymer membrane and an ion in solution but by extraction of ions into the membrane based on their lipophilic character. For example, dissociated anion exchanger–based electrodes using lipophilic quaternary ammonium salts as active membrane components are used commercially for the determination of Cl^- in whole blood, serum, and plasma, despite some limitations. Selectivity for this type of ISE is controlled by extraction of the ion into the organic membrane phase and is a function of the lipophilic character of the ion because no direct binding interaction occurs between the exchanger site and the anion in the membrane phase. Thus the selectivity order for a Cl^- ISE based on an anion exchanger is fixed as lipophilic anion R^- > perchlorate (ClO_4^-) > iodide (I^-) > nitrate (NO_3^-) > bromide (Br^-) > Cl^- > fluoride (F^-), where R^- represents anions with greater lipophilic character than ClO_4^-. Application of the Cl^- ion-exchange electrode is therefore limited to samples without significant concentrations of anions more lipophilic than Cl^-. For example, blood samples containing salicylate or thiocyanate will produce positive interference for the measurement of Cl^-. Repeated exposure of the electrode to the anticoagulant heparin will lead to loss of electrode sensitivity toward Cl^- because of the extraction of negatively charged heparin into the membrane.
2. For charged, associated carriers, there is specific ion exchange or complexation between the carrier in the polymer membrane and an ion in solution. An early charged-associated, ion-exchanger type ISE for Ca^{2+} was developed and commercialized for clinical application in the 1960s based on the Ca^{2+}-selective, ion-exchange/complexation properties of 2-ethylhexyl phosphoric acid dissolved in dioctyl phenyl phosphonate. A porous membrane was impregnated with this cocktail and mounted at the end of an electrode body. This type of sensor was referred to as the "liquid membrane" ISE. Later, a method was devised in which these ingredients could be cast into a plasticized poly(vinyl chloride) (PVC) membrane that was more rugged and convenient to use.

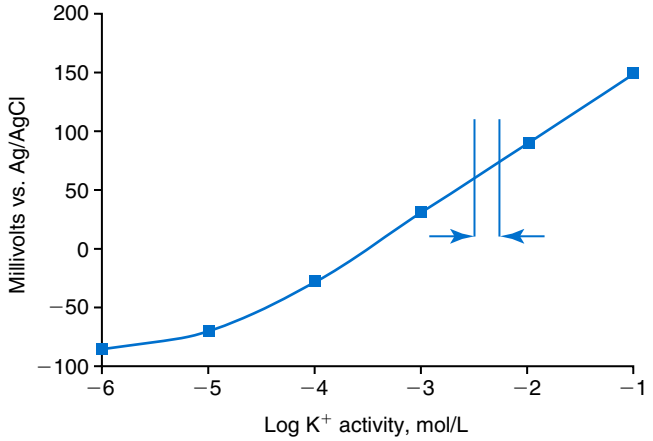
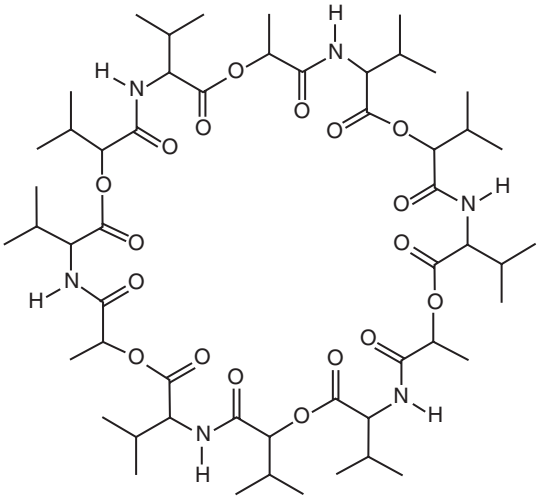
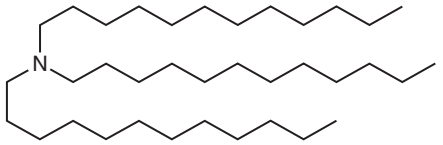


Fig. 10.2 Typical electromotive force response of potassium (K^+) selective membrane electrode to changes in activity of K^+ in the sample solution. Bracketed interval represents the normal reference interval of K^+ concentration in blood. *Ag/AgCl*, Silver–silver chloride. (From D'Orazio, P. [2002]. In K. Lewendrowski [Ed.]. *Clinical chemistry: laboratory management and clinical correlations* [p. 455]. Philadelphia: Lippincott, Williams and Wilkins.)

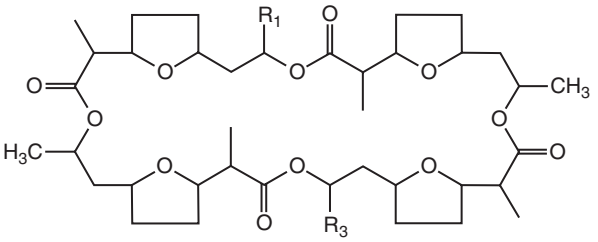
3. For neutral ion carriers (ionophores), the ionophore molecule in the polymer membrane can reversibly and selectively complex an ion of interest in the sample. The interaction is not based on charge, but rather the ion is accommodated based primarily on size within a cavity in the structure of the molecule. A breakthrough in the development and routine application of PVC-type ISEs was the discovery that the neutral antibiotic valinomycin could be incorporated into organic liquid membranes (and later plasticized PVC membranes), resulting in a sensor with high selectivity for K^+ over Na^+ ($K_{K/Na} = 2.5 \times 10^{-4}$). The K^+ ISE based on valinomycin was the first example of a neutral carrier ISE and is now used extensively for the routine measurement of K^+ in blood. Fig. 10.2 shows the response of the valinomycin-based K^+ ISE in the presence of physiologic concentrations of Na^+, Ca^{2+}, and Mg^{2+}. The wide linear range and excellent selectivity of this ISE over three orders of magnitude makes it suitable for the measurement of K^+ in blood and urine. The K^+ range in blood is only a small portion of the electrode linear range and is spanned by a total ΔEMF of approximately 9 mV. Interference from other cations, which is seen as deviation from linearity, is not apparent at K^+ activities more than 10^{-4} mol/L. Other, less selective polymer-based ISEs based on neutral carriers (e.g., for the measurement of Mg^{2+} and Li^+) are subject to interference from Ca^{2+}/Na^+ and Na^+, respectively, requiring simultaneous determination and correction for the presence of significant concentrations of these interfering ions. Studies regarding the relationship between molecular structure and ionic selectivity have resulted in the development of polymer-based ISEs using many naturally occurring and synthetic ionophores, with sufficient selectivity for application in clinical analysis. The chemical structures of several of these neutral ionophores are illustrated in Fig. 10.3.



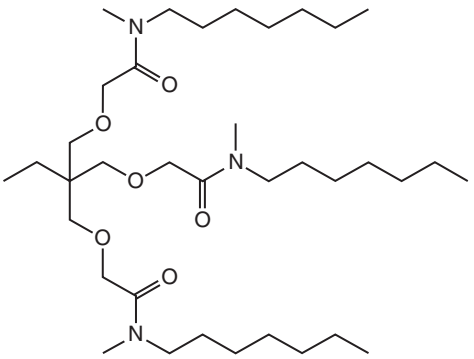
Valinomycin: K^+



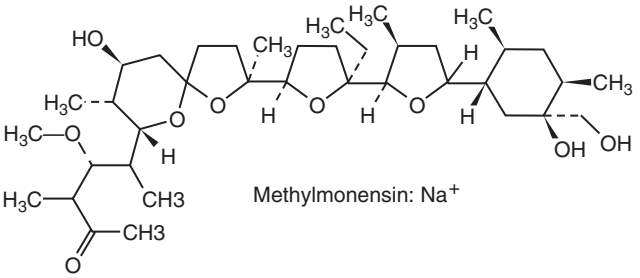
Tridodecylamine: H^+



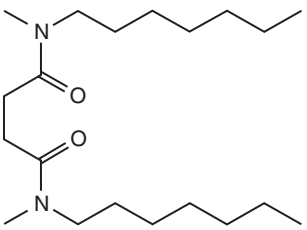
Noactin: NH_4^+



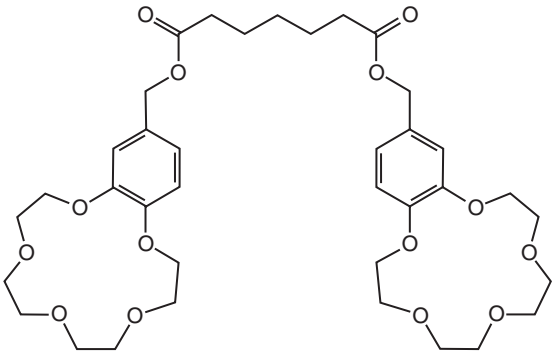
ETH 227: Na^+



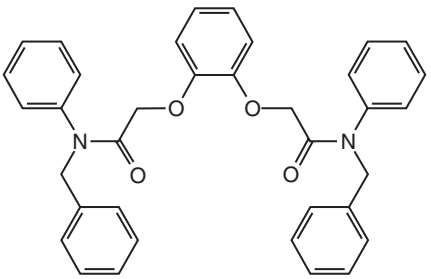
Methylmonensin: Na^+



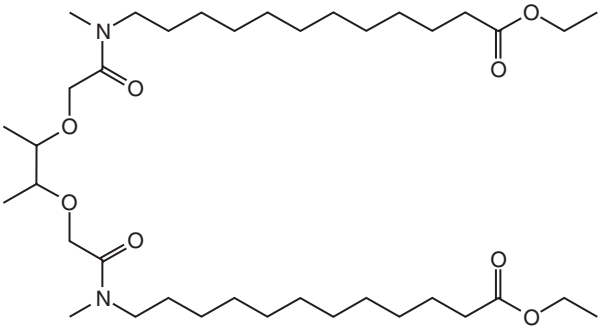
ETH 1117: Mg^{+2}



Bis(benzyl-15-crown-5)-heptanedoate: K^+



ETH 157: Na^+



ETH 1001: Ca^{+2}

Fig. 10.3 Structures of common ionophores used to fabricate polymer membrane–type ion-selective electrodes for clinical analysis.

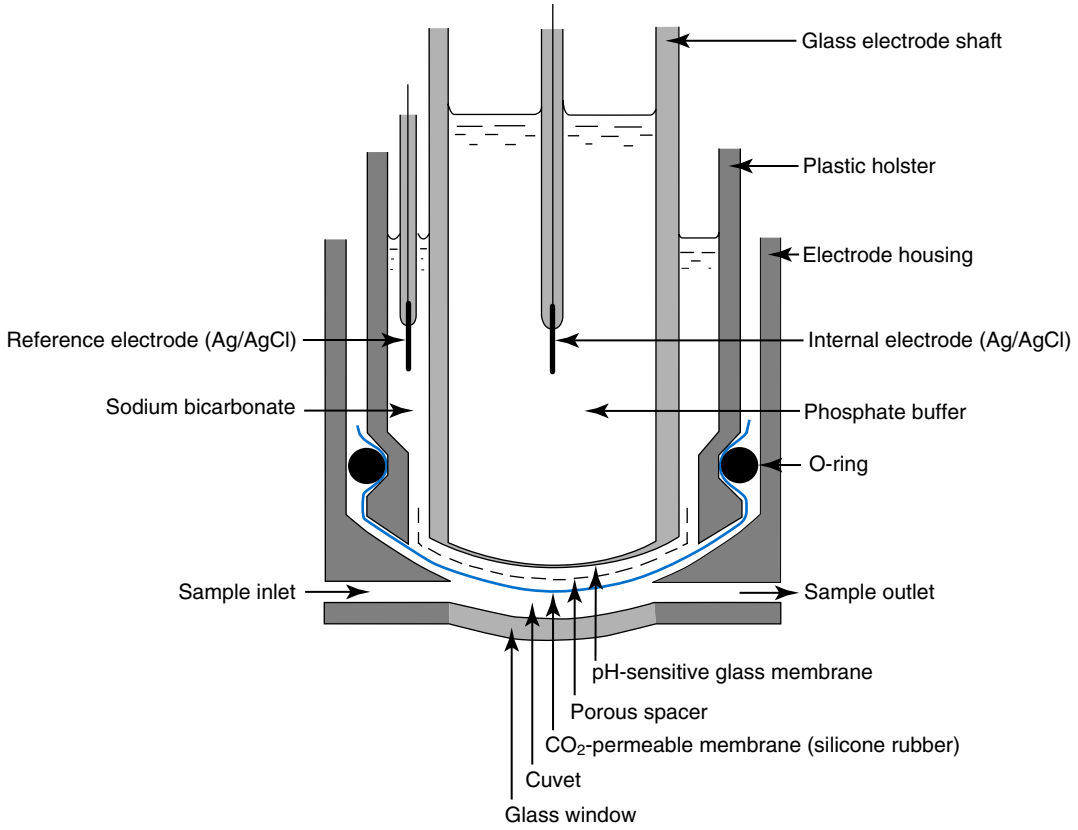


Fig. 10.4 Partial pressure of carbon dioxide sensor used to monitor carbon dioxide *(CO_2)* concentrations in blood samples. *Ag/AgCl*, Silver–silver chloride. (From Siggard-Andersen, O. [1974]. *The acid-base status of the blood* [4th ed., p. 172]. Baltimore, MD: Williams & Wilkins.)

High selectivity for the CO_3^{2-} anion can be achieved using a neutral carrier ionophore that possesses trifluoroacetophenone groups doped within a polymeric membrane. Such ionophores form negatively charged adducts with CO_3^{2-} anions, and the resulting electrodes have proven useful in commercial instruments for determination of total CO_2 in serum or plasma, after dilution of the sample to a pH value in the range of 8.5 to 9.0, where a significant fraction of total CO_2 exists as CO_3^{2-}.

Electrodes for Partial Pressure of Carbon Dioxide

Electrodes have been developed to measure PCO_2 in body fluids. The first PCO_2 electrode, developed in the 1950s, used a glass pH electrode as the internal element in a potentiometric cell. This important development paved the way for commercial availability of the three-channel blood analyzer (pH, PCO_2, PO_2) providing a complete picture of oxygenation and acid-base status of blood.

Fig. 10.4 shows a diagram of a typical electrode for PCO_2. A thin membrane (approximately 20 μm), permeable only to gases and water vapor, is in contact with the sample. Membranes of silicone rubber, Teflon, and other polymeric materials are suitable for this purpose. On the opposite side of the membrane is a thin electrolyte layer consisting of a weak bicarbonate salt (approximately 5 mmol/L) and a Cl^- salt. A pH electrode and an Ag/AgCl reference electrode are in contact with this solution. The PCO_2 electrode is a self-contained potentiometric cell. CO_2 gas from the sample or calibration matrix diffuses through the membrane and dissolves in the internal electrolyte layer. Carbonic acid is formed and dissociates, shifting the pH of the bicarbonate solution in the internal layer as follows:

$$CO_2 + H_2O \rightleftarrows H_2CO_3 \leftrightarrow H^+ + HCO_3^- \qquad \textbf{(10.12)}$$

and

$$\Delta \log PCO_{2(\text{sample})} \approx \Delta pH_{(\text{internal layer})} \qquad \textbf{(10.13)}$$

The relationship between the sample PCO_2 and the signal generated by the internal pH electrode is logarithmic and is governed by the Nernst equation. The electrode may be calibrated using precision gas mixtures or using solutions with stable PCO_2 concentrations. Although this style of electrode for PCO_2 has gained widespread use in modern blood gas analyzers, the format in which such sensors may be constructed is limited by size, shape, and the ability to fabricate the internal pH sensitive element.

A slightly different potentiometric cell for PCO_2 is shown in Fig. 10.5. This cell arrangement uses two PVC-type, pH-selective electrodes in a differential mode. The electrode membranes contain tridodecylamine neutral ionophore exhibiting high selectivity for H^+ (see Fig. 10.3). One electrode has an internal layer that is buffered, and the other is unbuffered, consisting of a low concentration of bicarbonate salt. CO_2 gas from the sample or calibration matrix diffuses across the

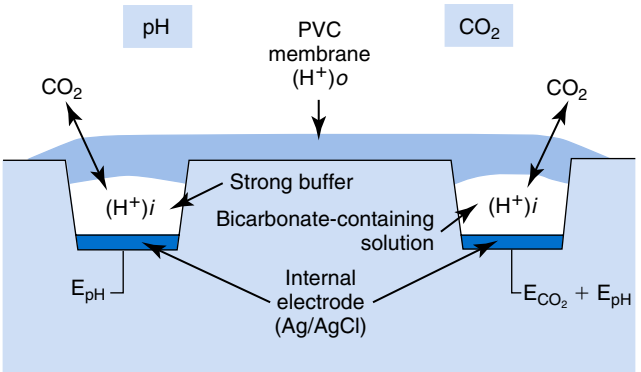


Fig. 10.5 Differential planar partial pressure of carbon dioxide potentiometric sensor design, based on two identical polymeric membrane pH electrodes, but with different internal reference electrolyte solutions. Both pH sensing membranes are prepared with a hydrogen (H^+)-selective ionophore. *Ag/AgCl*, Silver–silver chloride; *PVC*, poly(vinyl chloride).

outer H^+-selective PVC membranes of both sensors. On the unbuffered side, CO_2 diffusion produces a potential shift at the internal interface of the pH-responsive membrane proportional to the sample PCO_2 concentration. The signal at the electrode with the buffered internal layer is unaffected by the CO_2 that diffuses across the membrane. Consequently, one half of the sensor responds to pH alone, and the other half responds to both pH and PCO_2. The signal difference between the two electrodes cancels any contribution of sample pH to the overall measured cell potential. The differential signal is proportional only to PCO_2. Unlike the prior electrode, this differential potentiometric cell PCO_2 sensor has been commercialized in a planar format and is more easily adaptable to mass production in sensor arrays.

Direct Potentiometry by Ion-Selective Electrodes: Units of Measure and Reporting for Clinical Applications

Older analytical methods such as flame photometry for the measurement of electrolytes provide the total concentration *(c)* of a given ion in the sample, usually expressed in units of millimoles of ion per liter of sample. Molality *(m)* is a measure of the moles of ion per mass of water (millimoles per kilogram) in the sample. Using the Na^+ ion as an example, the relationship between concentration and molality is given by

$$c_{Na^+} = m_{Na^+} \times \rho H_2O \quad \textbf{(10.14)}$$

where ρH_2O = mass concentration of water in kilograms per liter. For normal blood plasma, the mass concentration of water is approximately 0.93 kg/L, but in specimens with increased lipids or protein, the value may be as low as 0.8 kg/L. In these specimens the difference between concentration and molality may be as great as 20%. A significant advantage of direct potentiometry by ISE for the measurement of electrolytes is that the technique is sensitive to molality and therefore is not affected by variations in the concentration of protein or lipids in the sample. Techniques such as flame photometry, ISE methods that require sample dilution (indirect potentiometry), and other photometric methods requiring sample dilution are affected by the presence of protein and lipids. In these methods, only the water phase of the sample is diluted, which produces results lower than molality as a function of the concentration of protein and lipids in the sample. Thus there is a risk for error, such as a falsely low Na^+ concentration (pseudohyponatremia), in cases of extremely increased protein and lipid concentrations (see also Chapter 24).

In addition to the difference between molality and concentration, measurement of ions by direct potentiometry provides yet another unit of measurement known as activity *(a)*, the concentration of free, unbound ion in solution. Unlike methods sensitive to ion concentration, ISEs do not sense the presence of complexed or electrostatically "hindered" ions in the sample. The relationship between activity and concentration, using Na^+ ion as an example, is expressed as

$$a_{Na^+} = \gamma_{Na^+} \times c_{Na^+} \quad \textbf{(10.15)}$$

where γ = dimensionless quantity known as the activity coefficient. The activity coefficient is primarily dependent on ionic strength of the sample as described by the Debye-Hückel equation:

$$\log \gamma = -\frac{\left(A \times z^2 \times I^{0.5}\right)}{1 + B \times a \times I^{0.5}} \quad \textbf{(10.16)}$$

where A and B = temperature-dependent constants ($A = 0.5213$ and $B = 3.305$ in water at 37°C); a = ion size parameter for a specific ion; and I = ionic strength ($I = 0.5\ \Sigma c \times z^2$, where z is the charge number of the ions). Eq. 10.16 shows that a decrease in the activity coefficient occurs with an increase in ionic strength. This effect is more pronounced when the charge (z) of the ion is higher. Activity coefficients for ions in biologic fluids, such as blood and serum, are difficult to calculate with accuracy because of the uncertain contribution of macromolecular ions, such as proteins, to the overall ionic strength. However, assuming that the normal ionic strength of blood plasma is 0.160 mol/kg, estimates of activity coefficients at 37°C are as follows: $Na^+ = 0.75$, $K^+ = 0.74$, and $Ca^{2+} = 0.31$. Referring to Eq. 10.15, activity and concentration will differ greatly in samples of physiologic ionic strength, especially for divalent ions.

Physiologically, ionic activity is assumed to be more relevant than concentration when chemical equilibria or biologic processes are considered. Practically, however, ionic concentration is the more familiar term in clinical practice, forming the basis of reference intervals and medical decision concentrations for electrolytes. Early in the evolution of ISEs as practical tools in clinical chemistry, it was decided that changing clinical reference intervals to a system based on activity instead of concentration was impractical and carried risk for clinical misinterpretation. A pragmatic approach for using ISEs in modern analyzers without changing established concentration-based reference intervals is to formulate calibration solutions with ionic strengths and ionic compositions as close as possible to those of blood plasma. In this way, the activity coefficient of each ion in the calibrating solutions approximates that in the sample matrix, allowing calibration and measurement of electrolytes in units of concentration instead of activity.

TABLE 10.2 Examples of Two-Value Calibrating Solutions for Measurement of pH and Electrolytes by Direct Potentiometry

Analyte	Calibration Point (mmol/L)	Slope Point (mmol/L)	Expected Signal Δ, (mV)
Na^+	140	110	6.6
K^+	4.0	8.0	18
Ca^{2+}	1.25	2.50	9
Cl^-	100	80	6
pH	7.38 (pH units)	6.84 (pH units)	32.4

Ionic strength adjusted to 160 mmol/kg with buffer salts and inert electrolytes.
Ca^{2+}, Ionized calcium; *Cl^-*, chloride; *K^+*, potassium; *Na^+*, sodium.

A typical set of solutions for multi-ISE calibration in an analyzer is shown in Table 10.2. Two points may be used to calibrate each ISE. The difference in the cell potential generated by these two solutions (ΔE) is used to calculate the response slope of the cell (slope = $\Delta E/\Delta \log c$), where c is the concentration of ion in each calibrating solution, substituted for activity. The standard electrode potential, E^0, is calculated as the y-intercept. Determination of the ion concentration in an unknown sample is then a straightforward solution of Eq. 10.9 after the cell potential generated by the sample is measured. The measured slope is used in place of the $2.303RT/z_iF$ term of Eq. 10.9. In the absence of significant influence from interfering ions on measurement of the primary ion (e.g., ≤1% interference on the measured value), contributions from a_j in Eq. 10.9 may be ignored.

Calibration of the cell is done in units of concentration; however, as mentioned earlier, direct potentiometry is sensitive to the molality of the ion, which is related to concentration by the water content of the sample (Eq. 10.14). The water content of the aqueous calibrating solutions shown in Table 10.2 is approximately 0.99 kg/L. The water content of normal blood plasma is approximately 0.93 kg/L. Molality is 7% greater than the concentration in this normal plasma specimen. The direct potentiometric cell will report results approximately 6% greater than the concentration in normal specimens because of this difference in water content between sample and the calibrator (0.99/0.93 = 1.06). Direct potentiometry presents an advantage in that the technique is not affected by the presence of protein and lipids in the sample; however, the application of clinical reference intervals based on concentration again poses a risk for confusion and clinical misinterpretation. Most manufacturers of electrolyte measurement systems have overcome this problem in a practical way by following Clinical and Laboratory Standards Institute (CLSI) guidelines that recommend the use of correlation factors to standardize ISE measurements to units of concentration. These factors may be obtained by standardizing the ISE measurement to certified reference materials based on human serum, with electrolyte values assigned in units of concentration. Appropriate correlation factors are then applied to sample calculations using algorithms resident in the instrument software.

POINTS TO REMEMBER

Advantages of Electrochemical Sensors for Whole Blood Measurements

- Electrochemical sensors measure the most important critical care analytes (gases, electrolytes, glucose) directly in whole blood without need for sample pretreatment or dilution.
- Measurement is rapid and nondestructive.
- Measurement is not affected by sample turbidity (red cells, lipids); only the analyte present in the water phase of plasma is measured.
- Simultaneous measurement of multiple analytes in the same blood sample is possible.

VOLTAMMETRY AND AMPEROMETRY

Voltammetric and amperometric techniques are among the most sensitive and widely applicable of all electroanalytical methods.

Basic Concepts

In contrast to potentiometry, voltammetric, and amperometric methods are based on electrolytic electrochemical cells in which an external voltage is applied to a polarizable working electrode versus a reference electrode and the resulting cathodic (for analytical reductions) or anodic (for analytical oxidations) measured current is proportional to the concentration of the analyte present in the sample. Current flows only if E_{appl} is greater than a certain voltage (decomposition voltage) determined by the thermodynamics for a given redox reaction of interest (Ox + ne^- {ReversReact} Red, which is defined by the E^0 value for that reaction [standard reduction potential]) and the kinetics for heterogeneous electron transfer at the interface of the working electrode. Often, slow kinetics of electron transfer for the redox reaction on a given inert working electrode (Pt, carbon, Au, etc.) mandates use of a much more negative (for reductions) or positive (for oxidations) E_{appl} than predicted based on the E^0 for the redox reaction. This is called an overpotential (η). Regardless of whether an overpotential for electron transfer exists, a specific oxidation or reduction reaction occurs at the surface of the working electrode in **voltammetry** and/or amperometry, and it is the charge transfer at this interface (current flow) that provides the analytical information.

For electrolytic cells that form the basis of voltammetric and amperometric methods:

$$E_{appl} = E_{cell} + \eta - iR_{cell} \quad \textbf{(10.17)}$$

where E_{cell} = thermodynamic potential between the working and reference electrodes in the absence of an applied external voltage. When the external voltage is greater or less than this equilibrium potential, plus or minus any overpotential (η), current will flow because of an oxidation or reduction reaction at the working electrode. A voltammogram is simply the plot of observed current, i, versus E_{appl} (Fig. 10.6). In **amperometry** (see later), a fixed voltage is applied, and the resulting current is monitored. The amount of current is inversely related to the resistance of the

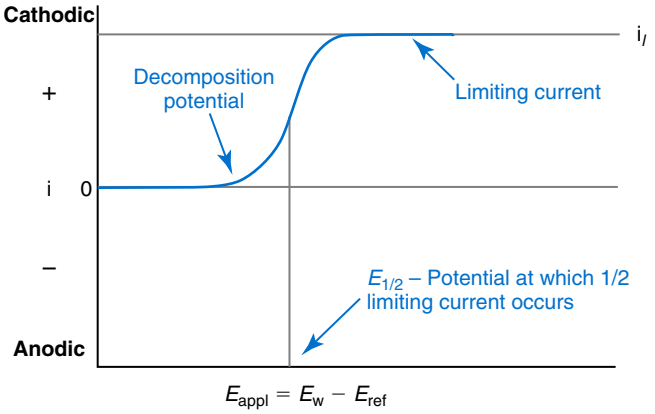


Fig. 10.6 Current versus voltage curve (voltammogram) obtained for oxidized species that were reduced at the surface of the working electrode, as the E_{appl} is scanned more negatively and the solution is stirred to yield a steady-state response. Decomposition potential is the applied potential where the oxidation or reduction reaction begins. Limiting current is the signal at steady state for the oxidation or reduction reaction and is proportional to concentration of the analyte of interest. E_{appl}, Applied voltage; E_w, potential at the working electrode; E_{ref}, potential at the reference electrode.

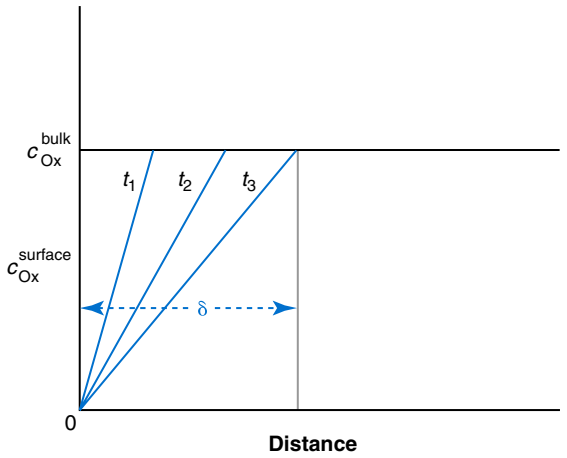


Fig. 10.7 Concept of electrochemical reaction increasing diffusion layer thickness (concentration polarization) of analytes via reduction or oxidation *(Ox)* at the surface of the working electrode. As time increases (t_1, t_2, t_3,...), the diffusion layer thickness at the surface of the working electrode *(δ)* grows quickly to a value determined by the degree of convection in the sample solution. c_{ox}^{bulk} and $c_{ox}^{surface}$ are concentrations of oxidizable species in bulk solution and at the surface of the working electrode, respectively.

electrolyte solution and to any "apparent" resistance that develops because of mass transfer of the analyte species to the surface of the working electrode. Because electrochemical reactions are heterogeneous, occurring only at the surface of the working electrode, the amount of current observed is also dependent on the surface area *(A)* of the working electrode.

When a potential is applied to a working electrode that will oxidize or reduce a species in the solution phase contacting the electrode, the electrochemical reaction causes the concentration of electroactive species to decrease at the surface of the electrode (Fig. 10.7), a process termed concentration polarization. In turn, this causes a concentration gradient of analyte species between the bulk sample solution and the surface of the electrode. When the bulk solution is stirred, the diffusion layer of the analyte grows out from the surface of the electrode quickly to a fixed distance, which is controlled by how vigorously the solution is stirred. This diffusion layer is termed the Nernst layer and has a finite thickness *(δ)* after a relatively short time period when the solution is moving (convection). Voltammetry carried out in the presence of convection (by stirring the solution, rotating the electrode, flowing solution by the electrode, and so on) is called steady-state voltammetry. When the solution is not moving, the diffusion layer grows further and further with time (i.e., not constant), creating larger and larger δ values over time. This is termed non–steady-state voltammetry and often results in peak currents in *i* versus E_{appl} plots for electrolytic cells.

In steady-state voltammetry, when the potential of the working electrode is scanned past a value that will cause an electrochemical reaction, the current will rise rapidly and then will plateau, even as E_{appl} changes further. Fig. 10.6 illustrates such a wave for a hypothetical reduction of an oxidized species (Ox) via an n electron reduction to a reduced species (Red). When the applied potential is much more negative than required, the current reaches a limiting value (termed the limiting current, i_l). This limiting current is proportional to the concentration of the electroactive species (Ox in this case) as expressed by the following equation:

$$i_l = nFA\left(\frac{D}{\delta}\right)c_{Ox} \qquad \textbf{(10.18)}$$

where i = measured current in amperes; n = the number of electrons in the electrochemical reaction (reduction in this case); F = Faraday constant (96,487 coulombs/mol); A = electrochemical surface area of the working electrode (in square centimeters, assuming a planar electrode geometry); D = diffusion coefficient (in square centimeters per second) of the electroactive species (Ox in this case); δ = diffusion layer thickness (in centimeters); and c = concentration of the analyte species (in moles per cubic centimeter). Note that Eq. 10.18 indicates a linear relationship for limiting current and concentration. The same equation applies for detecting reduced species by an oxidation reaction at the working electrode. In this case, by convention, the resulting anodic current is considered a negative current. As shown in Fig. 10.6, the potential of the working electrode that corresponds to a current that is exactly one-half the limiting current is termed $E_{½}$. This value is independent of analyte concentration. The $E_{½}$ is determined by the thermodynamics (E^0) of the given redox reaction, the solution conditions (e.g., if protons are involved in reaction, then pH will influence $E_{½}$), and any overpotential caused by slow electron transfer at a particular working electrode surface. The $E_{½}$ values are indicative of a given species undergoing an electrochemical reaction under specified conditions; hence the $E_{½}$ values enable distinguishing one electroactive species from another in the same sample. If the $E_{½}$ values for various species differ significantly (e.g., more than 120 mV), then measurements of several limiting currents in a given voltammogram yield quantitative results for several different species simultaneously.

Electrochemical cells used to carry out voltammetric or amperometric measurements can involve a two- or three-electrode configuration. In the two-electrode mode, external voltage is applied between the working electrode and a reference electrode, and the current is monitored. Because current passes through the reference electrode, such current can alter the surface concentration of electroactive species that poises the half-cell potential of the reference electrode, changing its value by a concentration polarization process. For example, if an Ag/AgCl reference electrode were used in a cell in which a reduction reaction for the analyte occurs at the working electrode, then an oxidation reaction would take place at the surface of the reference electrode:

$$Ag^0 + Cl^- \rightarrow AgCl_{(s)} + 1e^- \qquad \textbf{(10.19)}$$

Hence the activity and/or concentration of Cl^- ions near the surface of the electrode would decrease, which would make the potential of the reference electrode more positive than its true equilibrium value based on the actual activity of the Cl^- ion in the reference half-cell, because the Nernst equation for this half-cell is:

$$E_{Ag/AgCl} = E^0_{AgCl/AgCl} - 0.059 \log\left(a^{surface}_{Cl^-}\right) \qquad \textbf{(10.20)}$$

Such concentration polarization of the reference electrode is prevented by keeping the current density (amperes per square centimeter) low at the reference electrode. This is achieved in practice by making sure that the area of the working electrode in the electrochemical cell is much smaller than the surface area of the reference electrode; hence the total current flow will be limited by this much smaller area, and current density values at the reference will be very small to prevent concentration polarization.

A three-electrode potentiostat is often used to completely eliminate changes in reference electrode half-cell potentials. In simple terms, the potentiostat applies a voltage to the working electrode, which is measured versus a reference electrode via a zero-current potentiometric-type measurement, but the current flow is between the working electrode and a third electrode, called the counter electrode. Thus if reduction takes place at the working electrode, oxidation would occur at the counter electrode, but no net reaction would take place at the surface of the reference electrode because no current flows through this electrode. A potentiostat circuit is relatively simple to construct using modern operational amplifiers.

In voltammetric methods, E_{appl} is varied via some waveform to alter the working electrode potential as a function of time and the resulting current measured. The current change occurs at the decomposition potential range, which is best when specific for a given analyte. However, the location of the current response as a function of E_{appl} provides information on the nature of the species present (e.g., $E_{½}$), along with a concentration-dependent signal. This scan of E_{appl} can be linear (linear sweep voltammetry), or it can have more complex shapes to enhance sensitivity for monitoring concentration of a given electroactive species (e.g., normal pulsed voltammetry, differential pulse voltammetry, square wave voltammetry).

Amperometric methods differ from voltammetry in that E_{appl} is fixed, generally at a potential value in the limiting current plateau region of the voltammogram, and the resulting current is proportional to concentration. Amperometry can be more sensitive than common voltammetric methods because background charging currents, which arise from changing the E_{appl} as a function of time in voltammetry, do not exist. Hence, when selectivity can be ensured at a given E_{appl} value, amperometry may be preferred to voltammetric methods for more sensitive quantitative measurements.

Applications

Molecular O_2 is capable of undergoing several reduction reactions, all with a significant overpotential at solid electrodes, such as Pt, Au, or Ag. For example, the following reaction:

$$O_2 + 2H_2O + 4e^- \rightarrow 4OH^-$$
$$\left(E^0 = +0.179 \text{ vs. Ag/AgCl};\ 1\ \frac{mol}{L}Cl^-\right) \qquad \textbf{(10.21)}$$

exhibits an $E_{½}$ at approximately −0.500 V on a Pt electrode (vs. a Ag/AgCl reference electrode), with a limiting current plateau beginning at approximately −0.600 V. This reaction is used to monitor the PO_2 in blood, which is the basis of the widely used amperometric O_2 sensor (Fig. 10.8). This device uses a small area planar Pt electrode as a working electrode (encased in insulating glass or other material) and an Ag/AgCl reference electrode, typically with a cylindrical design. This two-electrode electrolytic cell is placed within a sensor housing, on which a gas-permeable membrane (e.g., polypropylene, silicone rubber, Teflon) is held at the distal end. The inner working Pt electrode is pressed tightly against the gas-permeable membrane to create a thin film of internal electrolyte solution (usually buffer with KCl added). O_2 in the sample can permeate across the membrane and is reduced in accordance with the above electrochemical reaction. An E_{appl} of −0.650 or −0.700 V versus Ag/AgCl (within the limiting current regime) to the Pt working electrode will result in an observed current that is proportional to PO_2 present in the sample (including whole blood). In the absence of any O_2, the current at this applied voltage under amperometric conditions will be near zero.

The outer gas-permeable membrane enables the electrode to detect O_2 with high selectivity over other easily reduced species that might be present in a given sample (e.g., metal ions, cystine). Only other gas species or highly lipophilic organic species can partition into and pass through such gas-permeable membranes. One type of interference in clinical samples can be caused by certain anesthesia gases, such as nitrous oxide, halothane, and isoflurane. These species can also diffuse through the outer membrane of the sensor, can be electrochemically reduced at the Pt electrode, and can yield a falsely high value for the measurement of PO_2. However, optimized gas-permeable membrane materials and appropriate control of the applied potential to the cathode of the sensor have greatly reduced this problem in modern instruments. The outer gas-permeable membrane also helps restrict

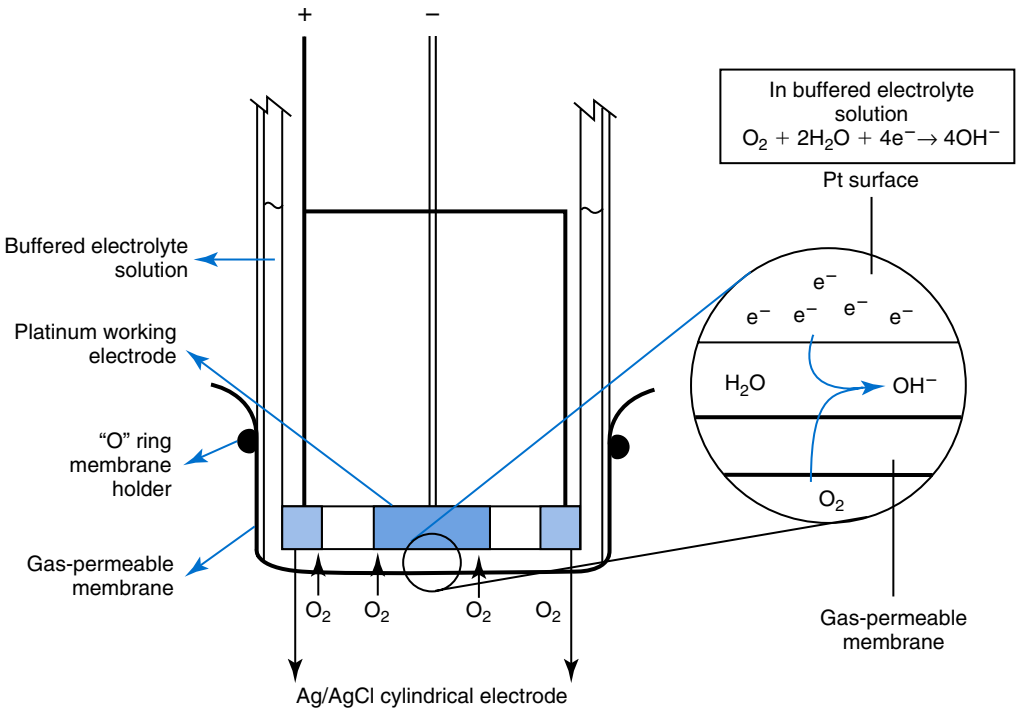


Fig. 10.8 Design of an amperometric oxygen sensor used to monitor partial pressure of oxygen in blood. *Ag/AgCl*, Silver–silver chloride; *H_2O*, water; *Pt*, platinum.

diffusion of analyte to the inner working electrode; hence the membrane can control the mass transport of analyte (D/δ term in Eq. 10.18), such that in the presence or absence of sample convection, mass transport of O_2 to the surface of the Pt working electrode is essentially the same.

Amperometric PO_2 sensors can be used to detect other gas species by altering the applied voltage to the working electrode. For example, it is possible to detect nitric oxide (NO) with high selectivity using a similar gas electrode design in which the Pt is polarized at +0.900 versus Ag/AgCl to oxidize diffusing NO to nitrate at the Pt anode. Such NO sensors can be used for a variety of biomedically important studies to measure the amount of NO locally at or near the surface of various NO-producing cells.

Beyond amperometric devices, one specialized method of detecting trace concentrations of toxic metal ions in clinical samples is anodic stripping voltammetry (ASV). In ASV a carbon working electrode is used (sometimes further coated with an Hg film), and the E_{appl} is first fixed at a negative E_{appl} voltage so that all metal ions in the solution will be reduced to elemental metals (M^0) within the Hg film and/or on the surface of the carbon. Then the E_{appl} is scanned more positively, and reduced metals deposited in and/or on the surface of the working electrode are reoxidized, giving a large anodic current peak proportional to the concentration of metal ions in the original sample. The potential at which these peaks are observed indicates which metal is present, and the height of the stripping peak current is directly proportional to the concentration of the metal ion in the original sample. Such ASV techniques can be used to detect the total concentration of lead in whole blood samples, providing a rapid screening method for lead exposure and poisoning.

Although voltammetric and/or amperometric techniques can be applied to detect a wide range of species, the selectivity offered for measurements in complex clinical samples—where many species can be electroactive—is rather limited. For example, as stated in the previous discussion relevant to the O_2 sensor, in the absence of the gas-permeable membrane, other species that can be reduced at or near the same E_{appl} as O_2 would cause significant interference.

To expand the range of analytes that can be detected with voltammetric and/or amperometric methods, electrochemical techniques can be used as highly sensitive detectors for high-performance liquid chromatography systems (see Chapter 12). In liquid chromatography with electrochemical detection (LC-EC), eluting solutes are detected by flow-through electrodes (usually carbon or Hg) designed to have extremely low dead volumes (Fig. 10.9). The electrodes can be operated in amperometric or voltammetric modes (with high scan speeds), and several electrodes can be operated simultaneously in series or in parallel flow arrangements to gain additional selectivity. For example, homocysteine can be measured with (1) the addition of reducing agents to a serum sample to generate free homocysteine, (2) precipitation of proteins in the sample (with trichloroacetic acid), and (3) separation of the serum components on a reversed-phase octadecylsilane high-performance liquid chromatography column. The eluting homocysteine is detected and measured with online electrochemical detection via homocysteine oxidation to the corresponding mercuric dithiolate complex:

$$2RSH + Hg \rightarrow Hg(RS)_2 + 2H^+ + 2e^- \qquad \textbf{(10.22)}$$

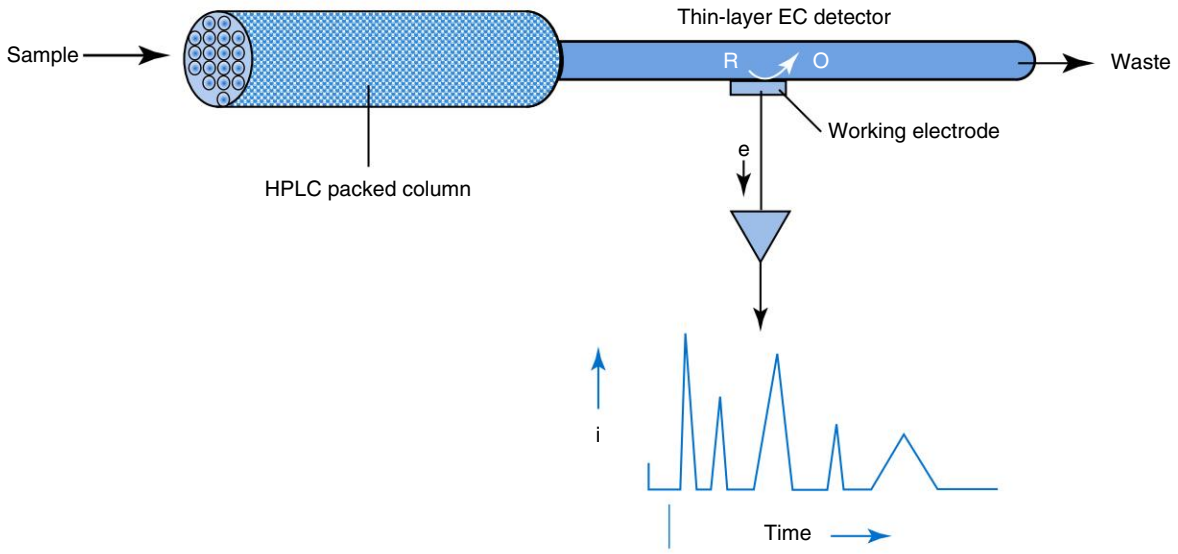


Fig. 10.9 Schematic of liquid chromatography with electrochemical detection system, with electrochemical detector monitoring elution of analytes from a high-performance liquid chromatography *(HPLC)* column by their oxidation or reduction at a suitable thin-layer working electrode. *EC*, Electrochemical.

using a thin-layer Hg/Au amalgam electrode poised at +0.150 V versus Ag/AgCl. Integration of the eluting band for homocysteine provides quantitative results, with high selectivity. Similarly, catechols and catecholamines can be readily detected in serum by a similar LC-EC method, with eluted catechols oxidized to quinones at a flow-through carbon working electrode poised at potentials typically more than +0.200 V versus Ag/AgCl. Furthermore, a host of therapeutic drugs can also be quantified in serum or urine via LC-EC methods.

POINTS TO REMEMBER

Predominant Types of Electrochemical Sensor Technologies Used in Whole Blood Analyzers

- Potentiometry: for measurement of pH, PCO_2, and electrolytes
- Amperometry: for measurement of PO_2 and as the basis for whole blood glucose and lactate biosensors
- Conductometry: for measurement of hematocrit

CONDUCTOMETRY

Conductometry is an electrochemical technique used to determine the quantity of an analyte present in a mixture by measuring its effect on the electrical conductivity of the mixture. It is the measure of the ability of ions in solution to carry current under the influence of a potential difference. In a conductometric cell, potential is applied between two inert metal electrodes. An alternating potential with a frequency between 100 and 3000 Hz is used to prevent polarization of the electrodes. A decrease in solution resistance results in an increase in conductance, and more current is passed between the electrodes. The resulting current flow is also alternating. The current is directly proportional to solution conductance. Conductance is considered the inverse of resistance and may be expressed in units of ohm^{-1} (siemens). In clinical analysis, conductometry is frequently used for measurement of the volume fraction of erythrocytes in whole blood (hematocrit) and as the transduction mechanism for some biosensors.

Erythrocytes act as electrical insulators because of their lipid-based membrane composition. This phenomenon was used first in the 1940s to measure the volume fraction of erythrocytes in whole blood (hematocrit) by conductivity and is used currently to measure hematocrit on multianalyte instruments for clinical analysis. The conductivity of whole blood depends not only on the volume fraction and shape of the erythrocytes but also on the conductivity of the surrounding plasma. An increase in the volume fraction of erythrocytes that are less conductive than the surrounding plasma leads to a decrease in conductivity shown by the following relationship:

$$G_b = \frac{a}{1 + \frac{H}{100 - H} \times c} \qquad (10.23)$$

where G_b = conductivity of whole blood; a = plasma conductivity; H = hematocrit in percent; and c = factor for erythrocyte orientation. In practice, plasma conductivity also contains correction factors for Na^+ and K^+ concentrations. These cations are usually measured in conjunction with hematocrit on systems designed for clinical analysis.

Conductivity-based hematocrit measurements have limitations. Abnormal protein concentrations will change plasma conductivity and interfere with the measurement. Low protein concentrations resulting from dilution of blood with protein-free electrolyte solutions during cardiopulmonary bypass surgery will result in erroneously low hematocrit values by conductivity. Preanalytical variables, such as insufficient mixing of the sample, will also lead to errors. Hemoglobin is the preferred analyte to monitor blood loss and the need for transfusion during trauma and surgery. However, electrochemical measurement of hematocrit in conjunction with blood gases and electrolytes remains in use mainly because of its simplicity and convenience, despite some limitations.

Another clinical application of conductance is for electronic counting of blood cells in suspension. Termed the Coulter principle, it relies on the fact that the conductivity of blood cells is lower than that of a salt solution used as a suspension medium. The cell suspension is forced to flow through a tiny orifice. Two electrodes are placed on either side of the orifice, and a constant current is established between the electrodes. Each time a cell passes through the orifice, resistance increases; this causes a spike in the electrical potential difference between the electrodes. The pulses are then amplified and counted.

COULOMETRY

Coulometry measures the electrical charge passing between two electrodes in an electrochemical cell. The amount of charge passing between the electrodes is directly proportional to oxidation or reduction of an electroactive substance at one of the electrodes. The number of coulombs transferred in this process is related to the absolute amount of electroactive substance by the Faraday law:

$$Q = n \times N \times F \qquad (10.24)$$

where Q = amount of charge passing through the cell (unit: C = coulomb = ampere × second); n = the number of electrons transferred in the oxidation or reduction reaction; N = the amount of substance reduced or oxidized in moles; and F = Faraday constant (96,487 coulombs/mol).

Current is the amount of charge passed per unit time (ampere = coulomb per second). Coulometry is used in clinical applications for the determination of Cl^- in serum or plasma and as the mode of transduction in certain types of biosensors.

Commercial coulometric titrators have been developed for determination of Cl^-. A constant current is applied between a silver wire (anode) and a platinum wire (cathode). At the anode, Ag is oxidized to Ag^+. At the cathode, H^+ is reduced to H_2 gas. At a constant applied current, the number of coulombs passed between the anode and the cathode is directly proportional to time (coulombs = amperes × seconds). Therefore the absolute number of Ag^+ ions produced at the anode may be calculated from the amount of time current passes through it. In the presence of Cl^-, Ag^+ ions are precipitated as $AgCl_{(s)}$, and the amount of free Ag^+ in solution is low. When all Cl^- ions have been complexed, a sudden increase in the concentration of Ag^+ in solution is noted. Excess Ag^+ is sensed amperometrically at a second Ag electrode, which is polarized at negative potential. The excess Ag^+ is reduced to Ag, producing a current. When this current exceeds a certain value, the titration is stopped. The absolute number of Cl^- ions present in the sample is calculated from the time during which titration with Ag^+ was in progress. Because the volumetric amount of serum or plasma sample is known, it is possible to calculate the concentration of Cl^- in the sample. Coulometric titration is one of the most accurate electrochemical techniques because the method measures the absolute quantity of electroactive substance in the sample.

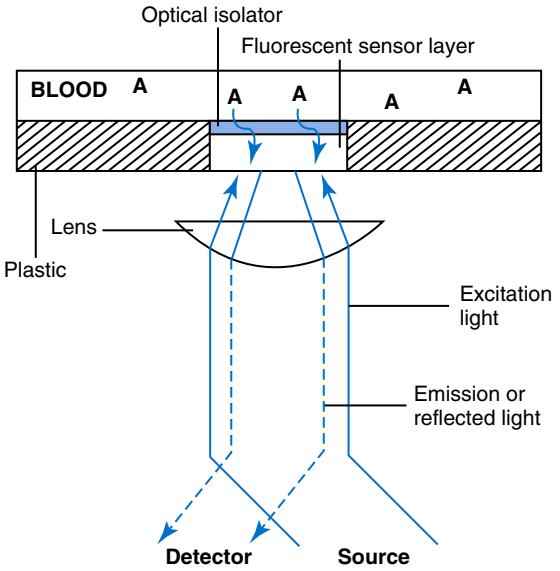


Fig. 10.10 General design for an in vitro optical sensor to detect a given analyte *(A)* in blood. Polymer film contains dye that changes spectral properties in proportion to the amount of analyte in the sample phase. Example shown is for a sensing film that changes luminescence (fluorescence or phosphorescence).

Coulometry is considered the "gold standard" for determination of Cl^- in serum or plasma. However, the method is subject to interference from anions in the sample that have greater affinity for Ag^+ than Cl^- (e.g., bromide) and is not commonly used currently in clinical laboratories.

OPTICAL CHEMICAL SENSORS

An **optode** is an optical sensor used in analytical instruments to measure blood gases and electrolytes. Optodes have certain advantages over electrodes, including (1) ease of miniaturization, (2) less noise (no transduction wires), (3) potential long-term stability using ratiometric-type measurements at multiple wavelengths, and (4) they do not require a separate reference electrode. These advantages initially promoted the development of optical sensor technology for the design of intravascular blood gas sensors. However, the same basic sensing principles can be used in clinical chemistry instrumentation designed for more classic in vitro measurements on discrete samples. In such systems, light can be brought to and from the sensing site by optical fibers or simply by appropriate positioning of light sources (light-emitting diodes), filters, and photodetectors to monitor absorbance (by reflectance), fluorescence, or phosphorescence (Fig. 10.10).

Basic Concepts

Optical sensors devised for PO_2 measurements are typically based on immobilization of certain organic dyes (e.g., pyrene, diphenylphenanthrene, phenanthrene, fluoranthene) or metal ligand complexes (e.g., ruthenium[II] tris[dipyridine], Pt and palladium metalloporphyrins) within hydrophobic polymer films (e.g., silicone rubber) in which O_2 is soluble. The fluorescence or phosphorescence of such species at a given

wavelength is often quenched in the presence of paramagnetic species, including molecular O_2. In the case of embedded fluorescent dyes, the intensity of the emitted fluorescence of such films will decrease in proportion to the PO_2 in the sample in contact with the polymer film, in accordance with:

$$\frac{I_0}{I_{PO_2}} = 1 + kPO_2 \qquad (10.25)$$

where I_0 = fluorescence intensity in the absence of any O_2; I_{PO_2} = fluorescence intensity at a given PO_2; and k = a quenching constant for the particular fluorophore used.

Hence a linear relationship exists between the ratio I_0/I_{PO_2} and the PO_2 in the sample phase. The larger the quenching constant, the greater is the degree of quenching for the given fluorophore. However, it is important that the quenching constant is in a range that will yield linear behavior over the physiologically relevant range of PO_2 in blood. For example, if k is too large, then the maximum quenching possible will occur over a range of PO_2 that is less than physiologic.

Phosphorescence intensity or phosphorescence lifetime measurements of immobilized metal ligand complexes can also be used (i.e., binding of O_2 decreases excited state lifetimes). Sensors based on changes in luminescent lifetime have the inherent advantage of being insensitive to perturbations in the optical path length and the amount of active dye present in the sensing layer.

Optical pH sensors require immobilization of appropriate pH indicators (e.g., fluorescein, 8-hydroxy-1,2,6-pyrene trisulfonate, phenol red) within thin layers of hydrophilic polymers (e.g., hydrogels) because equilibrium access of protons to the indicator is essential. The absorbance or fluorescence of the protonated or deprotonated form of the species can be used for sensing purposes. One issue with respect to using immobilized indicators for accurate physiologic pH measurements is the effect of ionic strength on the acid dissociation constant (pK_a) of the indicator. Because optical sensors measure the concentration of protonated or deprotonated dye as an indirect measure of H^+ ion activity, variations in ionic strength of the physiologic sample can influence the accuracy of the pH measurement.

Applications

Optical sensors suitable for the determination of PCO_2 use optical pH transducers (with immobilized indicators) as inner transducers in an arrangement similar to a classical electrochemical sensor design (see Fig. 10.4). The addition of bicarbonate salt within the pH-sensing hydrogel layer creates the required electrolyte film layer, which varies in pH depending on the PCO_2 in equilibrium with the film. The optical pH sensor is covered by an outer gas-permeable hydrophobic film (e.g., silicone rubber) to prevent proton access, yet it allows CO_2 equilibration with the pH sensing layer. As the PCO_2 in the sample increases, the pH of the bicarbonate layer decreases, and the corresponding decrease in the deprotonated form of the indicator (or increase in the protonated form) is sensed optically.

Two approaches have been used to sense electrolytes optically in physiologic samples. One method uses many of the same lipophilic ionophores developed for polymer membrane–type ISEs (see Fig. 10.3). These species are doped into thin hydrophobic polymeric films, along with a lipophilic pH indicator. In the case of cation ionophores (e.g., valinomycin for sensing K^+), when cations from the sample are extracted by the ionophore into the thin film, the pH indicator (RH) loses a proton to the sample phase to maintain charge neutrality within the organic film (yielding R^-). This results in a change in the optical absorption or fluorescence spectrum of the polymer layer. If the thickness of the films is kept at less than 10 μm, equilibrium response times on the order of less than 1 minute have been achieved. The main limitation of this design is that the pH of the sample phase also influences the overall extraction equilibrium for ions into the film. Thus simultaneous and independent measurement of sample pH is required or buffered dilution and/or pH control of the sample phase is necessary to obtain accurate measurements of electrolytes.

A second technique used to sense electrolyte ions is immobilization of a cation and/or anion recognition agent within a hydrogel matrix, similar to the pH sensors described earlier. The recognition agent in this case is not usually lipophilic; therefore it must be covalently anchored to the hydrogel, so that it does not leach into the sample phase. The agent is designed so that selective cation or anion binding alters the absorbance or fluorescence spectrum of the species within the hydrogel. Typically this is achieved by linking ion recognition and chromophoric properties within a single organic molecule. Such ion sensors have been used successfully in at least one commercial blood gas–electrolyte analyzer using an array of sensors of the generic design similar to that in Fig. 10.10.

BIOSENSORS

A **biosensor** is a specific type of chemical sensor consisting of a biologic recognition element and a physicochemical transducer, often an electrochemical or an optical device. The biologic element is capable of recognizing the presence and activity and/or concentration of a specific analyte in solution. The recognition may be a *biocatalytic reaction (enzyme-based biosensor)* or a *binding process (affinity-based biosensor)* when the recognition element is, for example, an antibody, DNA segment, or cell receptor. Interaction of the recognition element with a target analyte results in a measurable change in a solution property locally at the surface of the device, such as formation of a product or consumption of a reactant. The transducer converts the change in solution property into a quantifiable electrical signal. The mode of transduction may be one of several, including electrochemical or optical measurement and measurement of mass or heat. The present discussion is limited to biosensors based on electrochemical and optical modes of transduction because they constitute most biosensors used for clinical applications.

POINTS TO REMEMBER

Biosensors

- Biosensors measure substances lacking direct electroactive or optical properties.
- Biosensors consist of a biologic recognition element and a physicochemical (e.g., electrochemical or optical) transducer.
- Interaction of a biologic recognition element and an analyte results in either a biocatalytic reaction or a binding process, producing a measurable signal.
- Biosensors are used in most commercial blood glucose meters.
- Sensors for whole blood measurements without a biologic recognition element (e.g., ISEs) are not considered biosensors.

Enzyme-Based Biosensors With Amperometric Detection

Enzyme-based biosensors based on electrochemical transducers, specifically amperometric electrodes, are most commonly used for clinical analyses and are frequently cited in the literature. Most of the current blood glucose meters are based on measurements using enzyme-based biosensors with amperometric detection, with glucose sensors produced in single-use formats. The first amperometric biosensor was used to measure glucose in blood and was based on immobilized glucose oxidase on the surface of an amperometric PO_2 sensor. A solution of glucose oxidase was physically entrapped between the gas-permeable membrane of the PO_2 electrode and an outer semipermeable membrane (Fig. 10.11). The outer membrane was of a low-molecular-weight cutoff to allow the substrate (glucose) and O_2 from the sample to pass, but not proteins and other macromolecules. In this way, enzymes could be concentrated at the sensor's surface. Oxidation of glucose, catalyzed by glucose oxidase as follows:

$$\text{Glucose} + O_2 \xrightarrow{\text{glucose oxidase}} \text{gluconic acid} + H_2O_2 \quad \textbf{(10.26)}$$

consumes O_2 near the surface of the sensor. The rate of decrease in PO_2 is a function of the glucose concentration and is monitored by the PO_2 electrode. A steady-state PO_2 can be achieved at the surface in a short period of time, yielding a steady-state current value that decreases as a function of glucose concentration in the sample.

If the polarizing voltage of the PO_2 electrode is reversed, making the Pt electrode positive (anode) relative to the Ag/AgCl reference electrode, and if the gas-permeable membrane is replaced with a hydrophilic outer membrane containing the immobilized enzyme, it is possible to oxidize the hydrogen peroxide (H_2O_2) produced by the glucose oxidase as follows:

$$H_2O_2 - 2e^- \rightarrow 2H^+ + O_2 \quad \textbf{(10.27)}$$

The steady-state current produced is now directly proportional to the concentration of glucose in the sample.

In practice, a sufficiently high voltage (overpotential) must be applied to the platinum anode to drive the oxidation of the H_2O_2. An applied voltage of +0.7 V or greater (relative to Ag/AgCl) is typically used. Fig. 10.12 illustrates this basic H_2O_2 detection design, which is suitable for use in devising clinically useful sensors for glucose but also for a host of other substrates for which suitable oxidase enzymes generate H_2O_2.

Immobilization of enzymes in the early biosensors was a simple entrapment method behind a membrane of low-molecular-weight cutoff; this approach is still used in some commercial applications. Many other schemes for enzyme immobilization for biosensor development have been suggested. The most common are cross-linking of the enzyme with an inert protein (e.g., bovine serum albumin) using glutaraldehyde, simple adsorption of enzyme to electrode surfaces, and covalent binding of enzymes to insoluble carriers (e.g., nylon or glass). Another immobilization technique involves bulk modification of an electrode material, mixing

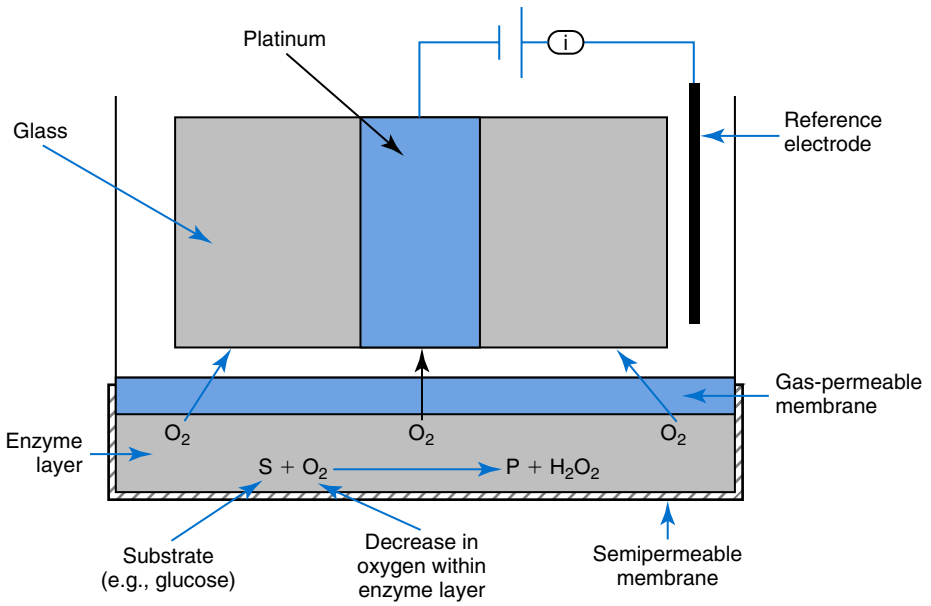


Fig. 10.11 Illustration of an enzyme-based biosensor prepared using oxidase enzyme immobilized at the surface of an amperometric partial pressure of oxygen sensor. Increase in substrate concentration (S) reduces the amount of oxygen present at the surface of the sensor. H_2O_2, Hydrogen peroxide.

enzymes with carbon paste, which serves as both the enzyme immobilization matrix and the electroactive surface.

One of the first biosensor-based systems for the measurement of glucose in blood used amperometric detection of H_2O_2 as the measurement principle. Dependence of the measured glucose value on O_2 concentration in the sample was a problem because significantly less than the stoichiometric amount of dissolved O_2 is present in blood to support the glucose oxidase reaction and to produce a linear relationship of signal with glucose concentration. This is especially true at high concentrations of glucose found in samples from patients with diabetes (greater than 500 mg/dL; 27.8 mmol/L). So, sample and calibration solutions were diluted at least 1:10 in buffer for equilibrium with atmospheric PO_2 to fix the O_2 concentration in the calibrator and sample to a constant value.

The problem of O_2 limitation for biosensors based on oxidase enzymes has been addressed by designing semipermeable membranes that restrict diffusion of the primary analyte (substrate) to the enzyme layer, which avoids saturation of the enzyme and keeps the ratio of O_2 to analyte always in excess of 1. This extends the linearity of response to analyte concentrations substantially higher than the K_m of the enzyme and reduces the signal dependence on O_2. Outer track–etched polycarbonate membranes are commonly used, as are membranes of PVC, polyurethanes, and silicone emulsions. Another approach has been to use an O_2-rich electrode material as a reservoir of O_2 to support the bioreaction. A fluorocarbon (Kel-F Oil) has been used to formulate a carbon paste electrode to act as a source of O_2 and as the working electrode.

Electron acceptors other than O_2 can serve as mediators in the glucose oxidase reaction and completely eliminate any dependence of the amperometric response on the O_2 concentration of the sample. The mediator, which is usually coimmobilized with the enzyme, transports electrons to the anode surface, where it is reoxidized, resulting in a cyclic reaction mechanism (Fig. 10.13). Mediators with electron transfer kinetics (little or no overpotential) more favorable than that of O_2 allow operation of the sensor at lower applied potentials (+0.2 V vs. Ag/AgCl or lower) than those that are typically used for the oxidation of H_2O_2. This approach not only eliminates dependency of the reaction rate on O_2, the lower applied potential serves to reduce the contribution from oxidizable interfering substances (e.g., uric acid, ascorbic acid, acetaminophen) to the sensor response. Examples of mediators that have been used include quinones and conductive organic salts, such as tetrathiafulvalene-tetracyanoquinodimethane. Ferricyanide and ferrocene derivatives have also been used, including early commercial applications for home blood glucose monitoring. Dimethylferrocene is impregnated into a graphite electrode to which glucose oxidase has been immobilized. Reduced glucose oxidase from the enzymatic reaction is reoxidized by the electrochemically generated ferricinium ion. Current produced during this cycling mechanism is proportional to the concentration of glucose in the blood sample.

Another technique used to decrease interferences from easily oxidized species in a blood sample when traditional H_2O_2 electrochemical detection is used is to employ selectively permeable membranes in proximity to the electrode surface that allow transport of H_2O_2 to the electrode surface but reject the interfering substances based on size exclusion (see Fig. 10.12B). An example is as simple as a

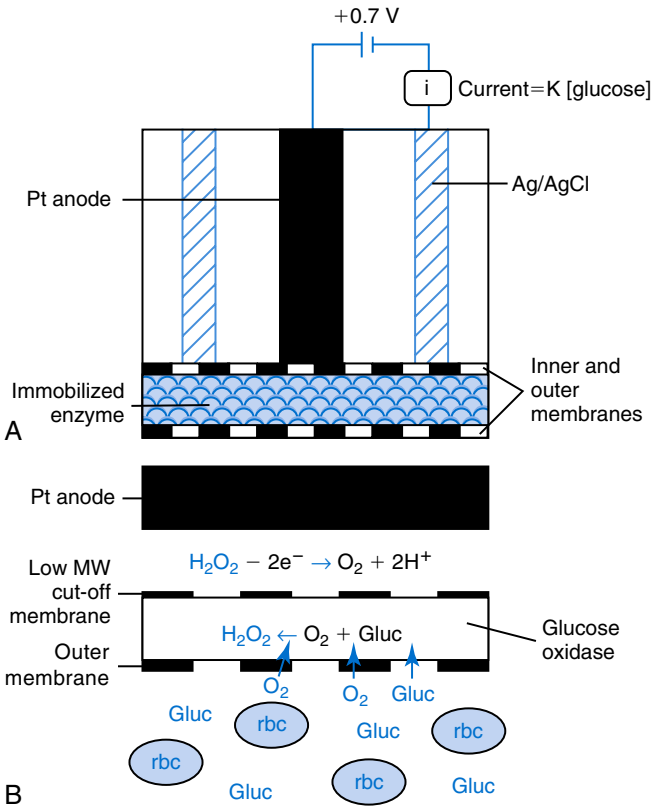


Fig. 10.12 (A) Design of an amperometric enzyme-based biosensor based on anodic detection of hydrogen peroxide *(H_2O_2)* generated from an oxidase enzymatic reaction (e.g., glucose oxidase). (B) Expanded view of the sensing surface shows the different membranes and electrochemical processes that yield the anodic current proportional to the substrate concentration in the sample. *Ag/AgCl*, Silver–silver chloride; *K*, proportionality constant; *MW*, molecular weight; *Pt*, platinum; *rbc*, red blood cells. (From Meyerhoff, M. [1990]. New in vitro analytical approaches for clinical chemistry measurements in critical care. *Clinical Chemistry, 36*, 1570.)

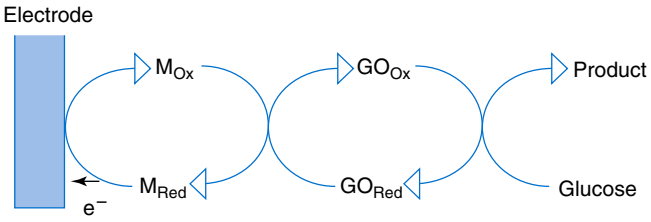


Fig. 10.13 The use of an electroactive mediator in the design of an amperometric enzyme-based biosensor. The mediator accepts electrons directly from the enzyme and is oxidized at the surface of the working electrode, creating more oxidized mediator to continue this process. *Ox*, Oxidant; *Red*, reductant. (From D'Orazio, P. [2002]. In K. Lewendrowski [Ed.]. *Clinical chemistry: laboratory management and clinical correlations* [p. 464]. Philadelphia: Lippincott, Williams & Wilkins.)

low-molecular-weight cutoff membrane, such as cellulose acetate, which is used in many commercial amperometric biosensors. Electropolymerized films, such as poly(phenylenediamine) formed in situ, are also used, to reject interfering substances based on size. Another approach used in a commercial application involves using a second correcting electrode, identical to the working electrode but without enzyme that is sensitive only to the presence of oxidizable interfering substances. The resulting differential signal is proportional to the concentration of analyte.

A novel approach used for elimination of electroactive interfering substances in a commercially available glucose sensor is to directly "wire" the redox center of the enzyme glucose oxidase to a metallic, amperometric electrode using an osmium (III/IV)-based redox hydrogel. Osmium sites effectively serve as mediators and can accept electrons directly from the entrapped enzyme, without need for O_2. This approach allows the operating potential of the electrode to be dramatically lowered to +0.2 V versus the saturated calomel reference electrode, where currents resulting from electrooxidation of ascorbate, urate, acetaminophen, and L-cysteine are negligible.

Substitution of other oxoreductase enzymes for glucose oxidase allows amperometric biosensors for other substrates of clinical interest to be constructed. For example, sensors have been developed to measure (1) blood lactate, (2) cholesterol, (3) pyruvate, (4) alanine, (5) glutamate, and (6) glutamine. In addition, by using a multiple enzyme cascade including an oxoreductase enzyme, an amperometric biosensor for creatinine has been developed.

Enzyme-Based Biosensors With Potentiometric and Conductometric Detection

ISEs can be used as transducers in potentiometric biosensors. An example is a biosensor for urea (blood urea nitrogen [BUN]) based on a polymer membrane ISE for NH_4^+ ion (Fig. 10.14). The enzyme urease is immobilized at the surface of the NH_4^+-selective ISE based on the antibiotic nonactin (see structure of ionophore in Fig. 10.3) and catalyzes the hydrolysis of urea to NH_3 and CO_2:

$$\text{Urea} \xrightarrow{\text{urease}} 2NH_3 + CO_2 \qquad \textbf{(10.28)}$$

The ammonia (NH_3) produced forms NH_4^+, which is sensed by the ISE. The signal generated by the NH_4^+ produced is proportional to the logarithm of the concentration of urea in the sample. The response may be steady state or transient. Typically, correction for background K^+ is required because the nonactin ionophore has limited selectivity for NH_4^+ over K^+ ($K_{NH_4/K} = 0.1$). Potassium is measured simultaneously with urea and is used to correct the output of the urea sensor using Eq. 10.9.

The approach already described for measurement of urea using an enzyme-based potentiometric biosensor assumes that the turnover of urea to NH_4^+ at steady state provides a constant ratio of NH_4^+ to urea, independent of concentration. This is rarely the case, especially at higher substrate concentrations, which results in a nonlinear sensor response.

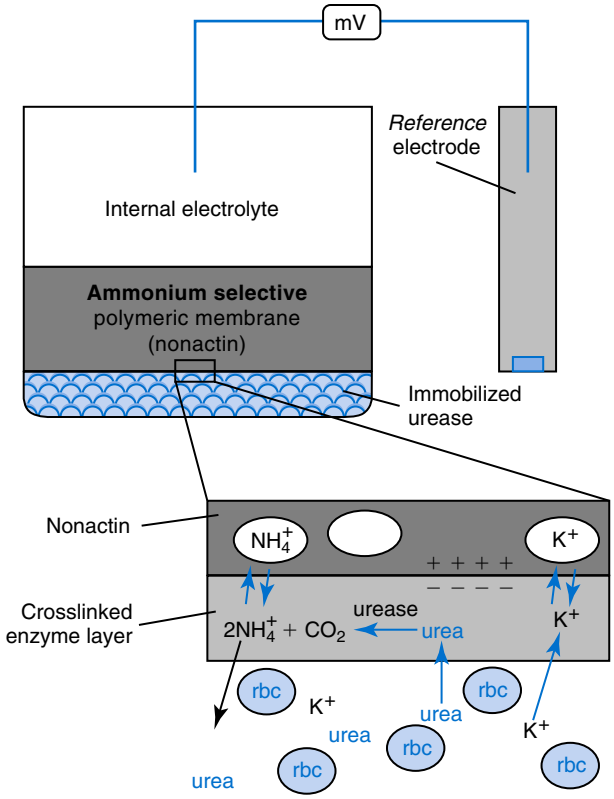


Fig. 10.14 Potentiometric enzyme-based biosensor for determination of blood urea, based on urease enzyme immobilized on the surface of an ammonium ion–selective polymeric membrane electrode. *CO_2*, Carbon dioxide; *K^+*, potassium; *NH_4^+*, ammonium; *rbc*, red blood cells.

The linearity of the sensor is also limited by the fact that hydrolysis of urea produces a local alkaline pH in the vicinity of the NH_4^+-sensing membrane, partially converting NH_4^+ to NH_3 (pK_a = 9.3). Ammonia is not sensed by the ISE. The degree of nonlinearity may be reduced by the placement of a semipermeable membrane between the enzyme and the sample to restrict diffusion of urea to the immobilized enzyme layer.

A change in solution conductivity has also been used as a transduction mechanism in enzyme-based biosensors. Examples include the measurement of glucose, creatinine, and acetaminophen using interdigitated electrodes. There are few practical applications of conductometric biosensors because of the variable ionic background of clinical samples and the requirement to measure small conductivity changes in media of high ionic strength. A commercial system for the measurement of urea in serum, plasma, and urine is based on the enzyme urease. Dissolution of products from Eq. 10.28 to NH_4^+ and HCO_3^- produces a change in sample conductivity. The initial rate of change in conductivity is measured to compensate for the background conductivity of the sample. This approach is limited to the measurement of analytes at relatively high concentrations because of small changes in conductivity produced by low concentrations of analyte.

Enzyme-Based Biosensors With Optical Detection

Optical sensors with immobilized enzymes and indicator dyes have been developed for the measurement of glucose

and other substrates of clinical interest. These biosensors are based on optical detection chemistries for pH and O_2, described earlier in this chapter, and rely on absorbance and/or reflectance, fluorescence, and luminescence as modes of detection. Enzyme immobilization methods resemble those used to construct electrochemical biosensors, including physical entrapment or encapsulation in a gel matrix, physical adsorption onto substrates, and covalent binding or absorption on an insoluble support. Using an example based on an optode for PO_2, a sensitive indicator is coimmobilized with an oxidase enzyme at the end of a fiberoptic probe. The probe is used to monitor fluorescence of the indicator. Quenching of fluorescence of the indicator by O_2 is followed. A decrease in PO_2 resulting from a reaction catalyzed by the enzyme will result in less quenching of the indicator and a fluorescent signal directly proportional to the concentration of the substrate. In an example of an optical biosensor probe for glucose, an O_2-sensitive cationic dye, $Ru(phen)_3^{2+}$, is immobilized, along with glucose oxidase, on the surface of an optical fiber. A decrease in PO_2 arising from the enzyme-catalyzed oxidation of glucose results in an increase in luminescence intensity of the ruthenium tris(phenanthrene).

Similar optical biosensors have been prepared for many other analytes. For example, a cholesterol optical biosensor has been devised based on fluorescence quenching of an O_2-sensitive dye that is coupled to the consumption of O_2 resulting from the enzyme-catalyzed oxidation of cholesterol by the enzyme cholesterol oxidase. Serum bilirubin has been detected using bilirubin oxidase, coimmobilized with a ruthenium dye, on an optical fiber. The bilirubin sensor was reported to exhibit a lower detection limit of 0.6 mg/dL (10 μmol/L), a linear range up to 1.8 mg/dL (30 mmol/L), and a typical reproducibility of 3% (coefficient of variation), adequate for clinical application.

The pH change resulting from enzyme-catalyzed reactions has also been measured optically. The indicator dye fluorescein is often used as a pH-sensitive indicator to construct such sensors. The protonated form of fluorescein does not fluoresce, but the conjugate base strongly fluoresces at 530 nm, when excited at 490 nm. Using glucose oxidase as the enzyme, a pH optode has been used to follow the formation of gluconic acid. A disadvantage of optical sensors based on pH changes is that they are strongly dependent on the pH and buffer capacity of the sample. Moreover, the working range of the sensor is determined by the pK_a of the indicator, 6.8 to 7.2 for fluorescein, depending on ionic strength of the sample matrix. A pH-sensitive indicator may also be used to follow enzymatic reactions that produce ammonia (e.g., urease action on urea).

Affinity Sensors

Affinity sensors are a special class of biosensors in which the immobilized biologic recognition element is a binding protein, antibody, or oligonucleotide with high binding affinity and high specificity toward a clinically important analyte. Such sensors are being developed as alternatives to conventional binding assays to enhance the speed and convenience of a wide range of assays that normally would be run on large, sophisticated instruments in a central laboratory. Affinity sensors may be more easily adapted than traditional assays to systems developed for point-of-care testing for infectious disease, cardiac markers, or other cases where speed and ease of use are required. Ideally, direct binding of the immobilized species with its target in a clinical sample should yield a sensor signal proportional to the concentration of the analyte. However, in practice, direct sensing of the binding event at analyte concentrations that would cover the full range for clinical application is difficult to achieve. Furthermore, high affinity of such binding reactions, which is required to achieve optimal sensitivity, limits the reversibility of such devices. Unlike ISEs, O_2 sensors and many of the enzyme-based biosensors described previously, affinity sensors based on electrochemical, optical, or other transduction modes are typically single-use devices. For repeated multiuse applications, some type of regeneration step (pH change, temperature change, and so on) to dissociate the tight binding between the recognition element and the target is required.

Most affinity-type sensors appearing in the literature with promise of clinical usefulness are based on reagents labeled with enzymes, fluorophores, and electrochemical tags, and hence they function more like traditional binding and/or immunoassays, except that a recognition element is immobilized on the surface of a suitable electrode or another type of transducer. For example, electrochemical O_2 sensors have been used to carry out heterogeneous enzyme immunoassays (sandwich or competitive type), using catalase as a labeling enzyme (catalyzes $H_2O_2 \rightarrow 2H^+$ and O_2) and immobilizing capture antibodies on the outer surface of the gas-permeable membrane. After binding equilibration and washing steps, the amount of bound enzyme is detected by adding substrate and following the increase in current generation caused by local production of O_2 near the surface of the sensor.

The number of research reports related to affinity biosensors continues to increase, with potential application for detection of markers of cardiac disease, cancer, and autoimmune disease. Commercialization has lagged behind basic research, and movement of affinity biosensors from the research laboratory to the clinical laboratory has been slow. Some commercial examples of immunosensors do exist, primarily in the unit-use, disposable format designed for point-of-care testing. One successful commercial example is a cartridge-based device for cardiac troponin I on the i-STAT handheld analyzer (Abbott Point of Care, Abbott Laboratories). This sensor uses a sandwich immunoassay format with electrochemical detection. A capture antibody is immobilized on a gold electrode for recognition and capture of troponin in the blood sample. A second reporter antibody is labeled with alkaline phosphatase and binds with surface-captured troponin antigen. Following incubation and washing steps, the substrate p-aminophenyl phosphate is introduced, and the product of the enzymatic reaction (p-aminophenol) is detected amperometrically by oxidation at the gold electrode. Magnitude of the oxidative current is proportional to concentration of troponin in the sample.

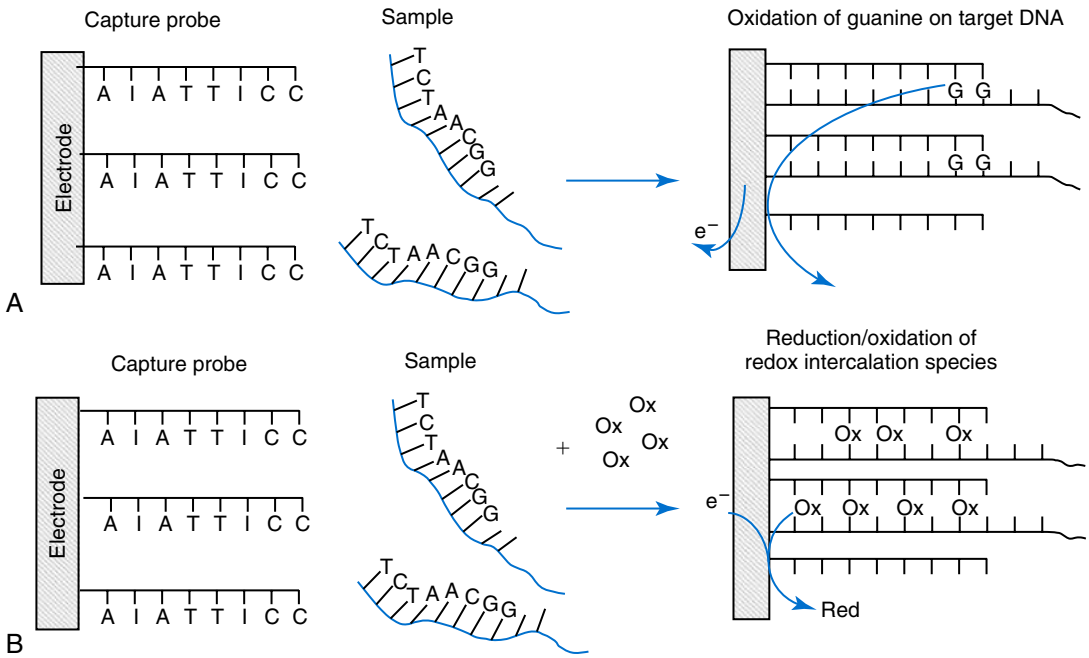


Fig. 10.15 Examples of DNA biosensor configurations. (A) Direct electrooxidation (Ox) detection of guanosine bases in target DNA after hybridization with immobilized capture probe on electrode surface. (B) Electrochemical detection of hybridization using exogenous redox (Ox) species that intercalate into hybridized complex between the immobilized capture DNA probe and target DNA.

Affinity Sensors for DNA Analysis Using Fluorescent Labels

Affinity-type sensors based on oligonucleotide binding represent perhaps the most promising application of affinity sensors to clinical chemistry. Fluorescent labels were the first to be used in an early commercial biosensor for DNA analysis, introduced in 1996 (GeneChip, Affymetrix, Santa Clara, CA). High-density oligonucleotide arrays are formed on glass substrates. Target DNA is isolated and amplified using the polymerase chain reaction (PCR) while a fluorescent label is incorporated. The sample is incubated on the array, and detection of hybridization of labeled DNA to its complementary strands takes place by monitoring light emitted by the fluorescent labels. The primary application of the device has been for DNA sequencing, so the focus has been primarily on increasing the density of the DNA arrays. Applications of the device continue to expand, with a variety of arrays developed for gene expression and whole genomic analysis related to (1) cancer, (2) inflammatory disease, (3) infectious disease, and (4) diabetes. Another commercial example of affinity sensors for DNA analysis using fluorescent labels is the BeadArray technology (Illumina, San Diego, CA). Silica microbeads with covalently attached, single-stranded DNA probes are randomly assembled on a fiberoptic bundle or planar silica surface. Each bead carries hundreds of thousands of copies of a gene-specific probe sequence. Due to the high-density packing of beads, with a possibility of approximately 50,000 beads on a 1.4-mm diameter array, a large number of probe sequences may be investigated in parallel, and multiple copies of each probe sequence are possible, increasing measurement precision. Fluorescently labeled target sequences in the sample to be analyzed hybridize with their complementary probe sequences. Arrays are scanned in a BeadArray reader for image analysis and data extraction. This versatile platform has been adapted for various genetic analyses, such as single nucleotide polymorphism genotyping, DNA methylation detection, and gene expression profiling.

Affinity Sensors for DNA Analysis Using Electrochemical Labels

DNA sensors with immobilized complementary DNA on a suitable electrochemical sensor can operate in direct (electrochemical oxidation of guanine in target DNA; Fig. 10.15A) or indirect (exogenous electrochemical markers and/or labels; Fig. 10.15B) transduction modes. Although most of the proposed electrochemical DNA biosensors require amplification methods, such as PCR, to multiply small amounts of DNA into measurable quantities, some are sensitive enough to eliminate the need for target amplification. Nanotechnology has been proposed, in an indirect format, for signal amplification. For example, a capture probe DNA is immobilized on an Au electrode. Reporter probes with electrostatically bound ruthenium complexes $(Ru[NH_3]_6)^{3+}$ are loaded onto Au nanoparticles (AuNP) and are capable of hybridizing with one of two sequences on target DNA. The other sequence on the target DNA is capable of hybridizing with the immobilized capture probe. Hybridization events on the electrode surface bring multiple reporter probes for each AuNP. Electroactive $(Ru[NH_3]_6)^{3+}$ is reduced at the electrode surface, and the coulometric signal is proportional to the concentration of target DNA. A commercial example of electrochemical DNA sensing, along with AuNP probes without

need for PCR amplification, is the Verigene system (Luminex, Austin, TX), which is capable of detecting single-nucleotide polymorphisms related to common genetic disorders, such as (1) thrombophilia, (2) alterations of folate metabolism, (3) cystic fibrosis, and (4) hemochromatosis. Another commercially available, electrochemically based platform uses a sensor array chip consisting of 16 nanoscale Au electrodes, modified with thiol self-assembled monolayers, optimized for biomolecule immobilization. Horseradish peroxidase is the preferred target label because of its fast electron transfer kinetics, when used with amperometric detection.

Another example of an electrochemical "gene" sensor array uses electrochemical probes that are selectively inserted into hybridized DNA duplexes. In one approach, after the immobilized capture of oligo anchored to the electrode surface is allowed to bind the target sequence, hybridization is detected by exposing the surface of the electrode to an exogenous electroactive species (e.g., Co[III]tris-phenanthroline or ruthenium complexes) that intercalates within the duplex but not to single-stranded DNA. After unbound electroactive species are removed by washing, the presence of hybridization is readily detected by voltammetry, scanning the potential of the underlying electrode to oxidize or reduce any intercalated electroactive species, with the current detected being proportional to the number of duplex DNA species on the surface of the electrode.

POINTS TO REMEMBER

Affinity Biosensors

- Affinity biosensors are a specific type of biosensor in which an immobilized recognition element exhibits a high binding constant for the analyte of interest.
- Examples of recognition elements are antibodies and oligonucleotides (DNA segments).
- Affinity biosensor binding reactions are typically irreversible and best suited to single-use sensors.
- Most affinity biosensors for clinical application use indirect modes of transduction, based on reagent label such as enzymes, fluorescent, and electrochemical labels.

REVIEW QUESTIONS

1. Ion-selective electrodes measure electrical potential difference across a membrane using the principles of:
   a. coulometry.
   b. potentiometry.
   c. amperometry.
   d. conductivity.
2. An optical sensor used in analytical instruments to measure pH, blood gases, and electrolytes is referred to as a(n):
   a. potientiometer.
   b. optode.
   c. affinity-type sensor.
   d. coulometer.
3. Voltammetry/amperometry measurements are based on which of the following electrochemical cell types?
   a. Ion-selective electrode cells
   b. Galvanic electrochemical cells
   c. Electromotive force cells
   d. Electrolytic electrochemical cells
4. An affinity-based biosensor is one in which:
   a. the biologic element recognizes the analyte on the basis of its binding to an immobilized substance.
   b. the fluorescence of a solution is coupled to an enzyme-catalyzed reaction.
   c. the biologic element reacts with an ion-selective electrode that is used as a potentiometric transducer.
   d. an electron acceptor is used to form a colored reaction product.
5. A voltmeter that measures the potential across an electrochemical cell (between the two electrodes) is referred to as a:
   a. conductometer.
   b. amperometer.
   c. direct-reading potentiometer.
   d. redox meter.
6. Membrane potentials caused by the permeability of certain types of membranes to selected anions or cations are measured by which type of electrode?
   a. Ion-selective electrode
   b. Saturated calomel electrode
   c. Platinum electrode
   d. Gas-sensing electrode
7. The maximum difference in potential between the two electrodes in an electrochemical cell obtained with the current at zero is the definition of:
   a. electrical potential.
   b. electromotive force.
   c. electrochemical gradient.
   d. Nernst equation.
8. In this type of measurement, a constant current is applied to a solution, and a substance in that solution is changed to a different state through oxidation or reduction. The time it takes for this to occur is measured. This electrochemical technique is called:
   a. amperometry.
   b. voltammetry.
   c. potentiometry.
   d. coulometry.

9. The measurement of ions by direct potentiometry provides a value for the concentration of free, unbound ion in solution that is called the ionic:
   a. force.
   b. potential.
   c. activity.
   d. action.

10. The predominant class of potentiometric electrodes used in the clinical laboratory is the:
   a. gas-sensing electrode.
   b. glass membrane electrode.
   c. coulometric electrode.
   d. polymer membrane ion-selective electrode.

SUGGESTED READINGS

Apple, F. S., Koch, D. D., Graves, S., et al. (1982). Relationship between direct potentiometric and flame photometric measurement of sodium in blood. *Clinical Chemistry*, *28*, 1931–1935.

Astrup, P., & Severinghaus, J. W. (1986). *The history of blood gases, acids and bases*. Copenhagen, Denmark: Munksgaard.

Bakker, E., Bühlmann, P., & Pretsch, E. (1997). Carrier-based ion-selective electrodes and bulk optodes. 1. General characteristics. *Chemical Reviews*, *97*, 3083–3132.

Bates, R. G. (1973). *Determination of pH: theory and practice*. New York: John Wiley & Sons.

Buck, R. P., & Lindner, E. (1994). Recommendations for nomenclature of ion-selective electrodes. *Pure Appl Chem*, *66*, 2527–2536.

Clark, L. C., Jr., & Lyons, C. (1962). Electrode systems for continuous monitoring in cardiovascular surgery. *Annals of the New York Academy of Sciences*, *102*, 29–45.

D'Orazio, P. (2003). Biosensors in clinical chemistry. *Clinica Chimica Acta*, *334*, 41–69.

D'Orazio, P. (2011). Biosensors in clinical chemistry—2011 update. *Clinica Chimica Acta*, *412*, 1749–1761.

Faulkner, L. R. (1983). Understanding electrochemistry: some distinctive concepts. *Journal of Chemical Ed*, *60*, 262–264.

Guilbault, G. G. (1984). *Handbook of immobilized enzymes*. New York: Marcel-Dekker.

Heller, A. (1996). Amperometric biosensors. *Current Opinion in Biotechnology*, *7*, 50–54.

Leiner, M. J. P. (1991). Luminescence chemical sensors for biomedical applications: scope and limitations. *Analytica Chimica Acta*, *255*, 209–222.

Luppa, P. B., Sokoll, L. J., & Chan, D. W. (2001). Immunosensors—principles and applications to clinical chemistry. *Clinica Chimica Acta*, *314*, 1–26.

Oesch, U., Ammann, D., & Simon, W. (1986). Ion-selective membrane electrodes for clinical use. *Clinical Chemistry*, *32*, 1448–1459.

Osswald, H. F., & Wuhrmann, H. R. (1981). Calibration standards for multi ion analysis in whole blood samples. In D. W. Lubbers, H. Acker, R. P. Buck, et al. (Eds.), *Progress in enzyme and ion-selective electrodes* (pp. 74–78). Berlin: Springer-Verlag.

Pioda, L. A., Simon, W., Bosshard, H. R., et al. (1970). Determination of potassium ion concentration in serum using a highly selective liquid-membrane electrode. *Clinica Chimica Acta*, *29*, 289–293.

Ronkainen, N. J., & Okon, S. L. (2014). Nanomaterial-based electrochemical immunosensors for clinically significant biomarkers. *Materials*, *7*, 4669–4709.

Seitz, W. R. (1984). Chemical sensors based on fiber optics. *Analytical Chemistry*, *56*, 16A–34A.

Stott, R. A. W., Hortin, G. L., Wilhite, T. R., et al. (1995). Analytical artifacts in hematocrit measurements by whole blood chemistry analyzers. *Clinical Chemistry*, *41*, 306–311.

Thevenot, D. R., Toth, K., Durst, R. A., et al. (2001). Electrochemical biosensors: recommended definitions and classifications. *Biosensors & Bioelectronics*, *16*, 121–131.

11

Electrophoresis

Lindsay A.L. Bazydlo, James P. Landers

OBJECTIVES

1. Define the following:
   - Agarose
   - Ampholyte
   - Amphoteric
   - Blot/blotting
   - Densitometry
   - Electroendosmosis
   - Electrophoresis
   - Electrophoretic mobility
   - Polyacrylamide
   - Wick flow
2. State the theory of electrophoresis.
3. List and describe the components of an electrophoresis system, including buffers, support media, separation techniques, stains, and detection methods.
4. Identify the individual protein fractions of serum from an electropherogram.
5. Quantify the individual fractions of a scanned gel when given the total protein value.
6. Identify how each of the following affects electrophoresis: size and shape of molecule, heat and current produced by the power supply, buffer pH, and type of gel used.
7. Describe and compare the following types of electrophoretic techniques, including clinical utility, system components, and detection methods: (1) capillary, (2) isoelectric focusing, (3) microchip, (4) two-dimensional, and (5) zone.
8. Compare and contrast Southern, Northern, and Western blotting, including analyte, electrophoresis technique, and clinical utility.
9. State three advantages of capillary electrophoresis over conventional electrophoresis.
10. Explain how the following affect the performance of electrophoresis: endosmosis, buffer and stain integrity, sample type or appearance, and over-application or under-application of the sample.

KEY WORDS AND DEFINITIONS

Ampholyte A molecule that is positively or negatively charged on the basis of the pH of the solution in which it resides; proteins, because they contain many ionizable amino and carboxyl groups, behave as ampholytes in solution and are considered amphoteric.

Capillary electrophoresis A method in which the classic techniques of electrophoresis are carried out in a small-bore, fused silica capillary tube coated with a polymeric covering.

Densitometry A measuring technique that uses an optical system to scan and quantify electrophoretic fractions separated on a gel or other medium.

Electropherogram A densitometric display of protein zones on a support material after separation and staining.

Electrophoresis The migration of charged solutes or particles within a liquid medium under the influence of an electrical field.

Electrophoretic mobility (μ) The rate of migration (cm/s) of a charged solute in an electrical field, expressed per unit field strength (volts/cm).

Endosmosis (electroendosmotic flow) Preferential movement of water in one direction through an electrophoresis medium due to selective binding of one type of charge on the surface of the medium.

Isoelectric focusing An electrophoretic technique that separates amphoteric compounds within a medium that possesses a stable pH gradient.

Isoelectric point (pI) The pH at which a molecule has no net charge and will not migrate during electrophoresis.

Wick flow Movement of water from the buffer reservoirs toward the center of an electrophoresis gel or strip to replace water lost by evaporation.

Zone electrophoresis Migration of charged molecules in an applied electric field.

BASIC CONCEPTS

The first **electrophoresis** method used to study proteins was the free solution or moving boundary method devised by Tiseus in 1937. This technique was used in research to measure electrophoretic mobility and to study protein-protein interactions. It was able to resolve the serum proteins into only four component mixtures, with the α_1 fraction incompletely separated from albumin.

Zone electrophoresis can be performed in a porous supporting medium, such as agarose gel film, such that each protein zone is sharply separated from neighboring zones by a protein-free area. Zones are visualized by staining with a protein-specific stain to produce an **electropherogram**, which is then scanned and quantified using a densitometer. The support medium can also be handled after drying and kept as a permanent record. This is the most commonly applied technique in clinical chemistry and is used to separate proteins in serum, urine, cerebrospinal fluid (CSF), other physiologic fluids, erythrocytes and tissue, and nucleic acids in various tissue cells.

Although electrophoretic separation of biologically relevant macromolecules in gels (or paper) has been the workhorse of modern biomedical research, the advent of **capillary electrophoresis** (CE) has revolutionized separations. Intense interest in carrying out electrophoretic separation in capillaries with inner diameters ranging from 20 to 75 μm has resulted from its unprecedented resolving power, separation speed, and small sample analysis capabilities. However, the true significance of CE to the separations community can be seen in its ability to apply these separation principles in a multimodal approach to a variety of analytes that obviously included proteins and polynucleic acids, but also peptides, small drug-like molecules, and even ions.

POINTS TO REMEMBER

Electrophoresis

- Refers to the migration of ions in an electrical field
- Separation occurs based on the inherent electrophoretic mobility of an analyte
- Electrophoretic mobility is directly proportional to the net charge and inversely proportional to the size of the molecule and the viscosity of the electrophoresis medium

THEORY OF ELECTROPHORESIS

Depending on the charge they carry, ionized solutes move toward either the cathode (negative electrode) or the anode (positive electrode) in an electrophoresis system. For example, positive ions (cations) migrate to the cathode, and negative ions (anions) to the anode. An **ampholyte** (a molecule that is either positively or negatively charged, formerly called a zwitterion) becomes positively charged in a solution that is more acidic than its isoelectric point (pI)[a] and migrates toward the cathode. In a more alkaline solution, the ampholyte becomes negatively charged and migrates toward the anode. Because proteins contain many ionizable amino (—NH2) and carboxyl (—COOH) groups, and because the bases in nucleic acids may also be positively or negatively charged, they both behave as ampholytes in solution.

The rate of migration of ions in an electrical field depends on the factors listed in Box 11.1. The equation expressing the driving force in such a system is given by

$$F = (X)(Q) = \frac{(\text{EMF})(Q)}{d} \quad \textbf{(11.1)}$$

where F = the force exerted on an ion; X = the current field strength (V/cm) (i.e., voltage drop per unit length of medium); Q = the net charge on the ion; EMF = the electromotive force [voltage (V) applied]; and d = the length of the electrophoretic medium (cm).

Steady acceleration of the migrating ion is counteracted by a resisting force characteristic of the solution in which migration occurs. This force, expressed by Stokes' law, is

$$F' = 6\pi r\eta v \quad \textbf{(11.2)}$$

where F' = the counter force; π = 3.1416; r = the ionic radius of the solute; η = the viscosity of the buffer solution in which migration is occurring; and v = the rate of migration of the solute = velocity (length *[l]* traveled per unit of time [cm/s]).

The force F' counteracts the acceleration that would be produced by F if no counter force were present, and the result of the two forces is a constant velocity. Therefore, when

$$F = F' \quad \textbf{(11.3)}$$

then

$$6\pi r\eta = (X)(Q) \quad \textbf{(11.4)}$$

or

$$v/X = 1 \times d/t \times E = Q/6\pi r\eta = \mu \quad \textbf{(11.5)}$$

where v/X is the rate of migration (cm/s) per unit field strength (E/cm), defined as the **electrophoretic mobility** and expressed by the symbol μ.

Electrophoretic mobility is directly proportional to the net charge and is inversely proportional to the size of the molecule and the viscosity of the electrophoresis medium. Mobility may be positive or negative, depending on whether a protein migrates in the same or the opposite direction as the electrophoretic field (defined as extending from the anode to the cathode).

In addition to the factors listed in Box 11.1, other factors that affect electrophoretic mobility include electroendosmosis (endosmosis) and wick flow. Electroendosmosis affects mobility by causing uneven movement of water through the support medium. An electrophoretic support medium, such as a gel in contact with water, takes on a negative charge caused by the adsorption of hydroxyl ions. These ions are fixed to the surface and are immobile. Positive ions in solution cluster about the fixed negative charge sites, forming an ionic cloud of mostly positive ions. The number of negative ions in the solution increases with increasing distance from

[a] The isoelectric point of a molecule is the pH at which it has no net charge and will not move in an electrical field.

BOX 11.1 Factors Affecting the Motility of Ions in an Electrophoretic System

Net charge of the molecules
Size and shape of the molecules
Strength of the electrical field
Support medium properties
Ionic strength of the buffer
Temperature

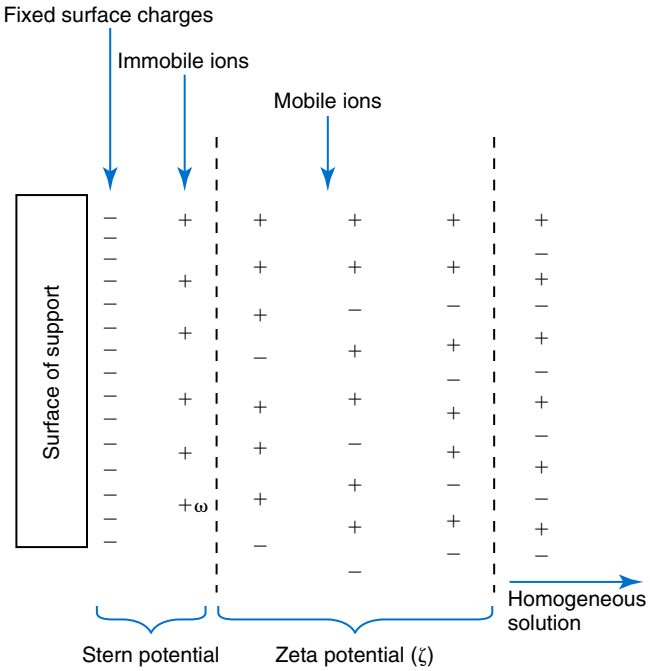


Fig. 11.1 Distribution of + and – ions around the surface of an electrophoretic support. Fixed on the surface of the solid is a layer of – ions. A second layer of + ions is attracted to the surface. These two layers compose the Stern potential. The large, diffuse layer containing mostly + ions is the electrokinetic or zeta potential. Extending farther from the surface of the solid is homogeneous solution. The Stern potential plus the zeta potential equals the electrochemical potential, or epsilon potential.

the fixed negative charge sites until eventually positive and negative ions are present in equal concentrations (Fig. 11.1).

When current is applied to such a system, charges attached to the immobile support remain fixed, but the cloud of ions in solution moves to the electrode of opposite polarity. Because ions in solution are highly hydrated, this results in movement of the solvent as well. Movement of solvent and its solutes relative to the fixed support is referred to as **endosmosis** and causes preferential movement of water in one direction. Macromolecules in solution that move in the direction opposite this flow may remain immobile or may even be swept back toward the opposite pole if they are insufficiently charged. In media in which endosmosis is strong, such as conventional cellulose acetate and unpurified agarose gel, γ-globulins are swept back from the application point. Because the inner surface of a glass capillary contains many such charged groups, endosmosis is very strong and is actually the primary driving force for migration in CE systems. However, the endosmosis in CE can be manipulated to modify the magnitude of the endosmotic effect. In electrophoretic media in which surface charges are minimal (starch gel, purified agarose gel, or polyacrylamide gel), endosmosis is minimal.

Wick flow results from the movement of buffer into the support medium. During electrophoresis, heat evolved because of the passage of current through a resistive medium can cause evaporation of solvent from the electrophoretic support. This drying effect draws buffer into the support, and, if significant, the flow of buffer can affect protein migration and hence the calculated mobility.

CLINICAL ELECTROPHORESIS

In this section, focus will be on the electrophoresis methodology that is frequently used in the clinical laboratory. Refer to Chapter 18 for a thorough discussion on the various electrophoretic fractions present in clinical protein electrophoresis.

Slab Gel Electrophoresis

Traditional methods, using a rectangular gel regardless of thickness, are referred to collectively by the term *slab gel electrophoresis*. Its main advantage is its ability to simultaneously separate several samples in one run. Starch, agarose, and polyacrylamide media have all been used in this format. It is the primary method used in clinical chemistry laboratories for the separation of various classes of serum or CSF proteins and DNA and RNA fragments. Gels (usually agarose) may be cast on a sheet of plastic backing or completely encased within a plastic-walled cell, which allows horizontal or vertical electrophoresis and submersion for cooling, if necessary.

General Operations

General operations performed in conventional electrophoresis include separation, detection, and quantification, as well as a number of "blotting" techniques.

Electrophoretic Separation

When electrophoresis is performed on precast microzone agarose gels, the following steps are typical: (1) excess buffer is removed from the support surface by blotting, taking care that bubbles are not present; (2) 5 to 7 μL of sample is applied using a comb or a plastic template and is allowed to diffuse into the gel; it is then blotted to remove the excess; (3) the gel is placed into the electrode chamber; (4) electrophoresis is performed at specified current, voltage, or power; (5) the gel is fixed, rinsed, and then dried; (6) the gel is stained and redried; and (7) the gel is scanned in a densitometer. If isoenzymes are to be determined, substrate dye solution is incubated on the gel to stain zones before fixing and drying. Alternative procedures would be required if the more sophisticated methods described later are used.

Detection and Quantification

Once separated, proteins may be detected by staining followed by quantification using a densitometer, or by direct measurement using an optical detection system.

TABLE 11.1 Suggested Wavelengths for Quantitation of Protein Zones by Direct Densitometry

Separation Type	Stain	Nominal Wavelength, nm
Serum proteins in general	Amido Black (Naphthol Blue Black)	640
	Coomassie Brilliant Blue G–250 (Brilliant Blue G)	595
	Coomassie Brilliant Blue R–250 (Brilliant Blue R)	560
	Ponceau S	520
Isoenzymes	Nitrotetrazolium Blue	570
Lipoprotein zones	Fat Red 7B (Sudan Red 7B)	540
	Oil Red O	520
	Sudan Black B	600
DNA fragments	Ethidium bromide (fluorescent)	254 (Ex) 590 (Em)
CSF proteins	Silver nitrate	—

CSF, Cerebrospinal fluid; *Em*, emission; *Ex*, excitation.

Staining. If staining is used to visualize separated proteins, the proteins usually are fixed first by precipitating them in the gel with a chemical agent such as acetic acid or methanol. This prevents diffusion of proteins out of the gel when submersed in the stain solution. The amount of dye taken up by the sample is affected by many factors, such as the type of protein and the degree of denaturation of the proteins by fixing agents.

Table 11.1 lists dyes commonly used in electrophoresis, along with suggested wavelengths for quantification by **densitometry**. Most commercial methods for serum protein electrophoresis use Amido Black B or members of the Coomassie Brilliant Blue series of dyes for staining. Isoenzymes are typically visualized by incubating the gel in contact with a solution of substrate, which is linked structurally or chemically to a dye before fixing. Silver nitrate stains proteins and polypeptides with sensitivity 10- to 100-fold greater than that of conventional dyes. Selective fixing and staining of protein subclasses can also be achieved by combining a stain molecule with an antiglobulin, as is done in IFE.

Improvements in conducting sensitive measurements have been achieved by linking an enzyme, such as alkaline phosphatase or peroxidase, to a single or double antibody specific for particular proteins, such as oligoclonal immunoglobulin (Ig), or by spraying separated proteins with luminal and peroxide to develop chemiluminescence, which, in turn, exposes x-ray film to form a permanent image. Chemiluminescence has been used in this way to quantify IgE, and DNA fragments have been detected by linking with a fluorescent dye label.

In practice, a typical stain solution may be used several times before it is replaced. A good rule of thumb is that a stain solution of 100 mL may be used for a combined total of 387 cm^2 (60 in^2) of agarose film. The stain solution may be considered faulty if leaching of stained protein zones occurs in the 5% acetic acid wash solution. Whenever protein zones appear too lightly stained, the stain or substrate reagent—in the case of isoenzymes—should always be suspected. Stain solution must be stored tightly covered to prevent evaporation.

Quantification. A *densitometer* is used to quantify stained zones. This instrument measures the absorbance of each fraction as the gel (or other medium) is moved past a photometric optical system and displays an electropherogram on a computer display. The software is able to automatically integrate the area under each peak and report each as a percent of total or as absolute concentration or activity computed from the total protein or activity of enzyme in the sample.

Reliable densitometric quantification requires: (1) light of an appropriate wavelength; (2) linear response from the instrument; and (3) a transparent background in the medium being scanned. Linearity may be tested with a neutral density filter designed with separated or adjacent areas of linearly increasing density. The densities are permanent and have expected absorbance values. The very small sample sizes used and the transparency of agarose gels satisfy the requirement for a clear background. Nevertheless, problems can occur with densitometry because of differences in the quantity of stain taken up by individual proteins and differences in protein zone sizes.

Essential features of a densitometer include: (1) the ability to scan gels 25 to 100 mm in length; (2) electronic adjustment of the most intense peak to full scale; (3) automatic background zeroing (peaks are not lost or "cut off"); (4) variable wavelength control over the range of 400 to 700 nm; (5) variable slits to allow adjustment of the beam size; (6) an integrating device with both automatic and manual selection of cut points between peaks; and (7) automatic indexing, a feature that advances the electrophoresis strip from one sample channel to the next.

Desirable features of a densitometer include computerized integration and printout, built-in diagnostics for instrument troubleshooting, the choice of one of several scanning speeds, and the ability to measure in the reflectance mode. Models with a separate computer for data processing permit storage and reformatting of data, if desired, and reprinting or delayed transmission to a host computer.

DNA analysis requires the ability to scan larger gels, which may contain several dozen bands of DNA fragments of different length. Modern automated electrophoresis systems also use larger gels containing 30 or more samples, which are scanned on a new generation of densitometers referred to as *flatbed scanners* or *digital image analyzers.* These instruments are capable of scanning and storing digitized light intensity readings from large areas and use ultrasensitive charge-coupled device detectors having a resolution of up to 1200 dots per inch (21 μm). Sophisticated data processing software permits manipulation of stored image information to produce conventional scans and computations or more complex outputs, such as overlaying and subtraction of patterns from two different samples.

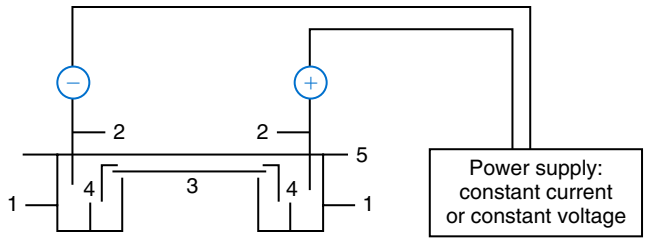


Fig. 11.2 A schematic diagram of a typical electrophoresis apparatus showing two buffer boxes with: *(1)* baffle plates, *(2)* electrodes, *(3)* electrophoretic support, *(4)* wicks, *(5)* cover, and power supply.

Blotting Techniques

In 1975, Edward Southern developed a technique that is widely used to detect fragments of DNA. This technique, known as *Southern blotting*, first requires electrophoretic separation of DNA or DNA fragments by agarose gel electrophoresis (AGE). Next, a strip of nitrocellulose or a nylon membrane is laid over the agarose gel, and the DNA fragments are transferred or "blotted" onto it by capillary, electro, or vacuum blotting. They are then detected and identified by hybridization with a labeled, complementary nucleic acid probe. This technique is widely used in molecular biology for identifying a particular DNA sequence; determining the presence, position, and number of copies of a gene in a genome; and typing DNA.

Northern and *Western* blotting techniques, named by analogy to Southern blotting, were subsequently developed to separate and detect ribonucleic acids (RNAs) and proteins, respectively. Northern blotting is carried out identically to Southern blotting except that RNA species are separated by electrophoresis. Western blotting is used to separate, detect, and identify one or more proteins in a complex mixture. It involves first separating the individual proteins by polyacrylamide gel and then transferring or "blotting" onto an overlying strip of nitrocellulose or a nylon membrane by electro-blotting. The strip or membrane is then reacted with a reagent that contains an antibody raised against the protein of interest.

Instrumentation

Although modern electrophoresis equipment and systems vary considerably in form and degree of automation, the essential components common to all systems (Fig. 11.2) include: (1) two reservoirs that contain the buffer used in the process, (2) a means of delivering current from a power supply via platinum or carbon electrodes contacting the buffer, and (3) a support medium in which separation takes place connecting the two reservoirs. In some systems, wicks may connect the medium to the buffer solution or directly to the electrodes. The entire apparatus is enclosed to minimize evaporation and protect both the system and the operator. The direct current power supply sets the polarity of the electrodes and delivers current to the medium.

Power Supplies

The power supply drives the movement of ionic species in the medium and allows adjustment and control of the current or the voltage. With more sophisticated units, the power may be controlled as well, and conditions may be programmed to change during electrophoresis. Capillary systems use power supplies capable of providing voltages in the kilovolt range.

Current flowing through a medium that has resistance produces heat:

$$\text{Heat} = (E)\,(I)\,(t) \qquad \textbf{(11.6)}$$

where E = electromotive force (EMF) in volts (V); I = current in amperes (A); and t = time in seconds (s).

This heat is released into the medium and increases the thermal agitation of all dissolved ions, and therefore the conductance of the system (decreases resistance). With constant-voltage power supplies, the resultant rise in current increases both protein migration and evaporation of water from the medium. Any water loss increases the ion concentration and further decreases the resistance *(R)*. Under these circumstances, the current and therefore the migration rate will progressively increase. To minimize these effects, it is best to use a constant-current power supply. According to Ohm's law,

$$E = (I)\,(R) \qquad \textbf{(11.7)}$$

therefore if R decreases, the applied EMF also decreases, keeping the current constant. This in turn decreases the heat effect and stabilizes the migration rate.

Buffers

Buffer ions have a twofold purpose in electrophoresis: they carry the applied current, and they fix the pH at which electrophoresis is carried out. Thus they determine: (1) the type of electrical charge on the solute, (2) the extent of ionization of the solute, and therefore (3) the electrode toward which the solute will migrate. The buffer's ionic strength determines the thickness of the ionic cloud (buffer and non-buffer ions) surrounding a charged molecule, the rate of its migration, and the sharpness of the electrophoretic zones. With increasing concentration of ions, the ionic cloud increases in size, and the molecule becomes more hindered in its movement.

According to Joule's law, power produced when current flows through a resistive medium is dissipated as heat. This heat increases in direct proportion to the resistance, but also in proportion to the square of the current. The reduction in resistance caused by a high ionic strength buffer therefore leads to increased current and excessive heat. These buffers yield sharper band separations, but the benefits of sharper resolution are diminished by the Joule (heat) effect that leads to denaturation of heat-labile proteins or the degradation of other components.

Ionic strength (also denoted by the symbol μ) is computed according to the following:

$$\mu = 0.5 \sum c_i z_i^2 \qquad \textbf{(11.8)}$$

where c_i = ion concentration in mol/L and z_i = the charge on the ion.

The ionic strength μ of an electrolyte (buffer) composed of monovalent ions is equal to its molarity (mol/L). The ionic

strength of a 1 mol/L electrolyte solution with one monovalent and one divalent ion is 3 mol/L, and for a doubly divalent electrolyte, it is 4 mol/L.

A relatively high ionic strength buffer used in *high-resolution electrophoresis* improves the separation of serum proteins into as many as 13 bands, with two or more bands in the α_1, α_2, and β-globulin regions and one or more additional bands seen in various pathologic conditions. Because of higher conductivity and the associated heat produced, it is necessary to reduce the temperature of the system to 10°C to 14°C. "Submarine" techniques, in which gels are submersed in circulating buffer cooled by an external cooling device, or are supported on an electrophoresis chamber cooled by circulating water or an integral Peltier plate, provide exact temperature control. Effective cooling with less precise temperature control may also be achieved using chambers designed with a sealed compartment of cooled ethylene glycol, which is in contact with the gel during running.

Because buffers used in electrophoresis are good culture media for the growth of microorganisms, they should be refrigerated when not in use. Moreover, a cold buffer is preferred in an electrophoretic run, because it improves resolution and decreases evaporation from the electrophoretic support. Buffer used in a small-volume apparatus should be discarded after each run because of pH changes resulting from the electrolysis of water that accompany electrophoresis. If volumes used are larger than 100 mL, buffer from both reservoirs may be combined, mixed, and reused up to four times.

Support Media

The support medium provides the matrix in which protein separation takes place. Various types of support media have been used in electrophoresis and range from pure buffer solutions in a capillary to insoluble gels (e.g., sheets, slabs, or columns of starch, agarose, or polyacrylamide) or membranes of cellulose acetate. Gels are cast in a solution of the same buffer to be used in the procedure and may electrophorese in a horizontal or vertical direction. In either case, maximum resolution is achieved if the sample is applied in a very fine starting zone. Separation is based on differences in charge-to-mass ratio of the proteins and, depending on the pore size of the medium, possibly molecular size.

Cellulose Acetate. Cellulose acetate, a thermoplastic resin made by treating cellulose with acetic anhydride to acetylate the hydroxyl groups, is primarily of historical interest. When dry, the membranes contain about 80% air space within the interlocking cellulose acetate fibers and are opaque, brittle films. As the film is soaked in buffer, the air spaces fill with liquid, and it becomes pliable. Samples are applied with a twin-wire applicator or the edge of a glass slide. Because of their opacity, stained membranes need to be made transparent (cleared) for densitometry by soaking in 95:5 methanol:glacial acetic acid. Cleared membranes are strong and could be stored as a permanent record, but because of the necessity for presoaking and clearing, cellulose acetate has largely been replaced by agarose in most clinical applications.

Agarose. Agarose is a linear polymer containing alternating D-galactose and 3,6-anhydro-L-galactose monomers. It is the purified, essentially neutral fraction of agar obtained by separating agarose from agaropectin, a more highly charged fraction containing acidic sulfate and carboxylic side groups. Because the pore size in agarose gels is large enough for all proteins to pass through unimpeded, separation is based only on the charge-to-mass ratio of the protein. Advantages of agarose gel include its low affinity for proteins and its native clarity after drying, which permits excellent densitometry. It is essentially free of ionizable groups and so exhibits little endosmosis.

Most routine procedures for AGE are now performed using commercially produced, prepackaged microzone gels, and the sample is applied by means of a comb or a thin plastic template, with small slots corresponding to sample application points. The template is placed on the agarose surface, and 5 to 7 μL samples are placed on each slot. The serum sample is allowed to diffuse into the agarose for 5 minutes, the excess sample is removed by blotting, and the template is removed. AGE separation for most routine serum applications requires an electrophoresis time of 20 to 30 minutes.

Polyacrylamide Gel. Polyacrylamide is a polymeric matrix consisting of linear chains of acrylamide cross-linked with bis-acrylamide. It is thermostable, transparent, strong, and relatively chemically inert, and—depending on concentration—can be made in a wide range of pore sizes. Its average pore size in a typical 7.5% gel is about 5 nm (50 Å), which is large enough to allow most serum proteins to migrate unimpeded. However, proteins with a molecular radius and/or length that exceeds critical limits will be impeded in their migration. Some of these proteins are fibrinogen, β-lipoprotein, α_2-macroglobulin, and γ-globulins; a schematic representation of serum protein electrophoresis by polyacrylamide gel electrophoresis (PAGE) is shown in Fig. 11.3. The separation is based on both charge-to-mass ratio and molecular size (a phenomenon referred to as *molecular sieving*), and serum proteins can be resolved into more individual fractions than with an agarose gel. Furthermore, these gels are uncharged, thus eliminating electroendosmosis. Precast minigels are available in a variety of concentrations and acrylamide-to-bis-acrylamide ratios suitable for most protein or nucleic acid separations. However, because of the known neurotoxicity of acrylamide, appropriate caution must be exercised when handling this material if the gels are prepared by hand.

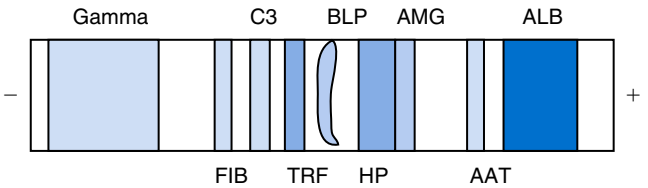


Fig. 11.3 A simplified schematic drawing of a protein pattern from the serum of a subject with haptoglobin type 2-1 (separation by polyacrylamide gel electrophoresis). Some zones contain more than the one protein shown, as demonstrated by immunologic techniques. *AAT*, α_1-Antitrypsin; *ALB*, albumin; *AMG*, α_2-macroglobulin; *BLP*, β-lipoprotein; *C3*, complement 3; *FIB*, fibrinogen; *Gamma*, γ-globulin; *HP*, haptoglobin; *TRF*, transferrin.

Attempts to improve the hydrophilic nature of polyacrylamide have led to the development of mono- and di-substituted monomers, one of which is *N*-acryloyl-tris(hydroxymethyl) aminomethane, or Poly(NAT). This material is more hydrophilic than polyacrylamide, and its matrix has larger pores, thereby presenting less resistance to the passage of large molecules. It is ideally suited to the separation of DNA fragments up to 20 kilobases (kb) in size using a homogeneous (non-pulsed) electric field. Fragments that differ in size by as little as 2% can be resolved. Gels are submersed in buffer during use, allowing temperatures to be tightly controlled at values between 50°C and 60°C. Use of increased temperatures results in shorter run times and more reproducible band migration.

Automated Systems

Because of increased volume of testing, primarily for serum proteins, many laboratories are converting to automated systems for electrophoresis. Such a system is the Helena SPIFE 4000 (Helena Laboratories, Beaumont, TX), an automated electrophoresis system providing automated reagent application and a variety of gel sizes that permit analysis of 10 to 100 samples simultaneously. It also features in-line sample application, automated electrophoretic separation and staining of analytes, multiple stain ports, and positive sample identification. The Interlab Microgel system (Interlab Srl, Rome, Italy) also fully automates the process and integrates sample application, temperature-controlled electrophoresis, staining, and densitometry into a single unit with the capability of managing four gels simultaneously. Other systems that have partially automated the procedure or incorporated the ability to process sequentially multiple gels of different compositions include the Phast System (Pharmacia LKB, Gaithersburg, MD), the HITE Fractoscan (Olympus, Invicon, München, Germany), the Hydragel-Hydrasys (Sebia Inc., Durham, NC), and the High-Performance Gel Electrophoresis (HPGE)-1000 system (LabIntelligence Inc., Belmont, CA). Most CE systems (see "Capillary Electrophoresis" section) have autosampling capability for sequentially processing specimens, but the Sebia CAPILLARYS permits simultaneous processing of seven samples by using multiple capillaries. Newer microchip-based analyzers like the Agilent 2100 Bioanalyzer (Agilent Technologies Inc., Santa Clara, CA) significantly miniaturize and increase the speed of the process for separating proteins, nucleic acids, or even entire cells. These advances substantially reduce the labor component associated with electrophoresis.

POINTS TO REMEMBER

Capillary Zone Electrophoresis

- Uses electroosmotic flow (EOF) to mobilize the sample plug pass the detector
- The magnitude and direction of EOF can be manipulated through methodology
- Higher voltages can be applied due to more efficient dissipation of heat; this can translate to faster separations

CAPILLARY ELECTROPHORESIS

With CE, classic techniques of are carried out in a small-bore (10- to 100-μm internal diameter) fused silica capillary tube, 20 to 200 cm in length.

Two distinct advantages of the capillary format include the ability to apply much higher voltages than in traditional electrophoresis and the ease of automation. Applications are also more extensive and include the separation of low molecular weight ions in addition to proteins and other macromolecules. Even uncharged molecules can be separated using CE in the micellar electrokinetic chromatography (MEKC) mode discussed later. CE has also proved useful for separation of inorganic ions, amino acids, organic acids, drugs, vitamins, porphyrins, carbohydrates, oligonucleotides, proteins, and DNA fragments.

General Operations

A schematic diagram of a typical instrumental configuration for CE is shown in Fig. 11.4. As indicated, the capillary serves as an electrophoretic chamber, analogous to a lane on a gel, which is connected to buffer reservoirs at both ends, which, in turn, are connected to a high-voltage power supply. It is important to note that at some point along the length of the capillary (typically close to the end), a detector is interfaced for online detection. Improved heat dissipation from the capillary (as opposed to a slab gel) permits the application of voltages in the range of 10 to 30 kV, which enhances separation efficiency and reduces separation time—in some cases to less than 1 minute. Only a few microliters of the sample are required, with injected volumes in the nanoliter range. The small sample plug volume minimizes distortions in the applied field caused by the presence of analytes or other sample species.

In contrast to the cumbersome and time-consuming tasks of conventional electrophoresis, CE is easily automated. Analogous to high-performance liquid chromatography (HPLC) technology, samples typically are stored in a temperature-controlled environment and are automatically injected into the capillary, with a variety of detector types available; the resulting electropherograms are analyzed and manipulated in much the same manner as chromatograms.

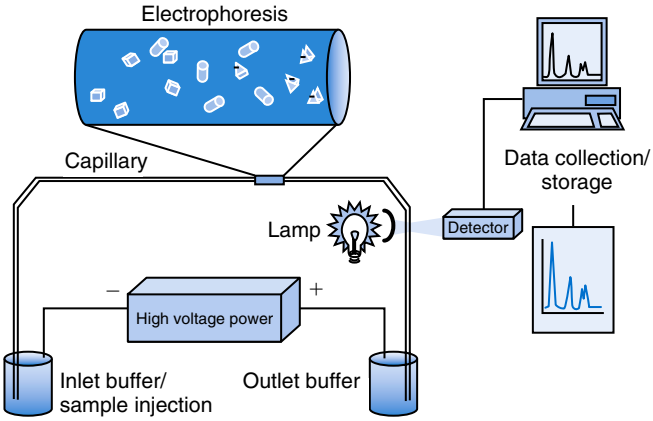


Fig. 11.4 A schematic for capillary electrophoresis instrumentation.

The capillaries used as separation columns in CE are most commonly made from fused silica (i.e., pure glass) coated with a thin exterior covering of polyimide to provide strength and flexibility. Although capillaries can be made from other materials, such as polyethylene or Teflon, such capillaries have seen limited use. The polyimide coating is usually removed from a small portion of the capillary close to the terminal end, creating a window for online optical detection. The outer diameter of the capillary tubing typically varies from 180 to 375 μm, the inner diameter from 20 to 180 μm, and the total length from 20 cm up to several meters. Noncylindrical capillary tubing suitable for CE is now available from some commercial providers. For example, rectangular capillaries (Polymicro Technologies, Phoenix, AZ) provide a flat surface that is more amenable to optical detection than that of their curved counterparts.

CE is distinguished from other forms of electrophoresis by the fact that extraordinarily high voltage (30,000 volts) has been used to obtain rapid, high-efficiency separations. The problems encountered with noncapillary platforms (e.g., slab gels) are avoided because of the effective dissipation of Joule heat by forced air convection or liquid cooling of the narrow-bore capillary.

Sample Injection

In CE, sample volumes of 1 to 50 nL are loaded into the capillary by *hydrodynamic injection* or *electrokinetic (EK) injection*. With hydrodynamic injection, an aliquot of a sample is introduced by applying positive pressure at the inlet vial or vacuum at the outlet vial. The volume of sample loaded is governed by a number of parameters, including (but not restricted to): (1) the inner diameter of the capillary; (2) buffer viscosity; (3) applied pressure; (4) temperature; and (5) time. With some earlier commercial or homemade systems, gravity was used as the source of pressure by raising the inlet vial (or lowering the outlet vial), thus allowing "siphoning" to occur for a timed interval. With EK injection, an aliquot of a sample is introduced by applying a voltage for a timed interval. The magnitude of the voltage is dependent on the analyte and buffer system used but typically involves field strengths three to five times lower than that used for separation. It is important to note that although hydrodynamic methods introduce a sample representative of the bulk specimen, EK injection favors the preferential movement of more electrokinetically mobile analytes into the capillary.

In practice, to maintain high separation efficiency, the sample plug length is usually less than 2% of the total capillary length.

Direct Detection

With CE, separated analytes are detected and measured as they migrate past a point on the capillary that is optically interrogated. Optical detection is based on classical methods, such as photometric absorbance, refractive index, and fluorescence (see Chapter 9). As with HPLC, ultraviolet-visible photometers are widely used as detectors to monitor CE separations. To interface such online detectors with the capillary, a detection window is created toward the outlet end of the capillary. This "window," which serves as an inline cuvette, typically is formed by burning off the polyimide with a small flame and cleaning the window with ethanol. Although this configuration allows high-efficiency separation, the inner diameter of the capillary tube defines the optical path length (OPL) of the inline cuvette. Because absorbance is directly proportional to the length of the cuvette used in an optical system, the 20 to 100 μm inner diameter of the capillary limits UV-visible absorbance detection limits to concentrations of 10^{-8} to 10^{-6} molar.

More sensitive optical techniques that have been used with CE include: (1) fluorescence; (2) refractive index; (3) chemiluminescence; (4) Raman spectrophotometry; and (5) circular dichroism. The most sensitive optical detection method used in CE is laser-induced fluorescence, which is capable of detection limits in the 10^{-9} to 10^{-12} molar (or better) range. This detection mode is easily accomplished with analytes that may be easily labeled with a fluorescent substrate (e.g., intercalators for double-stranded DNA) or may be naturally fluorescent (e.g., proteins or peptides containing tryptophan). CE systems have also been interfaced with mass spectrometers, and electrochemical detection methods have been developed, although such detectors must be isolated electrically from the electrophoretic voltages.

Indirect Detection

When strong chromophores are lacking in the analyte of interest, absorbance and fluorescence detection have been used in an indirect mode. In this mode, a strongly absorbing ion is added to the running electrolyte and is monitored at a wavelength that gives a constant, high background absorbance. As solute ions move into their discrete zones during the electrophoretic process, they displace the indirect detection agent through mutual repulsion, and this produces a decrease in background absorbance as the zone passes through the detector. Reagents with appropriate fluorescence properties have been used in a similar manner. Indirect detection of amino acids by CE has been demonstrated, with the potential for use in the diagnosis of amino acidurias. Investigators have demonstrated the direct extrapolation of this technique to microchip detection when UV detection is difficult, if not impossible.

TYPES OF ELECTROPHORESIS

Capillary Zone Electrophoresis

CZE, also called *open-tube* or *free-solution* CE, is the simplest form of CE. It includes *capillary ion electrophoresis*, which refers to the analysis of inorganic ions by CZE, often using indirect detection. The power of the CZE mode is its ability to electrophoretically resolve charged species without a sieving matrix; this applies to a broad spectrum of analytes ranging from proteins, peptides, and amino acids to small molecules (e.g., drugs) and ions.

Capillary Gel Electrophoresis

Capillary gel electrophoresis (CGE) is directly comparable with traditional slab or tube gel electrophoresis because the

separation mechanisms are identical. Size separation is achieved with a suitable polymer, which acts as a molecular sieve or sizing mechanism. As charged analytes migrate through the polymer network, they become hindered to a degree that is governed by their size (larger molecules are hindered more than smaller ones). Macromolecules, such as DNA and sodium dodecyl sulfate (SDS)-saturated proteins, cannot be separated without a gel or some other separation mechanism, because they have a mass-to-charge ratio that is size independent. The term *gel* in CGE is a misnomer, primarily because cross-linked "gels," as we know them in slab format, are not routinely used in CE. A more suitable term is a *sieving matrix* or *soluble polymer network*, a linear polymeric structure that is soluble, has reasonably low viscosity, and is capable of self-entangling in a manner that forms pores through which sieving can occur. A variety of polymeric matrices have been defined for DNA (e.g., polyacrylamide, cellulosic materials) and protein analysis (e.g., dextran-base matrices), provided that pores can be formed inherently that have diameters in the range of tens to hundreds of nanometers. One of the requirements that often accompany this type of analysis is reduction of electro-osmotic flow. This is accomplished by covalently, adsorptively, or dynamically coating the surface. Cross-linked polyacrylamide was the main polymer of choice for this but recently has been supplanted by a host of polymeric matrices that not only provide effective molecular sieving but also adsorptively coat the capillary surface.

TECHNICAL CONSIDERATIONS

Gel

In performing electrophoretic separations, a number of technical and practical aspects need to be considered, as they affect the process.

Sampling

To achieve a proper balance between sensitive measurements and resolution, the amount of serum protein applied to an electrophoretic support must be optimum. Albumin is about 10 times more concentrated in serum than the smallest fraction, the α_1-globulins. Therefore, the amount of serum applied should prevent overloading with albumin, but should still be adequate to quantify α_1-globulin. For the separation of serum proteins using PAGE, 3 μL of serum containing approximately 210 μg of total protein is applied. For alkaline phosphatase isoenzymes, up to 25 μL of a normal serum may be applied (less may be used if activity is greatly increased). Urine specimens require 50- to 100-fold greater concentrations or extended application times for adequate sensitivity, and CSF may or may not require concentration, depending on the staining approach used.

Discontinuities in Sample Application

Discontinuities in sample application may be caused by: (1) dirty applicators; (2) uneven absorption by sample combs; or (3) inclusion of an air bubble if the sample is pipetted onto the gel. The pipette tip should be checked for air bubbles before the sample is applied to the agarose gel template.

Unequal Migration Rates

Unequal migration of samples across the width of the gel may be caused by dirty electrodes, which may cause uneven application of the electric field, or by uneven wetting of the gel. If wicks are used to connect the gel to a power supply, uneven wetting of the wicks could cause unequal migration or bowing of sample lanes at the gel edges. Gels must be kept horizontal during storage to avoid sagging and uneven thickness. Finally, gels that may have been stored too close to heat sources (e.g., in a cabinet over a light fixture) could have partially and unevenly dried areas, contributing to similar problems.

Distorted, Unusual, or Atypical Bands

Distorted protein zones may be caused by: (1) bent applicators; (2) incorporation of an air bubble during sample application; (3) over-application; or (4) inadequate blotting of the sample. Excessive drying of the electrophoretic support before or during electrophoresis may also cause distorted zones. Irregularities (other than broken zones) in the sample application probably are due to excessively wet agarose gels. Portions of applied samples may look washed out.

In most cases, unusual bands are artifacts that may be easily recognized. Hemolyzed samples are frequent causes of increased β-globulin (where free hemoglobin migrates) or an unusual band between the α_2- and β-globulins, the result of a hemoglobin–haptoglobin complex. A band occurring at the starting point of an electropherogram may be fibrinogen. The sample should be verified as being serum before this band is reported as an abnormal protein. The α- and β-lipoproteins may migrate ahead of their normal positions in some samples. Occasionally, a split albumin zone is observed in the rare, benign, genetically related condition of bis-albuminemia. However, a grossly widened albumin zone could be due to albumin-bound medication and not to faulty electrophoresis.

Capillary

Temperature and surface effects influence the separation capabilities of CE. Artifacts also have been known to arise with CE.

Temperature Effects

In most slab or tube platforms for electrophoresis, moderate electric fields (up to 1000 volts) are used, because the Joule heating that accompanies the use of higher field strengths causes nonuniform temperature gradients, local changes in viscosity, and subsequent zone broadening. CE is distinguished from other forms of electrophoresis by the fact that extraordinarily high fields (30,000 volts) are used to obtain rapid, high-efficiency separations. The problems encountered with noncapillary platforms are prevented by effective dissipation of Joule heat by forced air convection or liquid cooling of the capillary, both of which are possible because of the narrow bore of the capillary. The Joule heat produced is a function of (1) buffer type, (2) concentration, (3) voltage applied, (4) capillary inner diameter, and (5) length, and can be determined

for any given system by generating an *Ohm's law plot*, which allows easy determination of the maximum voltage that can be used effectively. Reducing the inner diameter of the capillary, the ionic strength of the running buffer, or the applied voltage will reduce the heat produced by the electrophoretic process. It should be noted that reducing the inner diameter will compromise the detection limit of UV measurements (smaller OPL); reducing the applied field is less desirable in that resolution is directly proportional to the applied field. Consequently, attempts should be made to alter other parameters before reducing inner diameter or the applied field.

Surface Effects

As in electrophoresis in general, the flow of fluid (electro-osmotic or electroendosmotic flow [EOF]) in CE is a consequence of surface charge on the solid support. In CE, EOF can play a significant role in the separation process. The charge on the inner surface of a fused silica capillary is determined by the ionization state of the silanol groups (SiOH) that populate it. Interaction of positively charged buffer species with bound surface anions generates a layer of mobile cations that move toward the cathode when voltage is applied. This induces a very strong EOF that mobilizes all analytes in the same direction, regardless of their charge. Separation is consequently achieved because of differences in the electrophoretic migration rates of analytes superimposed on this EOF.

Because the driving force of the flow is distributed along the wall of the capillary, the flow profile is nearly flat or plug-like, contrasting with the laminar or parabolic flow generated by a pressure-driven system caused by shear forces at the wall. A flat flow profile is beneficial because it does not contribute to the dispersion of solute zones. The magnitude and direction of the EOF are influenced by several parameters, including: (1) type of electrolyte used; (2) pH; (3) ionic strength; (4) use of additives (e.g., surfactants, organic solvents); and (5) polarity and magnitude of the applied electric field.

Although advantageous for the dissipation of Joule heat, the large surface area-to-volume ratio of the inner capillary space increases the likelihood of analyte adsorption onto the surface of its inner wall. This causes phenomena such as peak tailing and even total and irreversible adsorption of the analyte. Adsorption is typically noted between cationic solutes and the negatively charged inner wall of the capillary, primarily through ionic interactions (with deprotonated silanols), but also involves hydrophobic interactions (with siloxanes). Because of the numerous charges and hydrophobic regions, significant adsorptive effects have been noted, especially for highly cationic proteins. In practice, adsorption of substances, whether from the sample or from the buffer, to the inner surface of the capillary will alter migration times and other separation characteristics; unaddressed, the capillary eventually may become "fouled." Buffer components and/or additives, such as surfactants, can often render permanent changes to the inner surface of the capillary (through adsorption) and may warrant dedication of specific capillaries for use with particular surfactants.

To minimize these inner wall effects, capillaries are conditioned by chemical treatment, most commonly with base, to remove adsorbates and rejuvenate the surface. A typical wash method includes flushing the chamber with 10 to 20 capillary volumes of 0.1 to 1.0 mol/L NaOH, followed by flushing with "run" buffer. To prevent exposing the capillary surface to drastic fluctuations in pH, conditioning procedures for separations at low pH may be better served by using strong acids (e.g., HNO_3), surfactants (e.g., SDS), or organic solvents, such as acetonitrile or methanol.

Serum Protein Analysis. Compared with AGE and cellulose acetate electrophoresis (CAE), CZE is more advantageous for serum protein analysis. Fig. 11.5 shows a comparison of the separation of serum proteins by CAE, AGE, and CE. The presence of the classical zones with CE is apparent, albeit in reversed order, as is the identification of serum protein abnormalities in gamma regions. Retrospective studies have shown CE to be effective for detecting monoclonal proteins, which could then be immunotyped by conventional techniques (IFE and **isoelectric focusing** [IEF]). Moreover, CE can do both serum protein electrophoresis and immunotyping in hundreds of samples simultaneously. These and other studies put forth the same conclusion: that CE is more sensitive than AGE in identifying abnormalities. Furthermore, CZE can be used effectively in serum protein analysis.

Artifacts in Serum Protein Analysis. With CE using online optical detection, artifacts can occur in the form of "system peaks." These often originate from the sample or from the interfaces between the sample and the separation buffer, because any species that absorbs at the detection wavelength will generate a response. This differs from protein slab gel

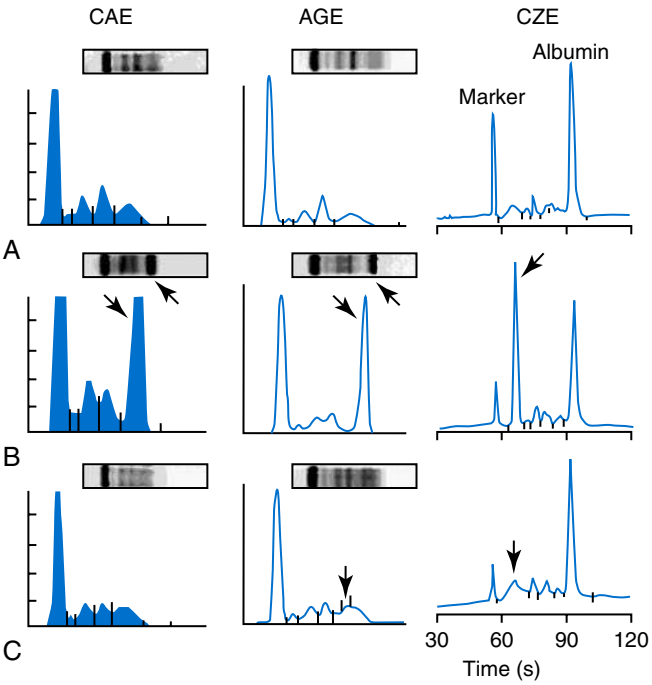


Fig. 11.5 Rapid protein electrophoresis of serum protein; comparison with scanning densitometry profiles obtained from cellulose acetate *(CAE)* and agarose *(AGE)* electrophoresis. (A) Normal serum. (B) Patient serum containing a large M-protein. (C) Patient serum containing a small monoclonal protein. *Arrows* indicate the position of the monoclonal proteins. *CZE*, Capillary zone electrophoresis.

electrophoresis, wherein detection specificity is governed by a protein-specific stain. It is not uncommon, for example, for buffer species present in the sample but not in the separation buffer to generate system peaks.

One problem associated with conventional electrophoresis of serum proteins is its proclivity for point-of-application artifacts. These are bands that result from the fact that electrophoretic mobility (e.g., with AGE) is bidirectional from the point of application. These precipitates may be euglobulin or cryoprecipitates and may or may not contain a monoclonal protein; only immunoelectrophoresis or IFE can identify the presence of monoclonal proteins. These bands must be immunotyped to distinguish real monoclonal proteins from artifacts—a process that is costly and time-consuming. CE prevents point-of-application artifacts in two ways. First, net mobility in CE results from the vectorial addition of both protein electrophoretic mobility and EOF. As a result of this unidirectional movement (toward the detector), the point of application remains removed from the detector. Second, unlike AGE, where precipitates cannot exit the loading well and enter the gel (thus appearing as a band in the scanned region of the gel), no gel matrix is present in CZE to impede electrophoretic migration because analysis occurs in free solution.

Improving Limits of Detection. Several approaches have been devised to improve the limit of detection of online CE detectors. These include increasing the length of the OPL and online concentration of the sample.

Increased Optical Path Length. Capillary tubes modified at the detector window with a "bubble" cell (a glass-blown expansion of the internal diameter of the capillary tube) can expand the OPL by almost an order of magnitude, with concomitant lowering of the system's limit of detection. Alternatively, a "Z" geometry has been developed that increases the OPL via detection down the core of the capillary, with possible lengths up to 1 mm.

Online Sample Concentration. Another technique used in CE systems to increase their limit of detection is preconcentration of the sample. One of the simplest methods for sample preconcentration is to induce a "stacking" effect with the sample components, which is easily accomplished by exploiting the ionic strength differences between the sample matrix and the separation buffer. This results from the fact that sample ions have decreased electrophoretic mobility in a higher conductivity environment. When voltage is applied to the system, sample ions in the sample plug instantaneously accelerate toward the adjacent separation buffer zone. Upon crossing the boundary, the higher conductivity environment induces a decrease in electrophoretic velocity and subsequent "stacking" of the sample components into a smaller buffer zone than the original sample plug. Within a short time, the ionic strength gradient dissipates and the charged analyte molecules begin to move from the "stacked" sample zone toward the cathode. Stacking has been used with hydrostatic or EK injection and typically yields a tenfold enhancement in sample concentration, resulting in a lower limit of detection.

An alternative approach to stacking is "focusing," which is based on pH differences between the sample plug and the separation buffer. This is very useful for the analysis of peptides, mainly because of their relative stability over a wide pH range. By increasing the pH of the sample to above that of the net pI of the analytes of interest and flanking the sample plug with low pH separation buffer zones (i.e., an equivalent volume of low pH separation buffer following introduction of the sample plug), negatively charged peptides are electrophoretically driven toward the anode. Upon entering the lower pH separation buffer, a pH-induced change in their charge state causes a reversal in their electrophoretic mobility, resulting in "focusing" of the peptides at the interface of the sample (high pH) and low pH buffer plugs (similar to those in isoelectric focusing). After the pH gradient dissipates, the peptides, again positively charged, migrate toward the cathode as a sharp zone. This approach has been applied to a variety of analytes, but it is limited to those that are able to withstand inherent changes in pH without substantial denaturation, and may yield as much as a five-fold enhancement of a system's limit of detection.

SPECIALITY ELECTROPHORESIS TECHNIQUES

Isoelectric Focusing Electrophoresis

IEF separates amphoteric compounds, such as proteins, with increased resolution in a medium possessing a stable pH gradient. The protein becomes "focused" at a point on the gel as it migrates to a zone where the pH of the gel matches the protein's pI. At this point, the charge of the protein becomes zero, and its migration ceases. Fig. 11.6 illustrates the procedure and shows the electrophoretic conditions before and after current is applied. The protein zones are very sharp, because the region associated with a given pH is very narrow. Ordinary diffusion is also counteracted by the acquisition

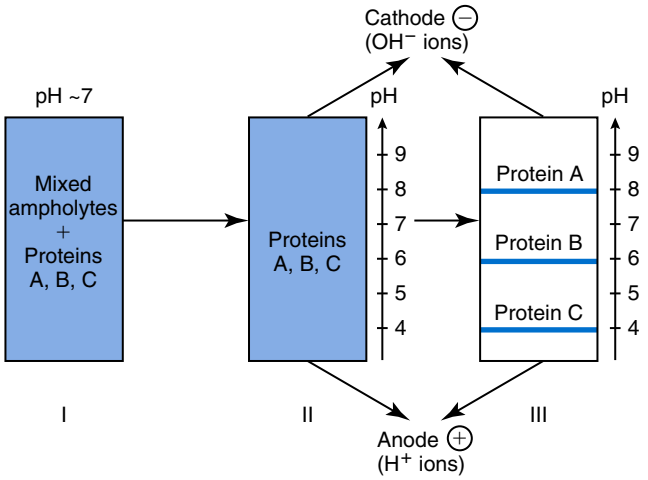


Fig. 11.6 Schematic of an isoelectric focusing procedure. (I) A homogeneous mixture of carrier ampholytes, pH range 3 to 10, to which proteins A, B, and C, with isoelectric points (pI) of 8, 6, and 4, respectively, were added. (II) Current is applied and carrier ampholytes rapidly migrate to pH zones where the net charge is zero (the pI value). (III) Proteins A, B, and C migrate more slowly to their respective pI zones, where migration ceases. The high buffering capacity of the carrier ampholyte creates stable pH zones in which each protein may reach its pI.

of a charge, as a protein varies from its pI position and subsequently migrates back because of electrophoretic forces. Proteins that differ in their pI values by only 0.02 pH unit have been separated by IEF.

The pH gradient is created with *carrier ampholytes*, a group of amphoteric polyaminocarboxylic acids that have slight differences in pKa value and molecular weights of 300 to 1000. Mixtures of 50 to 100 different compounds are added to the medium and create a "natural pH gradient" when individual ampholytes reach their pI values during electrophoresis. They establish narrow buffered zones, with stable but slightly different pHs, through which the slower-moving proteins migrate and stop at their individual pIs.

The anode is surrounded by a dilute acid solution and the cathode by a dilute alkaline solution. After focusing, the most negatively charged carrier ampholytes and proteins will be found at the anodal end, and the most positively charged near the cathodal end of the electrophoretic matrix. The other carrier ampholytes and proteins focus at intermediate points according to their pI values. Because carrier ampholytes are generally used in relatively high concentrations, a high-voltage power source (up to 2000 V) is necessary (power is in the vicinity of 2 to 50 W, depending on experimental conditions). As a result, the electrophoretic matrix must be cooled.

PAGE-IEF is widely used in analytical work, as it is essentially free of electroendosmosis. However, the polyacrylamide gel must have a large enough pore size so that protein migration will not be impeded by molecular sieving effects. In actual practice, impeded migration of some proteins, such as IgM, cannot be prevented. PAGE-IEF has the advantages that operating conditions are simple, and that large pore sizes make it unlikely that any proteins will be excluded on the basis of molecular size. IEF has been applied to the separation of alkaline phosphatase isoenzymes and is widely used in neonatal screening programs to test for variant hemoglobins. Off-gel techniques carry out the separation in free solution with sample containing ampholytes loaded into each of a linear series of wells separated by semipermeable membranes and in contact with a pH gradient strip. Electrophoresis separates sample proteins into different wells depending on their pI values. Separated fractions can be further resolved in a second similar focusing step or taken directly to further separation by two-dimensional electrophoresis or liquid chromatography-mass spectrometry. This technique has been useful in the study of the human proteome.

Micellar Electrokinetic Chromatography

MEKC is a hybrid of electrophoresis and chromatography. MEKC, a mode that is separate and distinct from capillary electrokinetic chromatography, is an effective electrophoretic technique, because it can be used for the separation of neutral and charged solutes. The separation of neutral species is accomplished by exploiting micelles formed in the running buffer when the concentration of surfactant exceeds the critical micelle concentration (e.g., 8 to 9 mmol/L for SDS). During electrophoresis, neutral micelles can interact with analytes in a chromatographic manner through hydrophobic interactions in which analytes are micellized based on their degree of hydrophobicity. Under these conditions, partitioning into the micelle is the driving force for separation. With charged micelles (e.g., SDS), analytes can also interact through electrostatic interactions via the charge on the surface of the micelle.

REVIEW QUESTIONS

1. For proteins to migrate on a gel during electrophoresis, the isoelectric point must be exceeded. The isoelectric point of an amino acid or protein is defined as the:
   a. ability of the amino acid or protein to carry both positive and negative charges.
   b. pH at which the amino acid or protein has a negative charge.
   c. pH at which a molecule has no net charge.
   d. point on an electrophoretic gel at which that specific protein migrates.
2. The term *amphoteric* means:
   a. the pH at which a protein has no net charge.
   b. that a protein will have a negative charge at a high pH.
   c. that a protein has both positive and negative charges because of its side chains.
   d. the positive electrode.
3. Unequal migration of samples down the gel in an electrophoresis assay is an interference usually caused by:
   a. dirty electrodes causing uneven application of the electrical field.
   b. dirty pipette tips or applicators.
   c. too much sample being applied.
   d. inappropriate binding of immunoglobulins to other proteins in a sample.
4. In serum protein electrophoresis, the greater the charge of a molecule:
   a. the slower it will migrate.
   b. the faster it will migrate.
   c. the less the migration rate will be affected.
5. Serum protein electrophoresis is what kind of basic technique?
   a. Immunoassay
   b. Standardization
   c. Qualitative technique
   d. Separation

6. Specimens with very low protein content sometimes require preconcentration before they are loaded onto an electrophoretic gel. An electrophoretic technique that *does not* require this concentration step is:
   a. disc electrophoresis.
   b. capillary electrophoresis.
   c. zone electrophoresis.
   d. two-dimensional gel electrophoresis.
7. What optical measuring technique allows for the scanning and quantification of electrophoretic fractions separated on a support medium?
   a. Spectrophotometry
   b. Chromatography
   c. Densitometry
   d. Nephelometry
8. A Southern blot is used to detect fragments of:
   a. RNA.
   b. protein.
   c. agarose.
   d. DNA.
9. A 1% agarose gel is best suited for separating DNA fragments in which one of the following size ranges:
   a. 500 bp (0.5 kbp) to 20 kbp
   b. 50 to 100 kbp
   c. 100 to 500 kbp
   d. larger than 500 kbp
10. The component of an electrophoresis system in which separation takes place is the:
   a. stain.
   b. support medium.
   c. densitometer.
   d. zone.

SUGGESTED READINGS

Bazydlo, L. A. L., & Landers, J. P. (2018). Electrophoresis. In N. Rafai, A. R. Horvath, & C. T. Wittwer (Eds.), *Tietz textbook of clinical chemistry and molecular diagnostics* (6th ed.) (pp. 250–265). St Louis: Elsevier Inc.

Breadmore, M. C. (2015). Recent advances in enhancing the sensitivity of electrophoresis and electrochromatography in capillaries and microchips. *Electrophoresis*, *36*, 36–61.

Dawod, M., Arvin, N. E., & Kennedy, R. T. (2017). Recent advances in protein analysis by capillary and microchip electrophoresis. *Analyst*, *142*, 1847–1866.

Gassmann, M., Grenacher, B., Rohde, B., & Vogel, J. (2009). Quantifying Western blots: pitfalls of densitometry. *Electrophoresis*, *30*, 1845–1855.

Gay-Bellile, C., Bengoufa, D., Houze, P., Le Carrer, D., Benlakehal, M., Bousquet, B., et al. (2003). Automated multicapillary electrophoresis for analysis of human serum proteins. *Clinical Chemistry*, *49*, 1909–1915.

Herpol, M., Lanckmans, K., Van Neyghem, S., Clement, P., Crevits, S., De Crem, K., et al. (2016). Evaluation of the Sebia Capillarys 3 Tera and the Bio-Rad D-100 systems for the measurement of hemoglobin A1c. *American Journal of Clinical Pathology*, *146*, 67–77.

Kleparnik, K. (2015). Recent advances in combination of capillary electrophoresis with mass spectrometry: methodology and theory. *Electrophoresis*, *36*, 159–178.

Landers, J. P. (1995). Clinical capillary electrophoresis. *Clinical Chemistry*, *41*, 495–509.

Landers, J. P. (2003). Molecular diagnostics on electrophoretic microchips. *Analytical Chemistry*, *75*, 2919–2927.

Landers, J. P. (2008). Introduction to capillary electrophoresis. In J. P. Landers (Ed.), *Handbook of capillary and microchip electrophoresis and associated microtechniques*. (3rd ed.) (pp. 3–74). Boca Raton, FL: CRC Press.

Li, S. F. Y., & Kricka, L. J. (2005). Clinical analysis by microchip capillary electrophoresis. *Clinical Chemistry*, *52*, 37–45.

Ouimet, C. M., D'amico, C. I., & Kennedy, R. T. Advances in capillary electrophoresis and the implications for drug discovery. *Expert Opinion Drug Discovery, 12,* 213–224.

Sharma, K., Lee, S., Han, S., Lee, S., Francos, B., McCue, P., et al. (2005). Two-dimensional fluorescence difference gel electrophoresis analysis of the urine proteome in human diabetic nephropathy. *Proteomics*, *5*, 2648–2655.

Willrich, M. A., Ladwig, P. M., Andreguetto, B. D., Barnidge, D. R., Murray, D. L., Katzmann, J. A., et al. (2016). Monoclonal antibody therapeutics as potential interferences on protein electrophoresis and immunofixation. *Clinical Chemistry and Laboratory Medicine*, *54*, 1085–1093.

Zhang, W., Hankemeier, T., & Ramautar, R. (2017). Next-generation capillary electrophoresis-mass spectrometry approaches in metabolomics. *Current Opinion in Biotechnology*, *43*, 1–7.

12

Chromatography

*David S. Hage**

OBJECTIVES

1. Define the following terms:
   - Chromatogram
   - Chromatograph
   - Chromatography
   - Column chromatography
   - Mobile phase
   - Planar chromatography
   - Retention factor
   - Retention time and retention volume
   - Stationary phase
   - Void time and void volume
2. Evaluate a chromatogram and its peaks to determine their separation factor and resolution; list three ways to improve resolution.
3. Define and calculate the following terms for a chromatographic separation: number of theoretical plates and plate height. Explain how the plate height can vary with the linear velocity of the mobile phase.
4. Explain how external or internal calibration can be used for quantitative measurements in chromatography. Define *internal standard* and discuss how this can be used in quantitative measurements by chromatography.
5. Define gas chromatography (GC) and describe three types of GC based on the type of stationary phase that is present. List some stationary phases for each of these methods.
6. Summarize the components of a gas chromatograph and describe the function of each component.
7. Explain the differences between packed columns and capillary columns in GC.
8. Define the following terms and explain how they are used in GC: headspace analysis, isothermal elution, and temperature programming.
9. List and describe five types of detectors that are used in GC, including the advantages and disadvantages of each.
10. Define the terms *liquid chromatography* and *high-performance liquid chromatography*. List several ways in which liquid chromatography differs from GC.
11. Describe the following types of liquid chromatography, including the principles of separation, the types of stationary phases and mobile phases that are used, and some possible clinical applications of each.
    - Adsorption chromatography
    - Affinity chromatography
    - Hydrophilic interaction liquid chromatography
    - Ion-exchange chromatography
    - Ion-pair chromatography
    - Normal-phase chromatography
    - Reversed-phase chromatography
    - Size-exclusion chromatography
12. Summarize the components of a system that are used in high-performance liquid chromatography and describe the function of each component.
13. List and describe five types of detectors that are used in liquid chromatography, including the advantages and disadvantages of each.
14. Explain the principles of planar chromatography. Define the terms *paper chromatography* and *thin-layer chromatography*.
15. Know how to calculate and use the retardation factor to identify compounds in planar chromatography.

KEY WORDS AND DEFINITIONS

Adsorption chromatography A type of liquid chromatography in which chemicals are retained based on their adsorption and desorption at the surface of an underivatized support.

Affinity chromatography (AC) A liquid chromatographic method that makes use of biologically related interactions for the retention and separation of chemicals.

* The author gratefully acknowledges the contributions of Drs. Glen L. Hortin, Bruce A. Goldberger, M. David Ullman, Carl A. Burtis, and Larry D. Bowers to this chapter in previous editions.

Anion-exchange chromatography A type of ion-exchange chromatography that uses positively charged groups as a stationary phase to bind and separate anions.

Bonded phase gas chromatography A type of gas chromatography in which the stationary phase is chemically linked or attached to the support.

Carrier gas A term often used to describe the mobile phase in gas chromatography.

Cation-exchange chromatography A type of ion-exchange chromatography that uses negatively charged groups as a stationary phase to bind and separate cations.

Chiral stationary phase (CSP) A stationary phase that is chiral and that can interact with compounds in a stereospecific manner.

Chromatogram A plot of detector response in chromatography as a function of the time or volume of the mobile phase that is needed to elute an analyte from a column.

Chromatograph The instrument that is used to perform a separation in chromatography.

Chromatography A method in which the components of a mixture are separated based on their differential interactions with two chemical or physical phases: a mobile phase and a stationary phase.

Column chromatography A type of chromatography that uses a column, or a tube, to contain the stationary phase and support.

Gas chromatography (GC) A type of chromatography that uses a gas as the mobile phase.

Gas chromatography/mass spectrometry (GC/MS) The use of a mass spectrometer as a detector for gas chromatography.

Gas-liquid chromatography (GLC) A type of gas chromatography in which the stationary phase is a liquid that is placed as a coating or layer on the support.

Gas-solid chromatography (GSC) A type of gas chromatography in which the same material acts as both the stationary phase and the support; chemicals are retained by their adsorption to the surface of the support.

Headspace analysis A sample introduction technique in which a portion of the vapor phase (or "headspace") that is above a liquid or solid sample is used in an analysis.

Height equivalent of a theoretical plate (HETP or *H*) A measure of efficiency in chromatography that is equal to length of a chromatographic system divided by the number of theoretical plates for the system; also known as the *plate height*.

High-performance liquid chromatography (HPLC) A type of liquid chromatography that uses small, efficient supports that produce narrow peaks.

Hydrophilic interaction liquid chromatography (HILIC) A type of partition chromatography that uses a polar stationary phase, and in which chemicals partition between an organic-rich region in the mobile phase and a more polar water-enriched layer that is at or near the surface of a polar support.

Internal standard A compound that is added in a constant amount to standard solutions and samples to help normalize the results for any variations that may occur during pretreatment steps or during injection into a chromatographic system.

Ion-exchange chromatography (IEC) A type of liquid chromatography in which ions are separated by their adsorption onto a support that contains fixed charges at its surface.

Ion-pair chromatography (IPC) A technique that combines columns that are used in reversed-phase chromatography with the ability to separate ionic compounds based on their charges, as achieved by adding an ion-pairing agent to the mobile phase.

Isocratic elution Use of a constant mobile phase composition throughout a separation.

Isothermal elution A chromatographic method that uses a constant temperature for elution.

Liquid chromatography (LC) A type of chromatography that uses a liquid as the mobile phase.

Liquid chromatography–electrochemical detection (LC-EC) The combination of an electrochemical detector with liquid chromatography.

Liquid chromatography–mass spectrometry (LC-MS) A technique that combines liquid chromatography with mass spectrometry.

Mobile phase The phase that travels through a chromatographic system and carries sample components with it.

Normal-phase chromatography A type of partition chromatography that uses a polar stationary phase; also known as "normal-phase liquid chromatography," or NPLC.

Number of theoretical plates (*N*) A measure of efficiency and band-broadening that is given for a Gaussian-shaped peak by using either $N = 16\,(t_R/W_b)^2$ or $N = 5.545\,(t_R/W_h)^2$, where t_R is the retention time for the peak, and W_b or W_h are baseline width or half-height width for the peak. The number of theoretical plates is also known as the *plate number*.

Partition chromatography A liquid chromatographic method in which solutes are separated based on their partitioning between the liquid mobile phase and a stationary phase that is coated or bonded onto a solid support.

Planar chromatography A type of chromatography in which the stationary phase is coated or placed onto a flat surface, or plane.

Resolution (R_s) A measure of the separation of two adjacent peaks in chromatography; the resolution equals the difference in retention time for the components in the two peaks divided by the average of the baseline widths for these peaks.

Retardation factor (R_f) In planar chromatography, the distance traveled by a chemical from its point of application divided by the distance traveled by the mobile phase in the same amount of time.

Retention factor *(k)* A measure of retention in chromatography, which is equal to the difference in the retention time and void time divided by the void time, or the difference in the retention volume and void volume divided by the void volume; also called the *capacity factor.*

Retention time (t_R) The average time that is required for a particular chemical to pass through a chromatographic column.

Retention volume (V_R) The average volume of mobile phase that is required for a particular chemical to pass through a chromatographic column.

Reversed-phase chromatography A type of partition chromatography that uses a nonpolar stationary phase; also known as *reversed-phase liquid chromatography* (RPLC).

Separation factor *(α)* A measure of the degree of separation in chromatography that is equal to the ratio of the retention factors for two neighboring peaks, where the larger retention factor appears in the numerator; also known as the *selectivity factor.*

Size-exclusion chromatography (SEC) A liquid chromatographic technique that separates analytes based on size through the use of a porous support that has an inert surface with little or no interactions with the sample components.

Solvent programming Use of a mobile phase composition that varies during a separation.

Stationary phase The phase in chromatography that is held within the system by a support and used to interact with and separate a sample's components as these components pass through the system.

Temperature programming A chromatographic method in which the temperature is changed over time.

Ultra-high-performance liquid chromatography (UPLC or UHPLC) A liquid chromatographic method that uses smaller diameter supports and pressures up to 15,000 psi.

van Deemter equation A relationship that shows how the overall value of plate height *(H)* for a chromatographic system is affected by the linear velocity of the mobile phase *(u)*, where $H = A + B/u + C\,u$ and the terms A, B, and C are constants that represent the contributions of several types of band-broadening processes.

Void time (t_M) The elution time for a compound that is not retained in a chromatographic system.

Void volume (V_M) The volume of mobile phase that is needed to elute a compound that is not retained in a chromatographic system.

BASIC PRINCIPLES OF CHROMATOGRAPHY

General Terms and Components of Chromatography

Chromatography is a method in which the components of a mixture are separated based on their differential interactions with two chemical or physical phases: a **mobile phase** and a **stationary phase** (Fig. 12.1). The mobile phase travels through the system and carries sample components with it once the sample has been applied or injected. The stationary phase is held within the system by a support. As a sample's components pass through this system, the components that have the strongest interactions with the stationary phase will be retained more and move through the system more slowly than components that have weaker interactions. This leads to a difference in the rate of travel for these components and their separation.

The type of chromatographic system that is shown in Fig. 12.1 is known as **column chromatography** because it uses a column (or a tube) to contain the stationary phase and support. The stationary phase may be the surface of the support, a coating on this support, or a chemical layer that is bonded to the support. In column chromatography, the support may be the interior wall of the column, or it may be a material that is placed or "packed" into the column. It is also possible to use a support and stationary phase that are present on a plane, or open surface. This second format is known as **planar chromatography**.

Besides being classified based on whether they use a column or plane, chromatographic methods can be classified based on the mobile phase. For instance, a chromatographic method that uses a mobile phase that is a gas is called **gas chromatography (GC)**, while a chromatographic method that uses a liquid mobile phase is known as liquid chromatography (LC). It is also possible to divide chromatographic methods according to the type of stationary phase that is present or the way in which this stationary phase is interacting with sample components. Examples include the GC methods of **gas-solid chromatography (GSC)** or **gas-liquid chromatography (GLC)**, and the LC methods of **adsorption chromatography** or **partition chromatography**.

The instrument that is used in chromatography is known as a **chromatograph**. This instrument can provide a response that is related to the amount of a compound that is exiting (or eluting) from a column as a function of the elution time or the volume of mobile phase that has passed through the system. The resulting plot of the response versus time or volume is known as a **chromatogram**, as illustrated in Figs. 12.1 and 12.2.

The average time or volume that is required for a chemical to pass through the column is known as that chemical's **retention time (t_R)** or **retention volume (V_R)**. The elution time or volume for a compound that has no interaction with the stationary phase and is not retained is known as the **void time (t_M)** or **void volume (V_M)**. In order for two chemicals to be separated by chromatography, it is necessary for these chemicals to have different values for t_R and V_R. Retention times or retention volumes can be used to help identify the eluting compounds, while the area or height of the compound's peak can be used to measure the amount of the compound that is present.

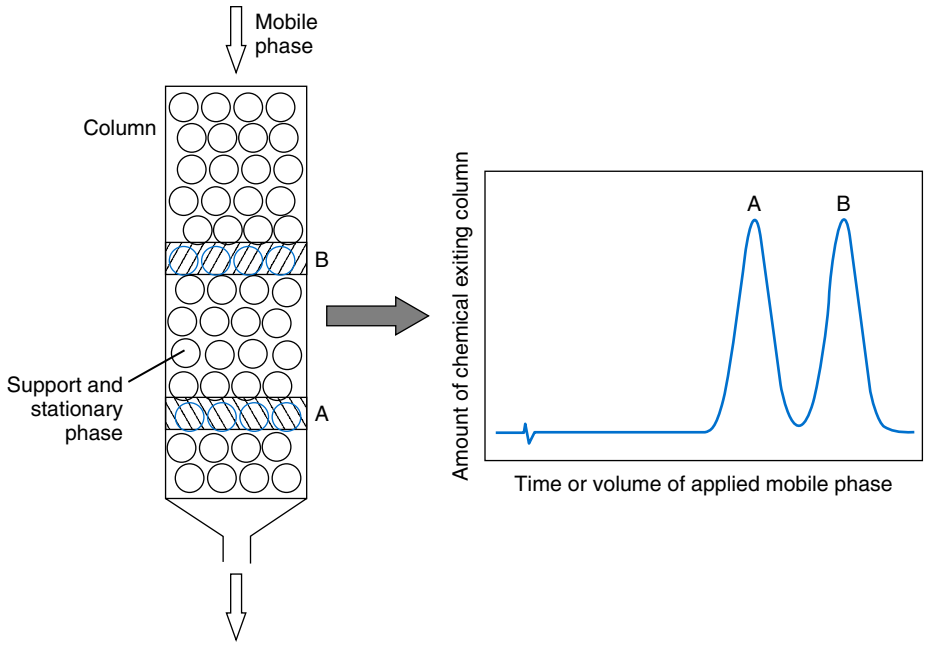
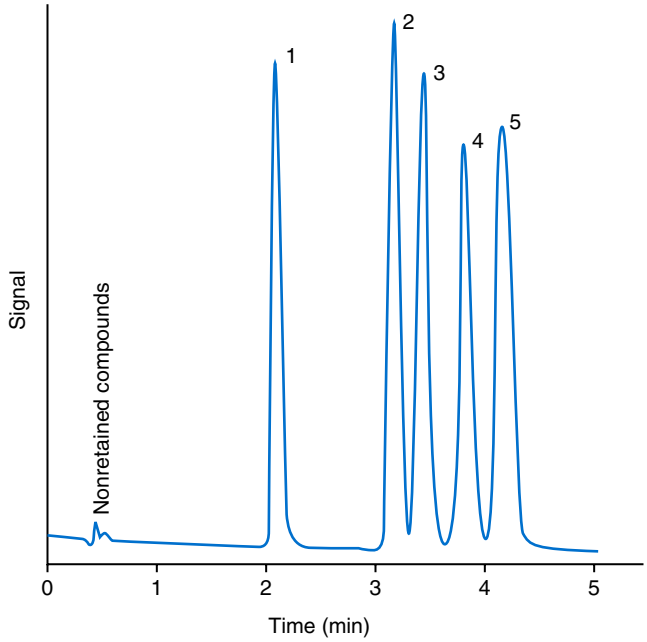


Fig. 12.1 The general components of a chromatographic system, as illustrated here by using a column to separate two chemicals, *A* and *B*.



Column: C18, 3 μm, 0.46 × 10 cm
Eluent: Isocratic, 0.025 M phosphate
Buffer: pH 3.0 in 25% acetonitrile
Flow rate: 2 mL/min
Detection: 215 nm, 0.1 AUFS

Compounds:
1. Doxepin
2. Desipramine
3. Imipramine
4. Nortriptyline
5. Amitriptyline

Fig. 12.2 Chromatogram from a separation of tricyclic antidepressants based on reversed-phase chromatography and high-performance liquid chromatography. Detection was based on the use of an absorbance detector that monitored the column eluent at 215 nm. (Note: the signal is displayed at 0.1 AUFS [absorbance units-full scale].) (Courtesy Grace Vydac Separations Group Inc., Hesperia, CA.)

The width of each peak reflects the separating "performance" or "efficiency" of the chromatographic system. The width of a peak in a chromatogram is often represented by its baseline width (W_b) or its half-height width (W_h) (Fig. 12.3). As the widths of the peaks in a chromatogram decrease, and the peaks become sharper, it becomes easier for the chromatographic system to separate two neighboring peaks and to separate multiple peaks in a given amount of time. Sharper peaks are also easier to measure than broader peaks and tend to produce better limits of detection.

Retention and Selectivity

For two chemicals to be separated by chromatography, these chemicals need to have some differences in how they are interacting with the stationary phase versus the mobile phase. These different interactions lead to differences in retention between the two chemicals. Besides using the retention time or retention volume, another way of representing retention in chromatography and identifying compounds is by using the **retention factor *(k)***, which is also sometimes called k' or the "capacity factor." This value can be calculated from experimental data by using either of the following equivalent relationships:

$$k = (t_R - t_M)/t_M$$
$$\text{or } k = (V_R - V_M)/V_M$$

The retention factor is a unitless number, where a value of zero indicates that no interaction is occurring between a chemical and the stationary phase. As a chemical undergoes greater interactions with the stationary phase, this will result in longer retention times and an increased value of k. The value of k is directly related to the strength of the interactions that are occurring between a chemical and the stationary phase or mobile phase, as well as the relative amount of stationary phase versus mobile phase that is present in the column.

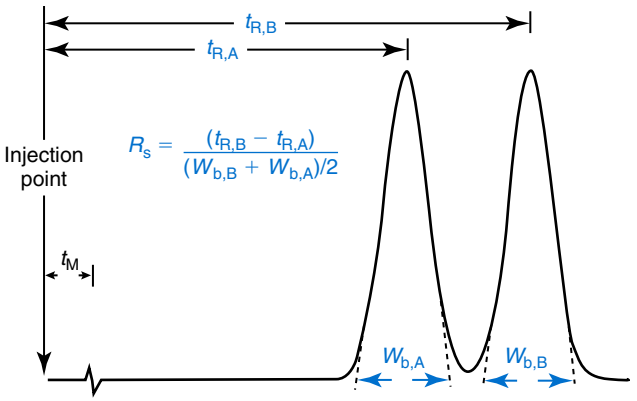


Fig. 12.3 An example of a general chromatogram that may be obtained when using a column. In this example, compound B is eluted later than compound A. R_s, Resolution; t_M, void time; $t_{R,A}$ and $t_{R,B}$, retention times for solutes A and B; $W_{b,A}$ and $W_{b,B}$, baseline peak widths for compounds A and B.

Any separation in chromatography requires that there be some difference in retention for the chemicals of interest. One way of describing this difference is by using the **separation factor** or "selectivity factor" *(α)*. The separation factor for two compounds (A and B) is equal to the ratio of their retention factors (k_A and k_B),

$$\alpha = k_B/k_A$$

where the retention factor for the later eluting component (k_B) is given in the numerator. If two chemicals, A and B, have the same retention in a chromatographic system, the value of α will equal one. If A and B have different retention, the value of α will be greater than 1 and will increase as the degree of separation increases.

Band-Broadening and Efficiency

The peaks for two sequentially eluting chemicals must also be sufficiently narrow to allow a difference in retention to be observed. The peaks observed from the injection of even a small sample will experience some increase in width, or band-broadening, as the chemical travels through a chromatographic system. This broadening is produced by various processes that are related to the rate of movement of the applied chemicals as they pass around or within the support and within or between the mobile phase and stationary phase.

The efficiency and degree of band-broadening in a chromatographic system are reflected by the width of a chemical's peak. This width can be described by using the baseline width (W_b) or the half-height width (W_h) of the peak. These values, in turn, can be used to find another measure of chromatographic efficiency that is known as the **number of theoretical plates** or *plate number (N)*. The value of N can be found for a Gaussian-shaped peak by using either of the following equivalent formulas:

$$N = 16(t_R/W_b)^2 \quad or \quad N = 5.545(t_R/W_h)^2$$

where t_R is the retention time for the peak, and the widths W_b or W_h are in the same units of time as t_R. The value of N can be thought of as representing the effective number of times that a chemical distributes between the mobile phase and stationary phase as the chemical passes through the chromatographic system. A larger value for N represents many such steps, which makes it easier to separate two chemicals that have only small differences in their retention.

Another way to describe the efficiency of a chromatographic system is to use the **height equivalent of a theoretical plate** or "plate height" (HETP or H). The value of H is found by dividing the length *(L)* of the chromatographic system by the number of theoretical plates for the system *(N)*:

$$H = L/N$$

The value of H represents the length of the column or chromatographic system that makes up one theoretical plate, or one distribution step, for a chemical between the mobile phase and stationary phase. A chromatographic system with high efficiency will have a small value for H.

A valuable feature of using H to describe efficiency is that this term can be related directly to the parameters and processes that affect band-broadening. An example of this is the **van Deemter equation**, showing how H is affected by the linear velocity of the mobile phase (u), which is directly related to the flow rate *(F)* through the relationship $u = (F \bullet L)/V_M$.

$$H = A + B/u + Cu$$

The terms A, B, and C in this equation are constants that represent the contributions of several types of band-broadening processes. The A term represents band-broadening processes that are independent of the linear velocity, the B term represents processes that produce more band-broadening at low linear velocities, and the C term represents processes that lead to an increase in H as the linear velocity is increased. The van Deemter equation predicts that the combined effect of all these processes will be an optimum range of flow rates and linear velocities over which the lowest plate heights, and best efficiencies, will be obtained.

Several factors affect chromatographic efficiency. For instance, efficiency can be improved by using longer columns, which increases the value of N but does not alter H. It is also possible to change the flow rate to its optimum value, use a smaller diameter or more efficient supports; or use a relatively narrow diameter capillary instead of a packed bed column. All of these latter factors help lower the value of H, which in turn increases the value of N for a given length of column or chromatographic system.

Resolution

The overall extent to which two peaks are separated in chromatography can be described by using the **resolution (R_s)**, as illustrated in Fig. 12.3. The resolution between two neighboring peaks can be found by using the following formula:

$$R_s = \frac{(t_{R,B} - t_{R,A})}{(W_{b,B} + W_{b,A})/2}$$

In this equation, $t_{R,A}$ and $t_{R,B}$ are the average retention times for compounds A and B (where B elutes after A), while $W_{b,A}$ and $W_{b,B}$ are the baseline widths for the peaks

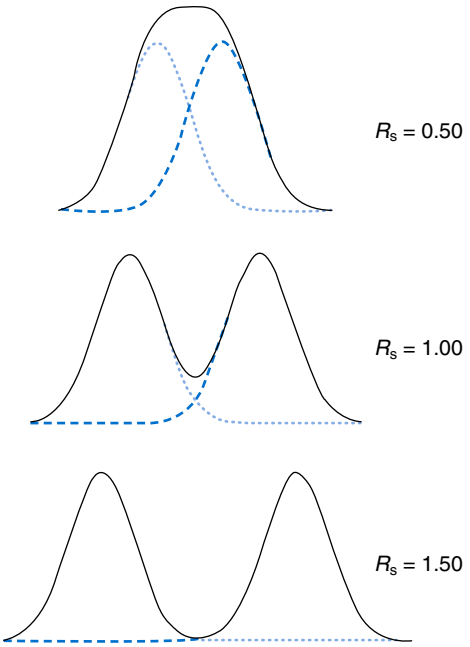


Fig. 12.4 Degree of separation obtained for two chromatographic peaks that are present in a 1:1 area ratio as the resolution between these peaks (R_s) is varied.

of these compounds (in time units, in this case). An equivalent equation can be written in terms of retention volumes and baseline widths in volume units. Either approach will give a unitless value for R_s that represents the average number of baseline widths that separate the centers of the two peaks.

Fig. 12.4 shows how the separation of two neighboring peaks increases as the value of R_s increases for these peaks. An R_s value of zero is obtained when there is no separation between the peaks and they have the exact same retention. An R_s value of 1.5 or greater is often said to represent a complete separation between two equally sized peaks, or "baseline resolution." However, for many separations, resolution values between 1.0 or 1.25 and 1.5 may be adequate, especially if the peaks are about the same size and are to be measured by using their peak heights rather than their peak areas.

There are several approaches that can be used to improve the resolution between two peaks in chromatography (Fig. 12.5). These approaches are indicated by the following expression, which is sometimes known as the *resolution equation* of chromatography:

$$R_s = \left[(N^{1/2})/4\right] \cdot \left[(\alpha - 1)/\alpha\right] \cdot \left[k/(1+k)\right]$$

In this equation, k is the retention factor for the second of two neighboring peaks, α is the separation factor for the peaks, and N is the number of theoretical plates for the chromatographic system. This relationship indicates that peak resolution can be increased in three ways: (1) by increasing the efficiency of the system, as represented by N; (2) by increasing the overall degree of retention, as represented by k; or (3) by increasing the selectivity of the column for one compound versus another, as represented by α.

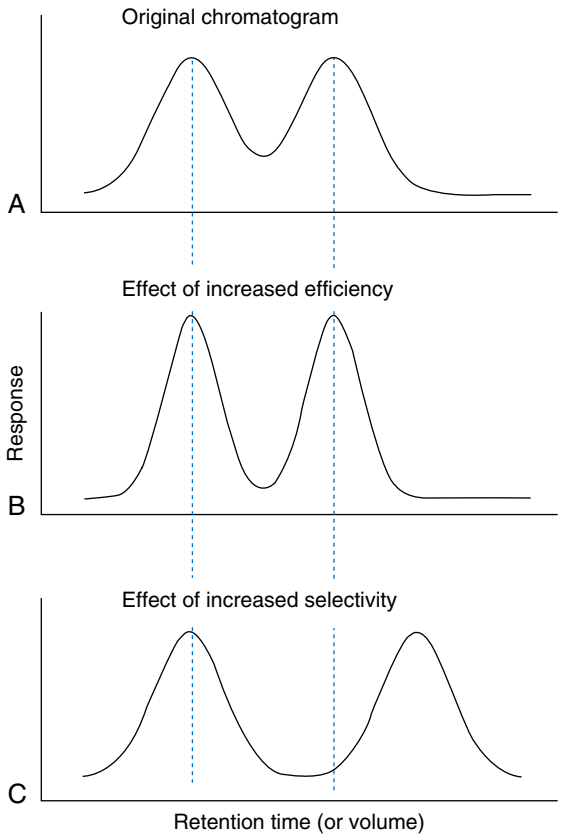


Fig. 12.5 Effects of selectivity and efficiency on the resolution of peaks in chromatography. These three situations represent cases where there is (A) poor or moderate resolution between two neighboring peaks, (B) good resolution between the peaks as a result of high column efficiency, or (C) good resolution between the peaks as a result of good column selectivity.

POINTS TO REMEMBER

General Ways to Improve Peak Resolution in Chromatography

Increase the efficiency of the system.
Increase the overall degree of peak retention.
Increase the selectivity of the column for the peak of one compound versus another.

Analyte Identification and Quantification

Chromatography is often used to identify analytes and/or to measure their concentrations. For example, the retention time, retention volume, and retention factor all reflect how an analyte is interacting within a particular chromatographic system. These retention values can be compared with those for a known sample of the same compound to confirm its identity. An alternative approach for identification is to use a detection method that provides structural information on the analyte, such as mass spectrometry (see Chapter 13).

The peak area or peak height can be used to produce quantitative information on an analyte. Peak areas tend to provide a more precise means for measuring an analyte, while peak heights are easier to use if there is not complete resolution between the analyte and its neighboring peaks. Both external

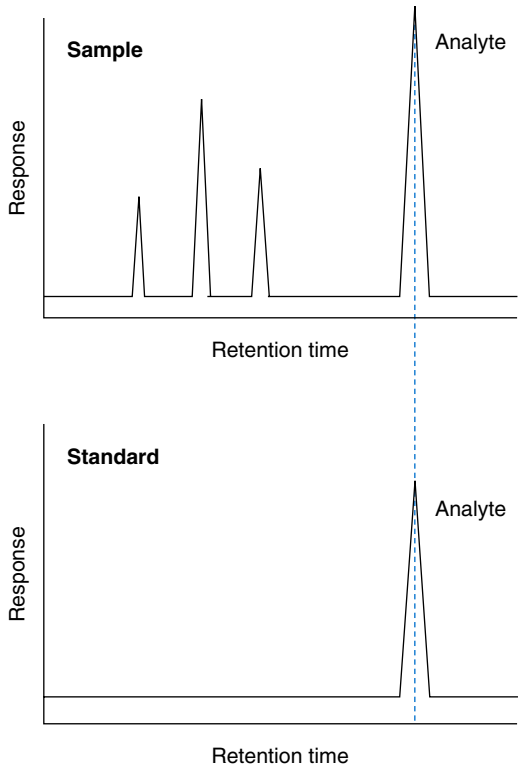


Fig. 12.6 Use of external calibration and standards to quantify an analyte based on its peak height or area in a chromatogram for an injected sample.

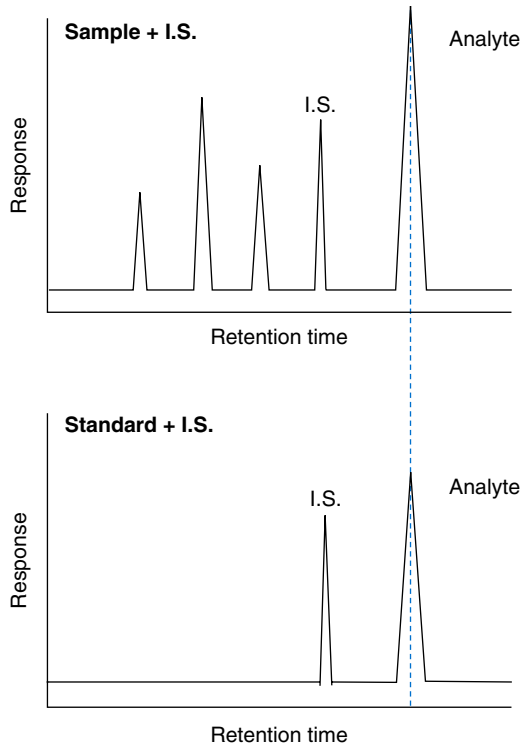


Fig. 12.7 Use of internal calibration and samples or standards containing an internal standard *(I.S.)* to quantify an analyte based on its peak height or area in a chromatogram for an injected sample.

and internal calibration techniques can be used for such measurements. In external calibration, standard solutions containing known quantities of the analytes are processed and separated in the same manner as samples that contain one or more of these analytes (Fig. 12.6). A calibration curve is then constructed by plotting the peak height or peak area (or the spot density, in the case of planar methods) versus the concentration or mass of analyte that was applied in the standard solutions. This curve can then be used with the peak area or peak height that is determined for the same analyte in the samples to find the concentration or the amount of the analyte that was present.

In the method of internal calibration, standard solutions of the analyte are again prepared; however, a constant amount of a different compound known as the **internal standard** is also added to each standard solution and sample (Fig. 12.7). The internal standard should be a chemical that was not originally present in either the sample or the standard, that is similar in its chemical and physical properties to the analyte, and that can be measured independently from the analyte. This internal standard is typically added to the samples and standards before they are processed by any pretreatment steps. The addition of this agent can help normalize the results for any variations that may occur during the pretreatment steps or during injection onto the chromatographic system. This normalization is made by constructing a calibration curve in which the *y*-axis is based on the ratio of the peak height or peak area for the analyte in a given standard or sample divided by the peak height or peak area for the internal standard in the same standard or sample. This ratio is plotted versus the concentration or amount of analyte that was in each standard. This calibration plot can then be used to find the concentration or the amount of the analyte in each sample.

GAS CHROMATOGRAPHY

GC is a type of chromatography in which a gas is used as the mobile phase. This method can be used to separate and analyze compounds that either are naturally volatile or can be converted into a volatile form, such as through derivatization. The mobile phase is typically an inert gas such as nitrogen, helium, or argon, or a low mass gas such as hydrogen. Due to the low densities of gases, the compounds that are injected onto a GC column do not have any appreciable interactions with the mobile phase but are simply carried by this gas through the column. As a result, the term **carrier gas** is commonly used to refer to the mobile phase in GC. Separations in GC are instead based on differences in the vapor pressures of the injected compounds and in the interactions of these compounds with the stationary phase.

Types of Gas Chromatography

Gas-Solid Chromatography

There are several ways of classifying GC methods based on the type of stationary phase that is present. GSC is a type of GC in which the same material acts as both the stationary phase and the support. In this method, chemicals are retained by their adsorption to the surface of the support. This support is often an inorganic material like silica or alumina. Other

TABLE 12.1 Stationary Phases Commonly Used in Gas-Liquid Chromatography and as Bonded Phases in Gas Chromatography

Composition	Polarity	Commercial Examples	Typical Applications
100% Methylpolysiloxane	Nonpolar	OV-1, SE-30	Drugs, amino acid derivatives
5% Phenyl–95% methylpolysiloxane	Nonpolar	OV-23, SE-54	Drugs
50% Phenyl–50% methylpolysiloxane	Intermediate polarity	OV-17	Drugs, steroids, glycols
50% Cyanopropylmethyl–50% phenylmethylpolysiloxane	Intermediate polarity	OV-225	Fatty acid methyl esters, carbohydrate derivatives
Polyethylene glycol	Polar	Carbowax 20M	Acids, alcohols, glycols, ketones

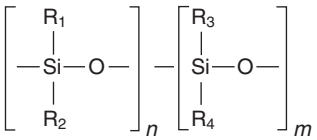


Fig. 12.8 General structure of a polysiloxane. The side groups are represented by R_1 through R_4, while *n* and *m* represent the relative lengths (or amounts) of each type of segment in the overall polymer.

supports that can be used are molecular sieves, which are porous materials that are made from a mixture of silica and alumina, or organic polymers like porous polystyrene. The retention of an analyte on a GSC support will be affected by factors such as the surface area of the support, the size of the pores in this support, and the types of functional groups that are present on the support.

Gas-Liquid Chromatography and Bonded Phases

In GLC the stationary phase is a liquid that is placed as a coating or layer on the support. This is the most common type of GC for chemical analysis. Various types of liquids can be used for this purpose (Table 12.1). Many GLC stationary phases are based on a polysiloxane (Fig. 12.8). The side chains in a polysiloxane can have structures that range from nonpolar methyl groups to polar cyanopropyl groups, and may be present in various ratios or as mixtures. The types of chemicals that will be retained by this type of stationary phase will be determined by the amounts and types of side chains that are present.

One issue in using a liquid as a stationary phase in GC is that some of this liquid will eventually leave the column over time in a process known as "column bleed." This process results in a change in the ability of the GC system to retain chemicals and may cause the signal of a detector to have a high background or to be noisy as the stationary phase leaves the column. Column bleed can be minimized by using a bonded phase (i.e., a stationary phase that is chemically linked or attached to the support) instead of a liquid as the stationary phase in the GC column. The resulting method is sometimes known as **bonded phase gas chromatography**.

POINTS TO REMEMBER

Types of Gas Chromatography Based on the Stationary Phase
Gas-solid chromatography
Gas-liquid chromatography
Bonded phase gas chromatography

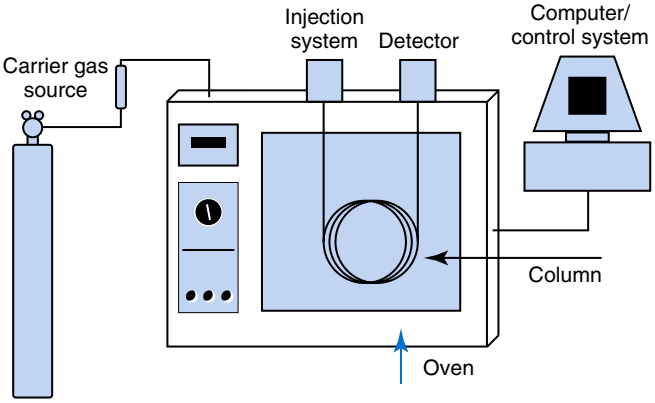


Fig. 12.9 General design of a gas chromatograph. (Modified from a figure courtesy Restek Corporation, Bellefonte, PA.)

Gas Chromatography Instrumentation

Carrier Gas Sources and Flow Control

The typical components of a gas chromatograph are shown in Fig. 12.9. The carrier gas is usually supplied by a standard gas cylinder. However, a gas generator can also be connected to the GC system to isolate nitrogen from air or to produce hydrogen gas through the electrolysis of water. Good flow control is needed to maintain good column efficiency and provide reproducible retention times. The flow can be controlled by using a simple pressure regulator or a more sophisticated electronic device.

The choice of the carrier gas will depend on the expense, purity, and chemical or physical properties of the gas. Carrier gas impurities such as water, oxygen, and hydrocarbons can harm or alter the column and influence the performance of some detectors. Molecular sieves and traps are often used to remove water, hydrocarbons, oxygen, and particulate matter that may be present in the carrier gas.

Many GC detectors work best with certain types of carrier gases. For instance, nitrogen is often used as the carrier gas when working with a flame ionization detector (FID), electron capture detector (ECD), or thermal conductivity detector (TCD), as described in more detail later. Helium is often used with a FID or TCD, and nitrogen/argon-methane mixtures are used with an ECD.

Injection Systems and Sample Derivatization

Most clinical GC methods make use of liquid-phase samples, from which the sample components are first extracted into a nonaqueous liquid or adsorbed onto a microextraction

fiber. This liquid or microextraction fiber is then placed into the GC system by using a precise and rapid online injector (e.g., an autosampler or automated injection system). With packed columns, a glass microsyringe is used to inject a sample through a septum and into a heated injection port. *Volatile* chemicals in the sample and the solvent are quickly vaporized in this heated port and swept into the column by the carrier gas. Common problems during injection include septum leaks and the adsorption of sample components onto the septum. In addition, because the injection port is heated, thermal decomposition may occur and result in spurious peaks in the chromatogram for some compounds.

Because of the low sample capacities and slow carrier gas flow rates that are used with capillary columns, split and splitless injection techniques are used to introduce samples into such columns. For a split injection, only a small portion of the vaporized sample enters the column, with the remainder being passed through a side vent. In splitless injection, most of the sample enters the column. The split flow injection mode is used for samples that contain relatively high concentrations of the analytes, while the splitless mode is used for samples that contain lower concentrations. Temperature-programmable injection ports are also available and may be used in either injection mode to reduce the time that thermally unstable compounds are exposed to high temperatures in the port.

Headspace analysis is a sample introduction technique that can be used with aqueous solutions or samples that contain some nonvolatile components. In this method, a portion of the vapor phase (or "headspace") that is above a liquid or solid sample is used for the analysis. This vapor phase contains some of the more volatile sample components and can be directly injected onto a GC system. Headspace analysis can be carried out by using either a static method or a dynamic method. In the static method, the sample is placed in an enclosed container and allowed to reach equilibrium for the distribution of its components between the sample and the vapor phase above the sample. A portion of the vapor phase is then injected onto the GC system. In the dynamic method, an inert gas is passed through the sample and used to sweep away the volatile components. These components are then captured by a solid adsorbent or a cold trap and later injected onto the GC system.

Although a fairly large number of low-mass chemicals can be injected directly onto a GC system, many more are not sufficiently volatile or thermally stable for direct injection. A way to make a chemical more volatile and thermally stable is to alter its structure through derivatization. This usually involves replacing one or more polar groups on the analyte with less polar groups. A common example is the replacement of an active hydrogen on an alcohol, phenol, amine, or carboxylic acid group with a trimethylsilyl (TMS) group, producing a TMS-derivative. Other examples include the use of alkylation (e.g., the formation of a methyl ester through the esterification of a carboxylic acid) or acylation (e.g., the production of an acetate derivative from an alcohol or amine). Some derivatization reactions can also be used to change the response of an analyte to certain detectors, such as an ECD through the addition of halogen atoms to a compound's structure.

Columns and Supports

Both packed columns and capillary columns are used in GC. Packed GC columns are filled with support particles that contain the stationary phase and have typical inner diameters of 1 to 4 mm and lengths of 1 to 2 m. Packed GC columns are useful when it is necessary to apply a relatively large amount of a sample onto the GC system. However, packed columns also tend to have lower efficiencies than capillary columns, so they are mainly used in applications for which a relatively small number of compounds are to be separated.

Capillary columns, or open-tubular columns, consist of a column that has the stationary phase attached to or coated on its interior surface. Capillary columns have typical inner diameters of 0.1 to 0.75 mm and lengths that often range from 10 to 150 m. Capillary GC columns are usually made from fused silica capillaries that have a polyimide or aluminum coating on the outside to give the capillary sufficient strength and flexibility for use in a GC system. Although capillary columns have lower sample capacities than packed columns, they provide better peak resolution and higher efficiencies. These properties make capillary columns the most common type of support that is used in GC for analytical applications.

Temperature Control

All GC systems require careful control of temperature with a column oven to obtain optimal performance and reliable results. The column may be maintained at a constant temperature (i.e., **isothermal elution**), or the temperature may be varied over time (i.e., **temperature programming**). Temperature programming is used for most clinical applications. In temperature programming, the sample components with the lowest boiling points and weakest interactions with the GC column will elute first, followed by chemicals that have higher boiling points and/or stronger interactions with the column. Temperature programming usually provides sharper and more distinct peaks in less time than can be obtained with isothermal elution. The main advantage of isothermal elution is that it can be faster for samples that do not contain a wide range of chemicals with different volatilities.

The thermal stabilities of the stationary phase and column are important to consider during the development of a GC method. Because each stationary phase has a specific temperature range over which it is stable, it is necessary to keep the column temperature within this usable range. Before any GC column is used for a routine analysis, it must be "thermally conditioned" by heating the column at various temperatures and for given lengths of time. This process helps remove volatile contaminants that may be initially present in the column or that have accumulated and can lead to unstable baselines. Preconditioned capillary columns are also available to minimize these problems.

Gas Chromatography Detectors

Flame ionization detector. A variety of detectors can be used in GC systems (Table 12.2). An FID is commonly used with GC in clinical laboratories for the analysis of ethanol and other volatiles in blood or aqueous samples (Fig. 12.10). During the operation of an FID, the carrier gas that is leaving the column is mixed with hydrogen, and the eluting

TABLE 12.2 Examples of Detectors Used in Gas Chromatography

Type of Detector	Principle of Operation	Selectivity	Approximate Limit of Detection
Flame ionization detector (FID)	Production of gas phase ions from combustion of organic compounds	General—Organic compounds	10^{-12} g Carbon
Nitrogen-phosphorus detector (NPD)	Heated alkali bead selectively ionizes N- or P-containing compounds	Nitrogen- or phosphorus-containing compounds	10^{-14}–10^{-13} g Nitrogen or phosphorus
Electron capture detector (ECD)	Capture of electrons by chemicals with electronegative groups	Chemicals with electronegative groups	10^{-15}–10^{-13} g
Mass spectrometry (MS)	Production of gas phase ions, followed by separation/analysis of these ions based on their mass-to-charge ratios	Universal—full-scan mode Selective—selected ion monitoring mode (SIM)	10^{-10}–10^{-9} g Full-scan mode 10^{-12}–10^{-11} g SIM mode
Thermal conductivity detector (TCD)	Measurement of change in thermal conductivity of carrier gas as compounds elute from the column	Universal	10^{-9} g
Flame photometric detector (FPD)	P- and S-containing compounds emit light when burned in a flame; emitted light is detected	Phosphorus and sulfur-containing compounds	10^{-12} g Phosphorus 10^{-11} g Sulfur

Portions of this table are based on data from Hage, D. S., & Carr, J. D. (2011). *Analytical chemistry and quantitative analysis*. New York: Pearson, and references cited therein.

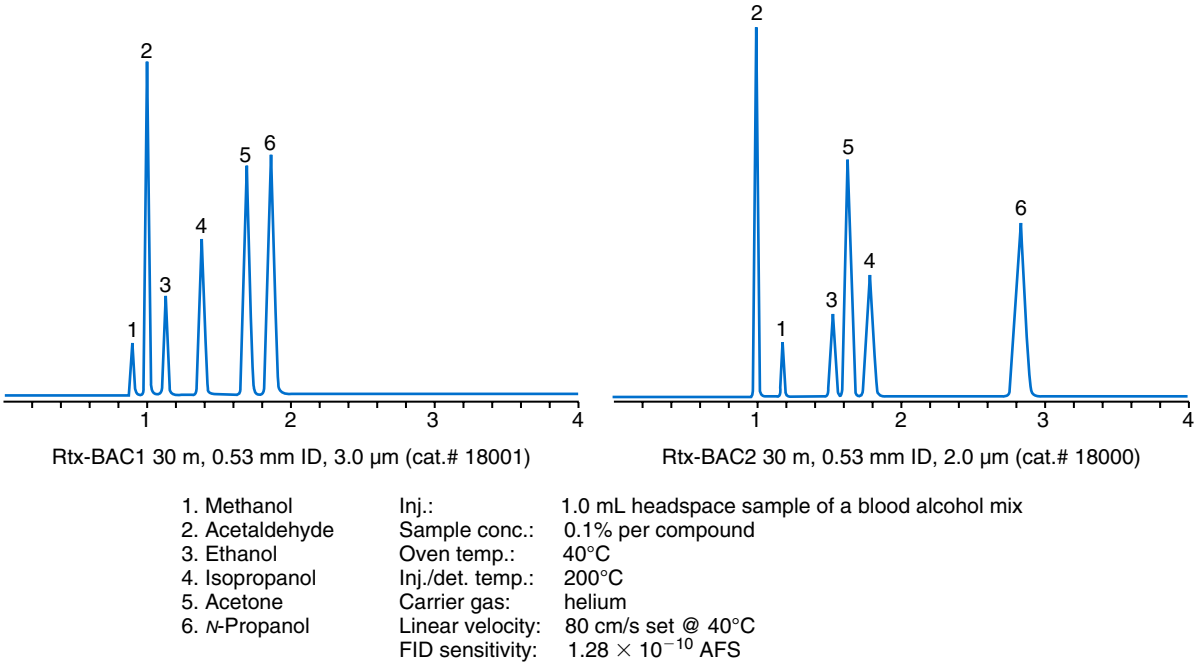


Fig. 12.10 Chromatograms obtained during the analysis of volatile organic compounds when using headspace analysis and gas chromatography with a flame ionization detector *(FID)*. (Courtesy Restek Corporation, Bellefonte, PA.)

compounds are burned by a flame that is surrounded by air and an oxygen-rich environment. This process produces gas-phase ions from organic compounds, with these ions being detected and measured by an electrode that is positioned above the flame.

The advantages of an FID include its simplicity, reliability, and ease of operation. Another advantage of an FID is that this detector gives little or no signal for common carrier gases (e.g., He, Ar, or N_2) or typical contaminants in such gases (e.g., O_2 and H_2O). An FID is easy to use with temperature programming and is a good general detector for the routine clinical analysis of organic compounds. One disadvantage of the FID is its destructive nature, so it cannot be connected directly to other GC detectors.

Nitrogen-phosphorus detector. The nitrogen-phosphorus detector (NPD) is similar to an FID but instead creates gas-phase ions by using an electrically heated alkali bead that is generally made of rubidium. This heated bead is placed directly above the point where a mixture of the carrier gas and hydrogen enter the detector. Ions are generated at or above the surface of the heated alkali bead, which supplies electrons to electronegative compounds and leads to the formation of negatively charged ions, which are then measured at an electrode.

Nitrogen- or phosphorus-containing compounds are especially good at creating ions in an NPD. This feature makes the NPD useful for monitoring low concentrations of analytes that have nitrogen or phosphorus in their structures, such as many organic bases and acids. This type of detector does not respond to common GC carrier gases or their impurities. However, it is necessary to regularly change the alkali bead in this detector because this bead will degrade over time.

Electron capture detector. The ECD is a selective GC detector based on the capture of secondary electrons by electronegative compounds that are eluting form the column. High-energy electrons, or beta particles, are provided in an ECD by a radioactive source such as ^{63}Ni or ^{3}H. These beta particles collide with the carrier gas and lead to the release of a large number of secondary electrons. When only the carrier gas is passing through this detector, a consistent supply of the secondary electrons is created, which is collected at an electrode and measured. When a chemical with electronegative groups elutes from the column, some of these secondary electrons are captured and fewer reach the electrode, producing a change in the signal.

An ECD can provide both selective and sensitive detection for chemicals that contain electronegative groups. This includes chemicals that contain halogen atoms (I, Br, Cl, and F) or nitro groups ($-NO_2$). It also includes chemicals that are polynuclear aromatic hydrocarbons, anhydrides, or conjugated carbonyl compounds, among others. Derivatization with reagents containing chlorinated or fluorinated groups can be used with some chemicals to allow them to be monitored with an ECD. Large carrier gases such as argon and nitrogen are usually employed with an ECD because it is easy for them to collide with beta particles and to produce secondary electrons. Some methane is also usually combined with these carrier gases to produce a stable detector response. Because an ECD uses a radioactive source, this source needs to be replaced on a regular basis by a certified technician.

Mass spectrometry. Mass spectrometers are also used as detectors for GC. This combination is known as **gas chromatography/mass spectrometry (GC/MS)**. GC/MS is a powerful method for identifying and quantifying analytes. Ionization methods that are often used in GC/MS include electron impact ionization (EI) and chemical ionization (CI). Mass analyzers that are commonly used in GC/MS are quadrupole mass analyzers and ion traps (see Chapter 13).

In the "full-scan mode" of GC/MS, information is acquired on a wide range of ions. This mode is used when the goal is to detect many compounds in a single run or to provide data that can be used to identify an unknown compound from its mass spectrum. GC/MS can also examine specific analytes by using selected ion monitoring (SIM) to monitor only a few ions that are representative of the analytes of interest. SIM is used when selective detection and low detection limits are desired.

Other gas chromatography detectors. There are a variety of other detectors that can also be part of a GC system. A TCD is a general detector that can monitor both inorganic and organic compounds based on their ability to change how the carrier gas/analyte mixture will conduct heat away from a hot wire. The carrier gas used with a TCD is often helium or hydrogen, which have large differences in their ability to conduct heat when compared with most organic or inorganic compounds. The primary advantage of a TCD is its ability to detect many types of chemicals. This detector is nondestructive and can easily be combined with a second detector. However, a TCD can also respond to contaminants in the carrier gas and give a change in the background response during temperature programming. In addition, the TCD tends to have much higher detection limits than other common GC detectors.

Another group of GC detectors make use of the interactions of chemicals with light. One example is a flame photometric detector (FPD). An FPD is a selective detector for phosphorus- or sulfur-containing compounds. Like an FID, an FPD passes the eluting chemicals into a flame, but the FPD then measures the release of light from excited-state phosphorus- or sulfur-containing species.

Data Acquisition and System Control

Computers are used to control and automate GC systems, as well as to collect and process data. The computer can regulate parameters such as the (1) carrier gas composition and flow rate; (2) column back pressure; (3) temperature; (4) sample injection process; (5) detector selection and operation; and (6) timing steps during system operation. In terms of data acquisition, the computer can monitor and store the signals that are generated by the GC system's detectors. The computer system can also be used to search databases to aid in the identification of analytes based on their retention times or their response at the detector.

LIQUID CHROMATOGRAPHY

Liquid chromatography (LC) is a type of chromatography in which the mobile phase is a liquid. Separations in LC are based on the distribution of chemicals between a liquid mobile phase and a stationary phase. LC is the dominant type of chromatography that is currently used for chemical analysis in clinical or biomedical laboratories. A key advantage of this method over GC is the ability of LC to work directly with liquid samples, as are often encountered with clinical or biological specimens.

The supports originally used in LC columns were based on large and irregularly shaped particulate supports like those used in packed columns for GC. These supports were useful in preparative work but were not suitable for many analytical applications because they tended to result in broad peaks and separations with low resolution. Work in the 1960s resulted in smaller, more mechanically stable and more efficient supports for LC, along with the instrumentation that could be used with such materials. This resulted in a method that is now known as **high-performance liquid chromatography (HPLC)**. The use of these more efficient supports made it possible to obtain narrower peaks, better separations, and lower limits of detection than traditional LC. These reasons, along with the ability of HPLC to be used as an automated method, have made this technique the method of choice for most routine chemical separations and analysis methods.

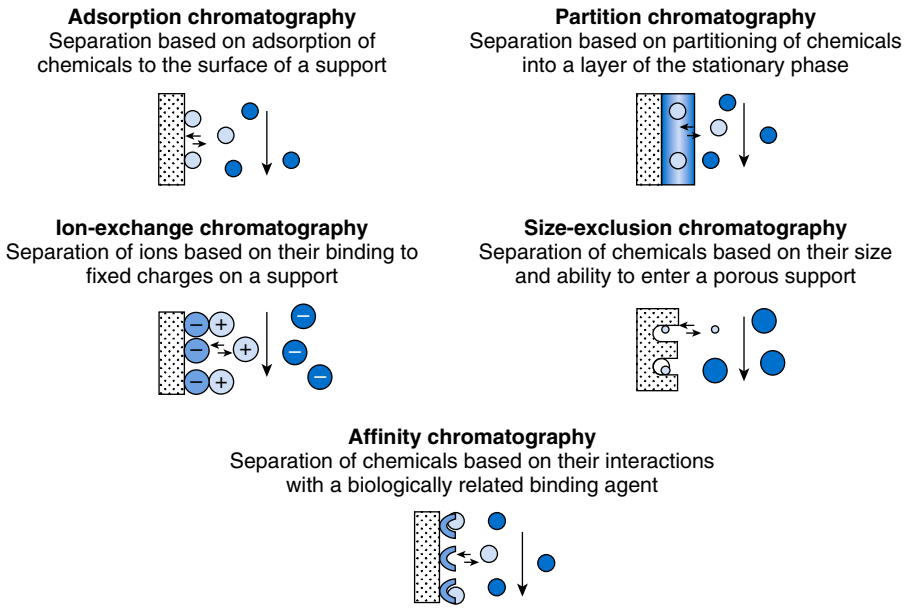


Fig. 12.11 Main types of liquid chromatography based on their separation mechanisms.

Because the pressure drop across a packed bed column increases with a decrease in the support particle's diameter, relatively high pressures can be required to pump liquids through HPLC columns. As a result, this technique has also been referred to as *high-pressure liquid chromatography*. Most modern HPLC systems can work at pressures up to 5000 to 6000 psi. Specialized systems that use even smaller diameter supports and that can operate at pressures up to 15,000 psi have recently been developed resulting in the technique called **ultra-high-performance liquid chromatography (UPLC or UHPLC)**.

One important difference between LC and GC is that the retention of chemicals in LC can depend on the interactions of these chemicals with both the mobile phase and stationary phase. This means the composition of the mobile phase is important to consider when adjusting the retention of a chemical in LC. The term *strong mobile phase* is used to describe a mobile phase that leads to weak retention for an analyte on a given type of stationary phase. The weakest retention for a chemical will occur when this substance favors staying in the mobile phase instead of the stationary phase. The term *weak mobile phase* refers to the opposite situation, in which a chemical favors the stationary phase versus the mobile phase and has its highest retention within a given column.

POINTS TO REMEMBER

Mobile Phase Strength in Liquid Chromatography

The mobile phase is important in liquid chromatography for determining chemical retention.

A "strong mobile phase" is a mobile phase that leads to weak retention for an analyte on a given type of stationary phase and column.

A "weak mobile phase" is a mobile phase in which a chemical favors the stationary phase and has a high retention on a given column.

Types of Liquid Chromatography

Adsorption Chromatography

Liquid chromatographic methods are classified according to the mechanisms by which they separate chemicals (Fig. 12.11). **Adsorption chromatography** is a type of LC in which chemicals are retained based on their adsorption and desorption at the surface of a support that is not derivatized (see Fig. 12.11). This method is sometimes referred to as *liquid-solid chromatography* (LSC). Retention in this method is based on the competition of the analyte with the mobile phase as both bind to the support. The degree of a chemical's retention in adsorption chromatography will depend on (1) the binding strength of this chemical to the support, (2) the surface area of the support, (3) the amount of mobile phase that is displaced from the support by the chemical, and (4) the binding strength of the mobile phase to the support.

Three types of adsorbents are generally used in adsorption chromatography: (1) polar acidic supports, (2) polar basic supports, and (3) nonpolar supports. The most common polar and acidic support is silica, which can adsorb polar compounds and works particularly well for basic substances. Alumina is a polar and basic adsorbent that also retains polar compounds, and especially polar acidic substances. Florisil is a polar and basic support that can be used in place of alumina. Nonpolar supports include charcoal and polystyrene.

Partition Chromatography

Partition chromatography is an LC method in which solutes are separated based on their partitioning between the liquid mobile phase and a stationary phase that is coated or bonded onto a solid support (see Fig. 12.11). The support in most types of partition chromatography is silica. This method originally involved coating a liquid stationary phase onto such a support. However, most current columns in partition

chromatography employ bonded stationary phases, which are more stable than coated stationary phases and provide better column efficiencies.

There are two main types of partition chromatography: normal-phase chromatography and reversed-phase chromatography. **Normal-phase chromatography** uses a polar stationary phase and is also known as *normal-phase liquid chromatography* (NPLC). The stationary phase in this method typically contains groups that can form hydrogen bonds or undergo dipole-related interactions. Examples of these bonded stationary phases are those that contain aminopropyl groups, cyano groups, and diol groups. Because this type of chromatography uses a polar stationary phase, it will have its highest retention for polar compounds. A weak mobile phase in this method will be a nonpolar liquid. A strong mobile phase is a more polar liquid, such as methanol or water.

Normal-phase chromatography can be used in many of the same applications as adsorption chromatography based on silica or alumina supports. These applications usually involve the separation or analysis of chemicals that are present in organic solvents and of substances that contain one or more polar functional groups. Examples of chemicals for which normal-phase chromatography has been used include steroids and sugars.

The second type of partition chromatography is **reversed-phase chromatography**, which is also known as *reversed-phase liquid chromatography* (RPLC). Reversed-phase chromatography uses a nonpolar stationary phase and is the most popular type of LC. One reason for this is that the weak mobile phase in reversed-phase chromatography is a polar solvent, such as water. This property makes this type of LC convenient for the analysis and separation of chemicals in aqueous-based systems, such as serum, urine, and blood. A strong mobile phase in this method is a liquid that is less polar than water, such as acetonitrile or methanol. Due to the presence of a nonpolar stationary phase, nonpolar compounds will have the highest retention in reversed-phase chromatography.

There are many applications for reversed-phase chromatography in clinical chemistry and biomedical research (see the example in Fig. 12.2). Examples of chemicals that have been separated or analyzed by this method include drugs, drug metabolites, amino acids, peptides, proteins, carbohydrates, lipids, and bile acids. It is important to consider both the type of analyte and support that are being used in these separations. For instance, large chemicals such as peptides or proteins will require reversed-phase supports with larger pore sizes than the supports that are used for separating small molecules.

The most common stationary phases used in reversed-phase chromatography are based on octadecyl (C_{18}), octyl (C_8), phenyl, or butyl (C_4) groups that are attached to a support such as silica. Alternative materials such as porous graphite, fluorinated hydrocarbons, and hydrophobic stationary phases with embedded polar groups can also be used. Silica tends to dissolve slowly at a pH greater than 8 or at a pH below 2, so separations that make use of silica supports are usually done in this pH range. Some of the other supports used in reversed-phase chromatography, such as polystyrene or porous graphite, are stable over a broader pH range (e.g., pH 2 to 13).

Ion-Exchange Chromatography

Ion-exchange chromatography (IEC) is a type of LC in which ions are separated by their adsorption onto a support that contains fixed charges at its surface (see Fig. 12.11). Depending on the charge of the groups that make up the stationary phase, the types of ions that bind to the column may be either cations (i.e., positively charged ions) or anions (i.e., negatively charged ions). These two methods are referred to as **cation-exchange chromatography** and **anion-exchange chromatography**, respectively.

Supports for cation-exchange chromatography contain negatively charged groups. These groups may be the conjugate bases of strong acids, such as sulfonate ions that are formed by the deprotonation of sulfonic acid, or the conjugate bases of weak acids, such as those produced from carboxyl or carboxymethyl groups. The supports used in anion-exchange chromatography are usually the conjugate acids of strongly basic quaternary amines, such as triethylaminoethyl groups, or the conjugate acids of weak bases, such as aminoethyl (AE) or diethylaminoethyl (DEAE) groups. Supports that can be modified to contain these charged groups include silica, polystyrene, and carbohydrate-based materials such as agarose or cellulose.

A strong mobile phase in IEC is usually a mobile phase that contains a high concentration of competing ions. A weak mobile phase is one that contains few or no competing ions, or that otherwise promotes binding by charged analytes to the column. The retention of ions in this method may also be affected by (1) pH, (2) the type of competing ion that is used, (3) the type of fixed charges that are used as the stationary phase, and (4) the density of these fixed charges on the support.

IEC has a number of clinical applications. Examples are the separation and analysis of amino acids and hemoglobin variants. IEC is also frequently used as a preparative tool for purifying proteins, peptides, and nucleotides. A modified form of IEC known as *ion chromatography* can be used with a conductivity detector to analyze small inorganic and organic ions.

Size-Exclusion Chromatography

Size-exclusion chromatography (SEC) is an LC technique that separates analytes based on size (see Figs. 12.11 and 12.12). In this method, a porous support is used that has an inert surface with little or no interactions with the sample components. This support should also have a range of pore sizes that are similar to the sizes of the compounds that are to be separated. As a sample travels through this type of column, small sample components can enter all or most of the pores while larger components may enter only a few or none of the pores. The result is a separation based on size or molar mass, in which the larger components elute first from the column.

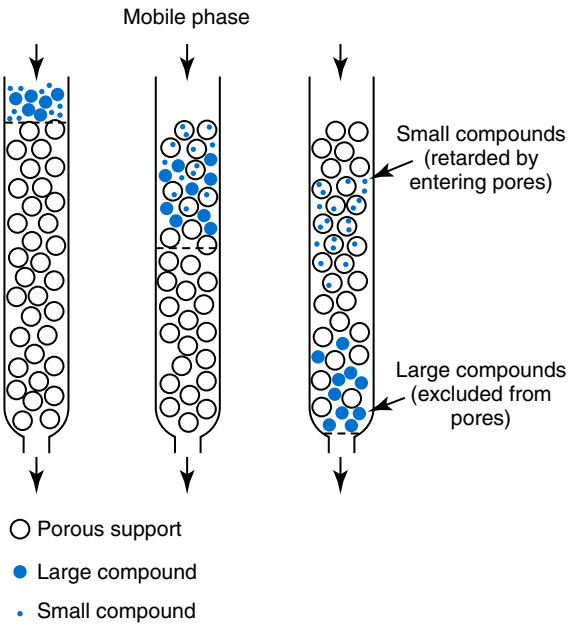


Fig. 12.12 General principle of size-exclusion chromatography. This method separates compounds based on their size by using a column that contains a porous support. (Modified from Bennett, T. P. [1968]. *Graphic biochemistry, vol. 1: chemistry of biological molecules*. New York: Macmillan.)

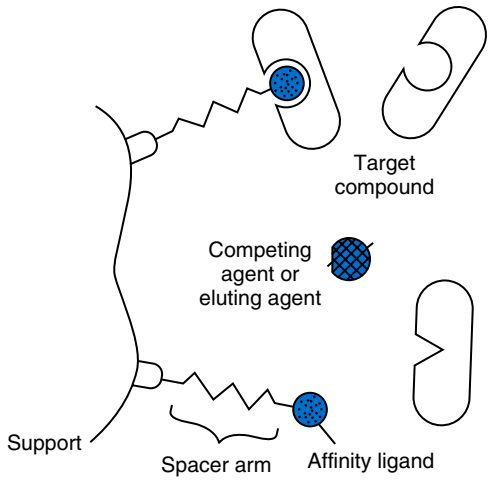


Fig. 12.13 General mechanism of separation in affinity chromatography. The target analyte is first allowed to bind to the immobilized affinity ligand in the column. After the nonretained sample components have been washed from the column, the retained analyte is later released by using an elution buffer. This elution may be based on a nonspecific method (e.g., the addition of a chaotropic agent or a change in pH), or it may be accomplished by adding a biospecific competing agent to the mobile phase to displace the analyte from the column.

Many types of porous supports have been used in SEC. Cross-linked carbohydrate-based supports like dextran and agarose are often used for work with aqueous-based samples and biological compounds such as proteins or nucleic acids. Polyacrylamide gel and modified silica or glass beads can also be used for aqueous samples and biological compounds. Polystyrene is usually employed as the support when SEC is to be used with synthetic polymers and samples that are present in organic solvents. For each of these supports, the range of pore sizes that are present will determine the sizes of the injected compounds that can be separated.

The mobile phase in SEC can be a polar solvent, such as water, or a nonpolar solvent, which is usually tetrahydrofuran. Because the stationary phase is based on a physical difference in the accessible pore volume for the solutes, rather than being based on chemical interactions, there is no weak mobile phase or strong mobile phase in this method. The choice of the mobile phase is instead determined by the solubility and the stability of the analytes and of the support and the column. If water or an aqueous mobile phase is used in SEC, the resulting method is often called *gel filtration chromatography*. If an organic mobile phase is used, the size-exclusion method is referred to as *gel permeation chromatography*.

SEC in the form of gel filtration is useful for identifying intact complexes of agents such as lipoproteins, antibody-antigen complexes, and the binding of proteins with their target compounds. SEC is also often used to exchange buffers, to remove salts from large sample components, or to remove small molecules from large biomolecules (e.g., the isolation of drugs, fatty acids, and peptides from proteins). In addition, SEC can be used to estimate the molecular weight of a biomolecule, such as a protein or nucleic acid, or to characterize the distribution of molecular weights for a polymer.

Affinity Chromatography and Chiral Separations

Affinity chromatography (AC) is an LC method that makes use of biologically related interactions for the retention and separation of chemicals (see Figs. 12.11 and 12.13). This method uses the selective, reversible interactions that are found in many biological systems, such as the binding of an antibody with an antigen or the interactions of an enzyme with a substrate or inhibitor. This type of binding is used in AC by immobilizing one of a pair of interacting compounds onto a support for use as the stationary phase. This immobilized agent is called the *affinity ligand* and creates a column that can bind and capture the complementary compound from samples.

AC is usually carried out by applying a sample to the column under conditions that allow strong and specific binding of the affinity ligand to its target compound. This is done in the presence of an application buffer, which mimics the pH and natural or preferred conditions for binding between the affinity ligand and the target. As the target binds to the column under these conditions, most other sample components are washed away. The retained target is then later released by using an elution buffer that either contains a competing agent that will displace the target from the affinity ligand or that uses a change in conditions such as pH, ionic strength, or polarity of the mobile phase to decrease the strength of binding between the target and affinity ligand. After the target has been eluted for detection or further use, the application buffer can be passed again through the column and the system is allowed to regenerate prior to the application of the next sample. For systems with weak-to-moderate binding strengths, it is also possible to use isocratic elution for both

sample application and target elution. This second type of elution is usually employed in chiral separations.

There are various types of AC. Bioaffinity chromatography is the most common type of AC and involves the use of a biological binding agent as the affinity ligand. Examples of bioaffinity chromatography are the purification of an enzyme by using an immobilized inhibitor, and the use of lectins (i.e., nonimmune system proteins that can bind carbohydrate residues) such as concanavalin A for the isolation of glycopeptides, glycoproteins, and glycolipids. Immunoaffinity chromatography (IAC) is a type of bioaffinity chromatography that uses an antibody or antibody-related agent as the affinity ligand. The selectivity of this method has made it popular for the isolation and analysis of many targets, as well as a tool for sample pretreatment prior to analysis by other methods.

Another group of methods in AC are those that use nonbiological binding agents. Immobilized metal-ion affinity chromatography (IMAC) is a method in which the affinity ligand is a metal ion, such as Ni^{2+}, that is complexed with an immobilized chelating agent. This technique is frequently used to isolate recombinant histidine-tagged proteins and has been used for the isolation or analysis of phosphorylated proteins in proteomics. Boronate affinity chromatography uses an affinity ligand that is boronic acid or a related derivative, which is useful in binding to compounds that contain *cis*-diol groups such as polysaccharides, glycoproteins, and catecholamines. An important clinical application of boronate AC is its use in the analysis and isolation of glycosylated hemoglobin in blood from diabetic patients.

Many biological molecules (including amino acids, peptides, and proteins) occur as stereoisomers, or in specific chiral forms. As a result, it is not unusual for the different chiral forms of a drug to have some variations in their interactions with these biological agents. The two mirror-image chiral forms of a drug (or enantiomers) generally have identical physical and chemical properties and cannot be separated by most types of chromatography, which tend to use non-stereoselective (or achiral) stationary phases. However, it may be possible to separate the enantiomers and stereoisomers of a drug or target compound if the stationary phase is also chiral and can interact with these compounds in a stereospecific manner. This type of medium is known as a **chiral stationary phase (CSP)**. Carbohydrates, peptides, enzymes, and other proteins have all been used as CSPs. Cyclodextrins, which are cyclic polymers of glucose, are an important set of carbohydrates that have been used for separating many types of chiral compounds in both LC and GC.

Hydrophilic Interaction Liquid Chromatography and Mixed-Mode Methods

There are a number of LC methods that combine several separation modes. One example is **hydrophilic interaction liquid chromatography (HILIC)**. HILIC is a type of partition chromatography that uses a polar stationary phase, and in which chemicals partition between an organic-rich region in the mobile phase and a more polar water-enriched layer that is at or near the surface of a polar support. The surface of the support, which can often undergo hydrogen bonding or dipole-related interactions with the applied solutes, may also have charged groups that can interact with these compounds. These features make HILIC a variation of normal-phase chromatography that is combined with some of the characteristics of reversed-phase chromatography and IEC. HILIC and related methods have become popular in areas such as proteomics and glycomics. Advantages of these methods include (1) their ability to give better separations for polar compounds than can be obtained by reversed-phase chromatography, (2) the greater ease with which they can be used with aqueous samples and in solubilizing polar compounds compared with normal-phase chromatography, and (3) the ability to couple these methods with mass spectrometry.

Ion-pair chromatography (IPC) is a technique that combines columns that are used in reversed-phase chromatography with the ability to separate ionic compounds based on their charges, as is done in IEC. This method is carried out by adding an ion-pairing agent to the mobile phase for a reversed-phase column. The ion-pairing agent is usually a surfactant that has a charged group at one end and a nonpolar tail or group at the other end. The purpose of the ion-pairing agent is to combine with ions of the opposite charge in the sample, thus also allowing their retention on the reversed-phase column through the ion-pairing agent's nonpolar tail or through the formation of a neutral complex. IPC is useful in separating charged compounds that are poorly resolved by IEC. Applications of IPC include the separation and analysis of catecholamines, drugs, and nucleic acids.

Liquid Chromatography Instrumentation

Mobile Phase Reservoirs and Delivery Systems

The major components of a system used in HPLC are shown in Fig. 12.14. The mobile phases in LC are contained in solvent reservoirs, which are often glass bottles or flasks into which "feed lines" to the pump are inserted. Filters are often placed at the inlets of these feed lines to prevent any particles in the mobile phase from moving on to the rest of the system. Most mobile phase reservoirs also have a means of "sparging" the mobile phase by bubbling through a gas such as helium or nitrogen to remove dissolved air or oxygen that may interfere with the response of some detectors. The removal of air and oxygen, or "degassing," can also be achieved by applying a vacuum to the reservoir by placing gas exchange devices or gas filters in the flow path leading from the reservoir. Degassing the mobile phase is important to avoid dissolved gases coming out of solution and forming bubbles as the pressure drops from the column to the detector.

The composition and "strength" of the mobile phase are factors that can be used to adjust and control a separation in LC. If the same mobile phase is used throughout the separation, this approach is known as **isocratic elution**. If the composition of the mobile phase is varied over time, the method is called **solvent programming** or gradient elution (Fig. 12.15). Solvent programming begins with a weak mobile phase, to allow chemicals with weak retention to interact with the column,

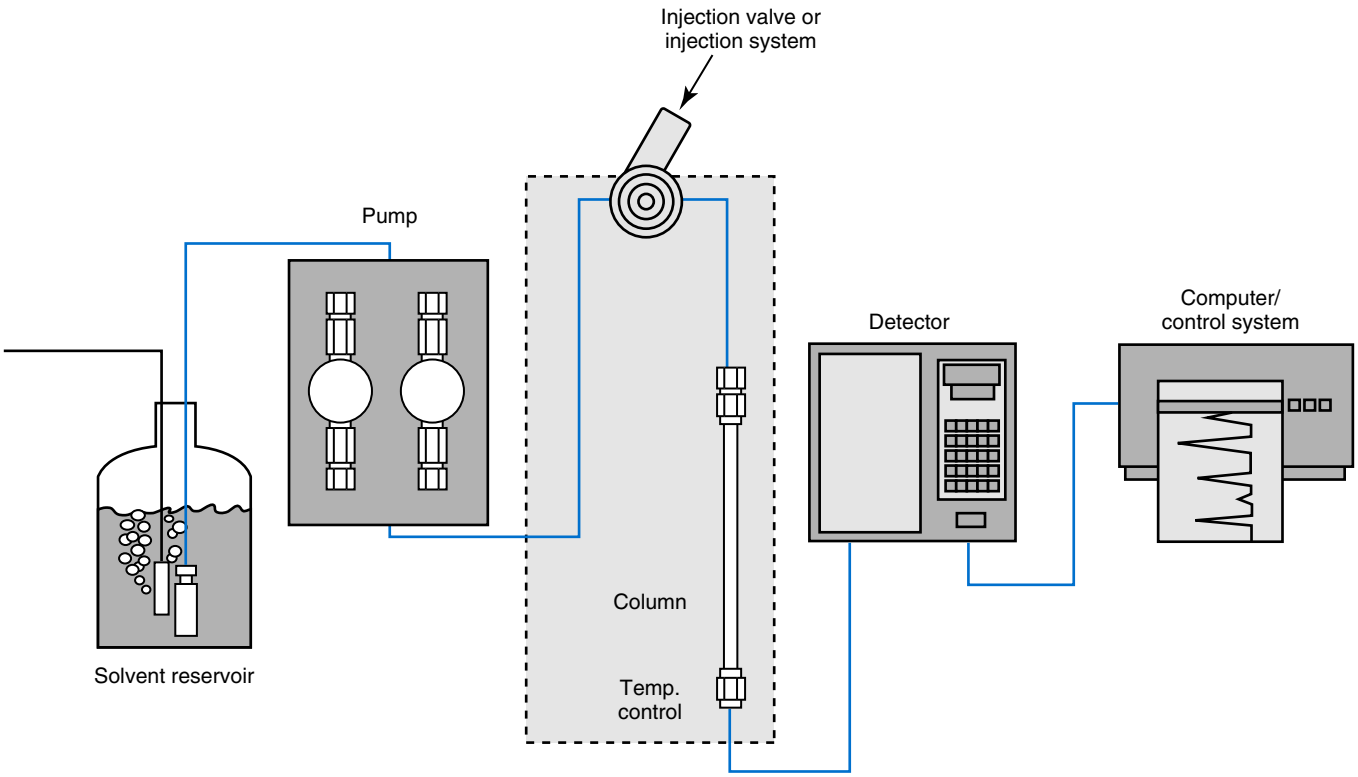


Fig. 12.14 General design of a liquid chromatograph, as used in high-performance liquid chromatography. (Modified from a figure courtesy Restek Corporation, Bellefonte, PA.)

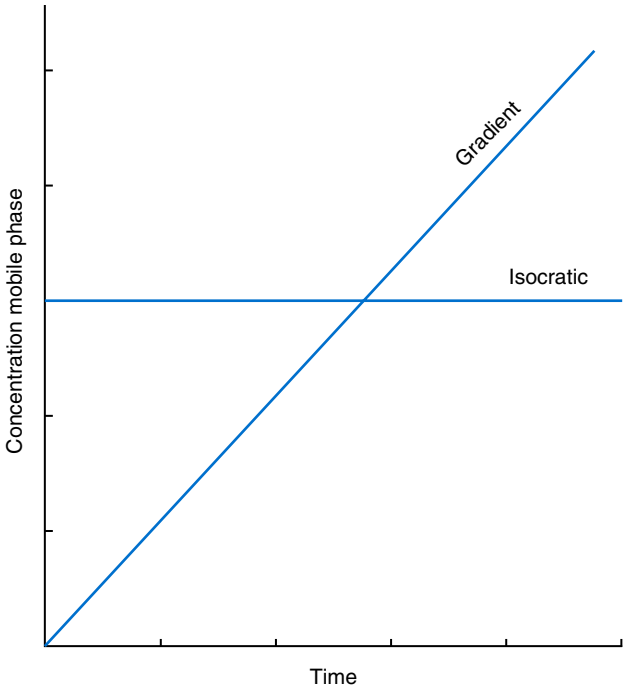


Fig. 12.15 Examples of isocratic elution (i.e., constant mobile phase composition) or linear gradient elution based on solvent programming (i.e., varying mobile phase composition).

and then goes to a stronger mobile phase to elute chemicals with moderate or high retention. This change in the mobile phase can be made in one or more steps and may involve a linear or nonlinear change over time.

Several types of pumps have been used in LC. Peristaltic and diaphragm-type pumps can be used with columns that are operated at low pressures, as are encountered in classic and low-to-medium performance LC. Reciprocating pumps and syringe pumps are used to achieve the higher pressures that are needed to deliver the mobile phase in HPLC. Reciprocating pumps are used at flow rates in the mL/min range and have a piston that moves in and out of the solvent chamber, with check valves being used to keep the flow of the mobile phase moving from the pump inlet to the outlet. The reciprocating action of the piston does generate some pulsation in the pressure and mobile phase flow, which can increase the baseline noise at the detector but can be minimized by electronic control of the pump and by placing pulse dampers in the flow path. Syringe pumps use a solvent chamber in a syringe to continuously deliver the mobile phase to the rest of the system. These pumps can deliver essentially pulse-free flow and can be used at much lower flow rates than reciprocal pumps (e.g., flow rates in the μL/min range).

Injection Systems and Sample Derivatization

Various approaches can be used to introduce a sample into an LC system. Important characteristics to consider when selecting an injection system are its (1) reproducibility, (2) the amount of sample carryover from one injection to the next, and (3) the range of volumes that can be injected. The most widely used approach in HPLC is a fixed-loop injector that is switched into or out of the flow path by manual control or through the use of an autoinjector.

TABLE 12.3 Typical Column Sizes Used in Analytical High-Performance Liquid Chromatography

Type of Column	Typical Inner Diameter (I.D.) and Lengths	Typical Flow Rate Range
Conventional packed column	4–5 mm I.D. × 5–30 cm	1–3 mL/min
Narrow bore column	2–3 mm I.D. × 5–15 cm	0.2–0.6 mL/min
Microbore column	1–2 mm I.D. × 10–100 cm	0.05–0.2 mL/min
Packed capillary	0.1–0.5 mm I.D. × 20–200 cm	0.1–20 μL/min
Open tubular column	0.01–0.075 mm I.D. × 1–100 cm	0.05–2 μL/min

Portions of the data in this table are based on Poole, C. F., & Poole, S. K. (1991). *Chromatography today*. New York: Elsevier.

Derivatization is sometimes used in LC to improve the response of compounds to the detector. It is also possible to use derivatization in LC to alter the retention of a compound on the column. There are two main ways of carrying out derivatization in LC: (1) precolumn derivatization and (2) postcolumn derivatization. Precolumn derivatization is done before the sample is injected and can be used to alter a compound's retention or to increase the signal it generates at the detector. Postcolumn derivatization is carried out online as compounds elute from a column and is used only to improve the response of the detector to those compounds.

Columns and Supports

The columns in LC can have various combinations of packing materials and diameters or lengths. In the clinical laboratory, most packed HPLC columns are made of tubes based on stainless steel or pressure-resistant polymers with internal diameters that range from 4 to 5 mm and lengths that range from 5 to 30 cm (Table 12.3). An inlet filter may be present to remove particulate matter, and a short guard column that contains the same packing material may be placed before the longer analytical HPLC column to protect and extend the usable life of the HPLC column.

Better efficiencies and lower detection limits are achieved with HPLC columns that have longer lengths and smaller inner diameters, including narrow bore columns (with approximate inner diameters of 2 to 3 mm) and microbore columns (with approximate inner diameters of 1 to 2 mm). These small-diameter columns may also require lower flow rates and smaller volumes of the mobile phase for their operation than the longer conventional packed columns. Capillary columns may also sometimes be used in LC for work with flow rates in the mid- to low μL/min range.

The most common type of support in LC is a packed bed of small particles, which are usually porous supports with diameters in the range of 1.8 to 10 μm. A smaller diameter for these supports, as is used in UPLC, provides better efficiency but also leads to an increase in back pressure across the column. Low to medium performance separations, which have much lower operating pressures than HPLC, typically use packing materials such as cross-linked dextran or agarose, which consist of particles with diameters of 50 to 200 μm.

When using porous support particles, the mobile phase flows around the support but not through the particle. This means compounds must travel within the particle by means of diffusion, which is a relatively slow process that can be a major source of band-broadening. The distance that these compounds must diffuse can be reduced by using a nonporous support or a pellicular support, in which the latter has a thin porous layer or porous shell. Another approach for minimizing this band-broadening is to use perfusion particles. This support has small pores that contain most of the stationary phase and larger pores that allow the mobile phase to pass both through and around the support particles. Another alternative support that can be used to improve efficiency is a monolithic support, which consists of a continuous porous bed that is usually prepared from silica or an organic polymer.

Temperature Control

The control of column temperature can be an important factor in determining the reproducibility and efficiency of an LC separation. Temperature control of an LC column can be achieved by using temperature-controlled column chambers, water jackets, blankets, and heating/cooling blocks.

Liquid Chromatography Detectors

There are many types of detectors that can be used in LC (Table 12.4). A key component for most of these detectors is the flow cell through which the mobile phase and eluting compounds from the column must pass to provide a response. A postcolumn reactor may also be present between the column and detector to derivatize some of the eluting compounds and to generate products that give a stronger and more specific signal at the detector.

Absorbance detectors. The absorption of ultraviolet (UV) or visible (vis) light is often used to detect compounds as they elute from an LC column. Many of the absorbance detectors that are used in LC can measure the absorption of UV light wavelengths in the range of 190 to 400 nm or of visible light in the range of 400 to 700 nm. Many organic compounds with aromatic groups or double or triple bonds absorb light between 250 and 300 nm. Other organic compounds, such as those containing amide bonds or carboxylic acids, can absorb in the range of 190 to 220 nm. In addition, some ions, inorganic compounds, and metal complexes can be detected by their absorption of light in the UV or visible range.

There are several types of absorbance detectors that can be used in LC. Fixed-wavelength absorbance detectors have the simplest design and are used to monitor absorbance at a particular wavelength or wavelength band. A variable-wavelength absorbance detector has a more flexible design that can operate at a wavelength that is selected from a broad wavelength range. A photodiode array detector (PDA) is an absorbance detector that uses an array of small detector cells to measure the change in absorbance at many wavelengths simultaneously, which can

TABLE 12.4 Examples of Detectors Used in Liquid Chromatography

Type of Detector	Principle of Operation	Range of Application	Detection Limit
Absorbance detector	Measures absorbance of light at a given wavelength or set of wavelengths	Compounds with chromophores that can absorb ultraviolet or visible light	10^{-10}–10^{-9} g
Fluorescence detector	Measures ability of chemicals to absorb and reemit light through fluorescence	Compounds with fluorophores	10^{-12}–10^{-9} g
Electrochemical detector	Measures current or charge as a result of chemical oxidation or reduction	Electrochemically active compounds	10^{-11}–10^{-9} g
Conductivity detector	Measures change in conductivity of the mobile phase as ions elute from the column	General for ionic solutes	10^{-9} g
Refractive index detector	Measures change in refractive index of the mobile phase as compounds elute the column	Universal	10^{-7}–10^{-6} g
Mass spectrometry	Production of gas phase ions, followed by separation/analysis of these ions based on their mass-to-charge ratios	Universal—full-scan mode Selective—selected ion monitoring (SIM) mode	10^{-10}–10^{-9} g (Full-scan mode) 10^{-12} g or lower (SIM mode)

Portions of the data in this table are based on Poole, C. F., & Poole, S. K. (1991). *Chromatography today*. New York: Elsevier.

be valuable in identifying overlapping peaks or in identifying analytes in complex samples such as urine and serum.

The use of an absorbance detector requires solvents and mobile phase components that have little or no absorption of light at the wavelengths of interest. Effective degassing of the mobile phase is important to minimize bubble formation, which can interfere with absorbance measurements. The use of some back pressure across the detector can further reduce bubble formation. However, care must be taken with this approach to avoid exceeding the usable pressure range of the detector.

Fluorescence detectors. Fluorescence detectors measure the ability of a chemical to absorb light at one wavelength and reemit the light at a different, longer wavelength. Fluorescence detectors in LC are generally much more selective and have better limits of detection than absorbance detectors for chemicals that are naturally fluorescent or that can be converted into a fluorescent derivative. Both precolumn and postcolumn derivatization have been used to modify chemicals (e.g., amino acids, peptides, and proteins) for this type of detector.

Electrochemical and conductivity detectors. The combination of an electrochemical detector with LC is often known as **liquid chromatography–electrochemical detection (LC-EC)**. In an amperometric electrochemical detector (see Chapter 10), an electroactive chemical that enters the flow cell may be oxidized or reduced at an electrode that is held at a constant potential. The current needed for or generated by this process is then measured, with the magnitude of the current being proportional to the concentration of the analyte. Electroactive compounds of clinical interest and that can be examined by this process include catecholamines, ascorbic acid, and thiol-containing compounds such as homocysteine. In addition, electrochemically active tags (e.g., bromine) can be added to compounds such as unsaturated fatty acids or prostaglandins for use with this type of detector.

Coulometric detectors measure the amount of charge that is required for a given electrochemical reaction. These detectors are selective, sensitive, and have reasonably wide linear ranges. Coulometric detectors are used in clinical laboratories for the analysis of metanephrines, vanillylmandelic acid, homovanillic acid, and 5-hydroxyindole acetic acid in human urine.

A conductivity detector measures the ability of the mobile phase and its contents to conduct a current when they are placed in an electrical field. This type of detector is often used in combination with IEC to monitor salt gradients, or in ion chromatography to monitor the elution of charged analytes. Conductivity detectors have been used to measure compounds such as sulfate in biological fluids.

Refractive index detectors. A refractive index (RI) detector in LC measures the change in the refraction of light as chemicals pass within the mobile phase through a flow cell. An important advantage for this type of detector is that it can monitor substances such as alcohols, salts, and sugars that do not absorb light at usable wavelengths or fluoresce. One disadvantage of this type of detector is it does not have limits of detection as low as absorbance or fluorescence detectors, and it has a response that can be sensitive to changes in the mobile phase composition and temperature.

Mass spectrometry. LC can be combined with mass spectrometry, giving a combined technique known as **liquid chromatography–mass spectrometry (LC-MS)**. This combination makes it possible to both measure chemicals and to identify them based on the masses of their molecular ions or fragment ions. When used in the full-scan mode, the mass spectrometer in LC-MS acts as a general detector. If the mass spectrometer is instead used for looking at particular ions, this device then acts as a selective detector. It is also possible to add another dimension based on mass spectrometry, as occurs in liquid chromatography-tandem mass spectrometry (LC-MS/MS), to look at fragment ions that are produced from a given precursor ion. This addition can further increase

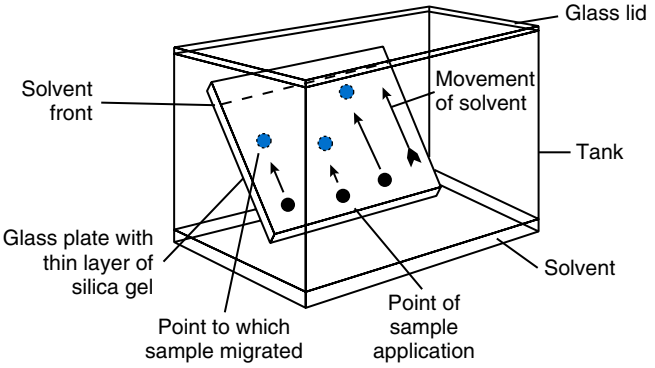


Fig. 12.16 General operation and system components of thin-layer chromatography. In this particular example, the mobile phase moves up the glass plate and a thin layer of adsorbent by means of capillary action. (Modified from Bennett, T. P. [1968]. *Graphic biochemistry, vol. 1: chemistry of biological molecules.* New York: Macmillan.)

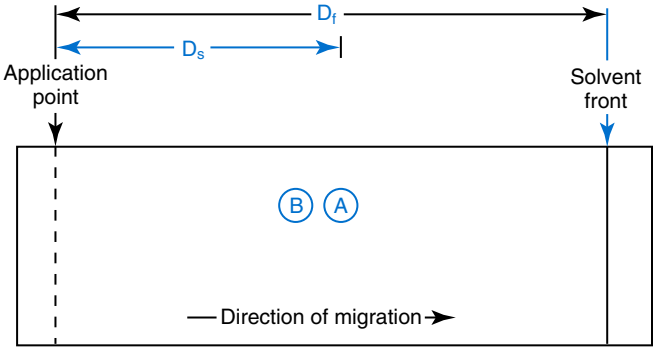


Fig. 12.17 General example of a separation obtained by planar chromatography. In this example, compound *B* is more strongly retained and migrates a shorter distance than compound *A*. D_s, Distance travelled by an analyte (A, in this example) from the point of sample application; D_f, distance travelled by the mobile phase, or solvent front, from the point of sample application in the same amount of time as allowed for sample migration.

the ability of the method to resolve or detect multiple components and can enable the analysis of hundreds or thousands of components in a sample.

Several types of ionization methods and mass analyzers can be employed in LC-MS (see Chapter 13). A common combination is the use of electrospray ionization with a quadrupole mass analyzer. Other possible ionization methods that can be used are CI or photoionization. The interface in LC-MS has to remove solvent from the mobile phase and place the sample components in a charged form in the gas phase that can be analyzed by the mass spectrometer. This process requires that the buffers used in LC-MS be sufficiently volatile to avoid overloading and contaminating the interface.

Data Acquisition and System Control

The system controller for an HPLC will often manage (1) sample injections, (2) solvent delivery, (3) the system flow rate and temperature, as well as (4) control the detectors, and (5) acquire data from these detectors. Data acquisition systems can collect thousands of data points from an individual run. These data can then be used to identify and measure the eluting compounds based on the retention times and areas or heights of their peaks. The hardware and software that are used for data analysis can become particularly important as the amount of collected data becomes large, as can often occur with the use of a PDA or LC-MS. Libraries of spectra or other databases can be searched to aid in the identification of chemicals based on the chromatograms and signals that are generated during the separation.

PLANAR CHROMATOGRAPHY

In planar chromatography, the stationary phase is coated or placed onto a flat surface, or plane. The sample is added as a small spot or band on this surface. This support is then placed into an enclosed container with one edge in contact with the mobile phase and the sample band located just beyond this point of contact (Fig. 12.16). The mobile phase is usually allowed to travel across the plane by means of capillary action. After this movement has occurred for a given period of time, the support is removed from the mobile phase and dried prior to the analysis or measurement of the separated sample components.

The planar surface that is used in this method may be a sheet of paper, known as *paper chromatography*, or some other type of surface, resulting in a method known as *thin-layer chromatography* (TLC). In paper chromatography, the stationary phase is the surface of the paper or a layer of a solvent or chemical that is coated or placed onto the paper. In TLC, a thin layer of particles (such as silica or alumina) is spread on a glass plate, plastic sheet, or aluminum sheet.

In planar chromatography, retention is described by the distance that compounds have traveled in a given amount of time (Fig. 12.17), rather than the time or volume of mobile phase required for elution, as is used in column chromatography. The distance traveled by a chemical from its point of application (D_s) in paper chromatography and TLC is often compared with the distance that has been traveled by the mobile phase in the same amount of time (D_f). This information can be used to calculate a measure of retention that is known as the **retardation factor (R_f)**.

$$R_f = D_s/D_f$$

The value for R_f will always be between zero and one. Chemicals that have high retention with the stationary phase will have low values R_f, and chemicals that have low retention will have R_f values that approach one. Chemicals can be identified in planar chromatography by comparing their retention and R_f values with those for reference compounds that have been examined on the same stationary phase and support and mobile phase. The detection characteristics of the reference compounds can also be compared with the chemicals in the unknown sample (e.g., the color of their bands or their response to a color-forming reagent). Additional confirmation can be obtained by comparing the unknown compound and the reference compound under a different set of separation conditions.

The separated components in planar chromatography can often be detected by their natural color, by their response to

UV light or through their fluorescence, or by their visualization with reagents that form colored products. In some cases, these chemicals may be allowed to react with labeled antibodies for their detection, or they may be detected by using radiolabels and autoradiography. Their bands may also be removed from the planar surface for analysis by a method such as mass spectrometry or nuclear magnetic resonance spectroscopy.

Paper chromatography and TLC tend to be used primarily for qualitative analysis. One application of these techniques is their use in the analysis of amniotic fluid to determine lecithin-to-sphingomyelin ratios. Another application is in the screening of urine for drugs or metabolites such as amino acids that accumulate during hereditary disorders.

REVIEW QUESTIONS

1. Which component in a chromatographic system carries the sample?
   a. Stationary phase
   b. Mobile phase
   c. Support
   d. Column
2. Which of the following is *not* an approach that can be used to improve the separation of two compounds by a chromatographic system?
   a. Increase the degree of peak retention
   b. Increase the selectivity of the system for one compound versus another
   c. Increase the amount of sample that is applied
   d. Increase the efficiency of the system
3. In column chromatography, unknown analytes are identified as they pass through the detector. Comparison of the _____ of unknown analytes with that of standard compounds can be used for identification of the analytes.
   a. separation factor
   b. retention factor
   c. resolution
   d. plate height
4. Before it can be used in a separation by GC, a compound must:
   a. be volatile or be converted into a volatile form.
   b. not be volatile.
   c. be water soluble.
   d. be present in a large quantity in a sample.
5. Which type of GC might use alumina particles as the stationary phase?
   a. Gas-liquid chromatography
   b. Gas-solid chromatography
   c. Bonded phase gas chromatography
   d. Capillary gas chromatography
6. Which of the following common detectors for GC can be used in more than one mode to either detect many compounds or examine specific analytes?
   a. Nitrogen-phosphorus detector
   b. Electron capture detector
   c. Mass spectrometer
   d. Flame ionization detector
7. Which of the following statements concerning LC is *not* correct?
   a. HPLC is the form of LC used in most routine chemical separation or analysis methods.
   b. Normal-phase chromatography uses a nonpolar stationary phase and is one of the most popular types of LC.
   c. Cation-exchange chromatography uses negatively charged groups as the stationary phase.
   d. A porous support is used in size-exclusion chromatography.
8. Some point-of-care tests use chromatography as a processing step in the assay. In point-of-care pregnancy testing, for example, the stationary phase is nitrocellulose paper and the hCG antigen is present in a urine sample, which is considered the mobile phase. The antigen in the mobile phase binds with an antibody in the stationary phase. This is an example of planar chromatography using which of the following as the separation mechanism?
   a. Partition chromatography
   b. Ion-exchange chromatography
   c. Size-exclusion chromatography
   d. Affinity chromatography
9. Which statement is true about the mobile phase in LC?
   a. The strength of the mobile phase is varied during solvent programming.
   b. A strong mobile phase gives strong retention for compounds on an LC column.
   c. An organic solvent is a weak mobile phase in reversed-phase chromatography.
   d. The mobile phase has few interactions with the injected compounds in LC.
10. Which of the following statements is *not* true about planar chromatography?
   a. Compounds may be identified by using their retardation factors.
   b. Paper may be used as a support in this method.
   c. This method is mainly used for quantitative analysis.
   d. Thin-layer chromatography is one form of this method.

SUGGESTED READINGS

Allenmark, S. (1991). *Chromatographic enantioseparations: methods and applications* (2nd ed.). New York: Ellis Horwood.

Drozd, J. (1981). *Chemical derivatization in gas chromatography*. Amsterdam: Elsevier.

Ewing, G. W. (1997). In *Analytical instrumentation handbook* (Ed.). (2nd ed.). New York: Marcel Dekker.

Gross, M. L., Caprioli, R. M., & Niessen, W. (2006). *The encyclopedia of mass spectrometry, vol. 8: hyphenated methods*. Amsterdam: Elsevier, 1–1068.

Hage, D. S. (Ed.). (2005). *Handbook of affinity chromatography* (2nd ed.). Boca Raton: CRC Press.

Hage, D. S., & Carr, J. D. (2011). *Analytical chemistry and quantitative analysis*. New York: Pearson.

Janson, J. C., (Ed.). *Protein purification: principles, high resolution methods, and applications*. (3rd ed.). Hoboken: Wiley.

Jorgenson, J. W. (2010). Capillary liquid chromatography at ultrahigh pressures. *Annual Review of Analytical Chemistry*, *3*, 129–150.

Karger, B. L., Snyder, L. R., & Horvath, C. (1973). *An introduction to separation science*. New York: Wiley.

Lough, W. J., & Wainer, I. W. (1995). *High performance liquid chromatography: fundamentals principles and practice*. New York: Blackie Academic.

Lunn, G., & Hellwig, G. C. (1998). *Handbook of derivatization reactions for HPLC*. New York: Wiley-Interscience.

Majors, R. E. (2003). A review of HPLC column packing technology. *America Laboratory*, *3*(10), 46–54.

McNair, H. M., & Miller, J. M. (2009). *Basic gas chromatography* (2nd ed.). Malden, Mass: Wiley Interscience.

Miller, J. M. (2009). *Chromatography: concepts and contrasts* (2nd ed.). Malden, Mass: Wiley Interscience.

Ravindranath, B. (1989). *Principles and practice of chromatography*. New York: Wiley.

Sherma, J., & Fried, B. (Eds.). (2003). *Planar chromatography*. New York: Taylor & Francis.

Shushan, B. (2010). A review of clinical diagnostic applications of liquid chromatography-tandem mass spectrometry. *Mass Spectrometry Reviews*, *29*, 930–944.

Snyder, L. R., Kirkland, J. J., & Dolan, J. W. (2009). *Introduction to modern liquid chromatography* (3rd ed.). New York: Wiley.

Svec, F., & Huber, C. G. (2006). Monolithic materials: promises, challenges, achievements. *Analytical Chemistry*, *78*, 2100–2108.

Taylor, L. T. (2010). Supercritical fluid chromatography. *Analytical Chemistry*, *2008*(80), 4285–4294.

13

Mass Spectrometry

*Mark M. Kushnir, Nigel J. Clarke, Alan L. Rockwood**

OBJECTIVES

1. Define the following terms:
   - Base peak
   - Ion trap
   - Mass analysis
   - Mass spectrometry
   - Mass-to-charge ratio (*m/z*)
   - Molecular ion
   - Product ion
   - Fragment ion
   - Quadrupole
   - Triple quadrupole
   - Tandem mass spectrometry
   - Time of flight (TOF)
   - Ion trap
   - Matrix-assisted laser desorption/ionization (MALDI)-TOF
2. Describe the main components of a mass spectrometer (MS), their functions, and principles of operation.
3. For the following ionization methods, state the principle and specific uses of each type:
   - Electron
   - Chemical
   - Electrospray
   - Atmospheric pressure chemical
   - Inductively coupled plasma (ICP)
   - MALDI
4. State the principles of operation of quadrupole, triple quadrupole, ion trap, and TOF mass analyzers.
5. List types of commonly used electron multiplier detectors and state the principle of this type of detection.
6. Explain the role of mass spectrometry in clinical laboratories.

KEY WORDS AND DEFINITIONS

Base peak An ion with the highest abundance in the mass spectrum; it is assigned a relative abundance of 100%.

Electrospray ionization Using this ionization technique, a sample is passed through a capillary to which a high voltage is applied, resulting in expulsion of ions from a spray of fine droplets.

Extracted ion chromatogram A plot consisting of a sum of selected mass-to-charge ratios displayed as a function of time.

Fragment ion An ion formed by fragmentation of a molecular ion.

Gas chromatography–mass spectrometry (GC-MS) A technique that uses a gas chromatograph coupled to a mass spectrometer.

Internal standard A structural analog of a substance that is added to calibrators, controls, and unknown samples at a constant amount. An internal standard corrects for variable recovery between samples during the sample preparation and analysis steps. The internal standard should have similar physical and chemical properties to the targeted molecule, with a distinct mass spectrometry signal. Internal standards typically contain stable isotopes of atoms.

Ion An atom or a molecule that has acquired an electrical charge by losing or gaining one or more electrons or other charged species such as protons.

Ionization Required for all mass spectrometry techniques; "ionization" describes the production of ions from neutral atoms or molecules using various techniques including but not limited to electron ionization, chemical ionization, or electrospray ionization.

Ion trap Type of mass analyzer in which ions are held in a spatially confined region using, for example, a magnetic or electrostatic field. Manipulation of the trap allows m/z measurements to be performed, as well as fragmentation of the molecular ion to allow multiple levels of the fragmentation (MSn).

Isotope A variant of a chemical element; each variant differs in the number of neutrons in the nucleus and therefore in atomic weight.

*The authors gratefully acknowledge the original contributions by Thomas M. Annesley and Nicholas E. Sherman on which portions of this chapter are based.

Isotope dilution mass spectrometry (IDMS) An analytical technique used to quantify a compound relative to isotopic analogs added to samples as internal standards at fixed concentrations.

Liquid chromatography–mass spectrometry (LC-MS) A technique that uses a liquid chromatograph coupled to a mass spectrometer.

Mass-to-charge ratio (*m/z*) The ratio formed by dividing the molecular mass of an ion by its charge.

Mass analysis The process of resolving ionic species based on their mass-to-charge ratios.

Mass spectrometer An analytical instrument that ionizes atoms or molecules; separates the ions and detects the *m/z* of these ionized atoms or molecules or their ion fragments. Mass spectrometers are often interfaced with other instruments including but not limited to a second mass spectrometer, a liquid chromatograph, or a gas chromatograph.

Mass spectrometry (MS) Study of matter through the formation of gas-phase ions that are characterized by their mass, charge, structure, and/or physicochemical properties.

Mass spectrum A plot of the relative abundance of ions plotted as a function of their mass-to-charge ratio.

Matrix-assisted laser desorption/ionization (MALDI) A soft ionization technique allowing analysis of organic molecules (proteins, peptides, DNA, polysaccharides), and microorganisms. MALDI ionization is performed by irradiation of sample mixed with a matrix using pulsed laser.

Product ion A fragment ion formed when a molecular ion breaks into smaller pieces. In a tandem mass spectrometer, the fragmentation process takes place after molecular ions have been separated by their m/z value in the first stage.

Selected ion monitoring (SIM) A mass spectrometry technique where signal from only specified ions of interest are acquired.

Tandem mass spectrometer A mass spectrometer capable of multistage separation of ions in an ion beam based on the *m/z* values (tandem in space) or separation of trapped ions according to *m/z* values (tandem in time).

Torr A non-SI unit of pressure with 760 torr equal to 1 standard atmosphere. One torr is essentially equal to the pressure exerted by a millimeter high column of mercury.

Total ion chromatogram (TIC) A chromatogram created by summing up intensities of all *m/z* responses acquired by a mass analyzer during chromatographic separation. The chromatogram is displayed as a function of time.

BASIC CONCEPTS

A **mass spectrometer** is an analytical instrument that ionizes molecules and then separates and measures the **mass-to-charge ratio (*m/z*)** of the ionized molecules and their fragments. **Mass analysis** is the process of identification of the *m/z* ratios of the ions. This analysis may be qualitative, or it may be both qualitative and quantitative. **Mass spectrometry** (MS) is useful for identifying and determining elemental composition and structure of compounds, as well as for performing quantitative analysis of components of samples.

A **mass spectrum** is represented by the relative abundance of **ions** plotted as a function of their *m/z* ratio (Fig. 13.1). When an ion has a single charge ($z = 1$) the *m/z* ratio is equal to the mass of the ion. The unfragmented ion of a molecule is called the *molecular ion*. **Fragment ions** are formed when a molecular ion breaks into smaller pieces. The ion with the highest abundance in the mass spectrum is assigned a relative value of 100% and is called the **base peak**. The base peak may be the molecular ion or a fragment ion. Fragmentation of ions at specific bonds depends on their chemical nature; in some cases it is possible to determine the structure of a molecule from the fragment ions present in the mass spectrum. Because patterns of relative abundance among the fragment ions of a given compound are often unique to that compound, computer-based libraries of spectra take advantage of this for identification of the analytes from the mass spectra.

In **tandem mass spectrometry** fragmentation takes place after ions have been separated by their *m/z* values in a first stage of mass spectrometry; these fragment ions are named **product ions**.

Mass spectrometers are often considered to be "universal detectors" because most molecules can be ionized and can therefore be detected by a MS. When interfaced with a liquid or a gas chromatograph, the MS often provides highly specific detection and structural information. In such instruments multiple mass spectra are acquired across the chromatographic peaks, and the sum of all the ions detected across the peak displayed as a function of time represents a **total ion chromatogram (TIC)**. It is possible to program the data system to display the sum of ions over a restricted mass range rather than the full mass range or multiple *m/z* values; the resultant chromatogram is called an ***extracted ion chromatogram*** (EIC; Fig. 13.2). Both TICs and EIC are displayed as signal intensity plotted as a function of time.

For quantitative analysis when only a few analytes are of interest, and their mass spectra are known, the MS is typically setup to monitor only the ions of interest. This is known as *selected ion monitoring* (SIM). The MS cycles repeatedly and rapidly through a list of selected *m/z* values, monitoring the abundance of each *m/z* for a specified time. Because SIM focuses only on specified ions, more ion signal is collected for each selected *m/z*, as compared with acquisition of the full mass spectrum, increasing the signal-to-noise ratio for the targeted analytes and improving the lower limit of detection.

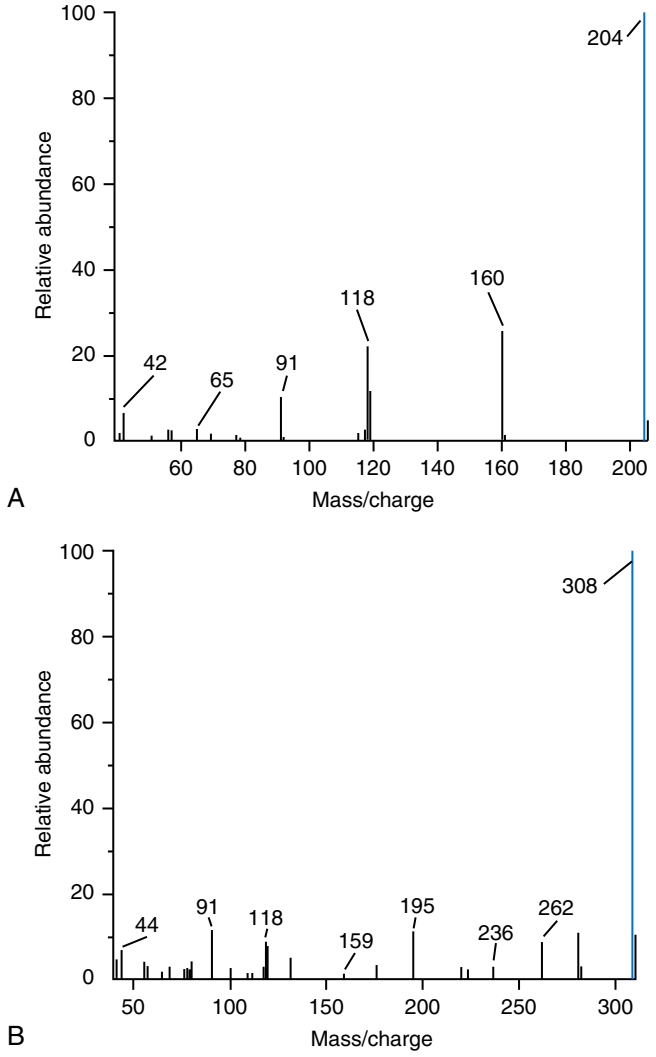


Fig. 13.1 Mass spectrum of: (A) pentafluoropropionyl, and (B) carbethoxyhexafluorobutyryl derivatives of D-methamphetamine.

By convention, an unknown is considered to be identified if the relative abundances of three or four ion fragments agree within ±20% of those of a reference mass spectrum.

Mass spectra of most molecules contain a series of peaks, most often differing by approximately one *m/z* unit, with the differing masses of the peaks attributed to the presence of atoms of different elemental **isotopes** within the molecule. These peaks are commonly referred to as isotope peaks or isotopic peaks. For example, the mass spectrum of $C_2H_7O^+$ consists primarily of three isotope peaks at nominal *m/z* values of 47, 48, and 49, with relative abundances of 100%, 2.4%, and 0.2%, respectively. The higher isotope peaks of $C_2H_7O^+$ contain contributions from ^{12}C, ^{13}C, ^{1}H, ^{2}H, ^{16}O, ^{17}O, and ^{18}O in various combinations. For example, the *m/z* 48 peak of $C_2H_7O^+$ is composed of $^{12}C^{13}C^{1}H_7{}^{16}O^+$, $^{12}C_2{}^{1}H_6{}^{2}H^{16}O^+$, and $^{12}C_2{}^{1}H_7{}^{17}O^+$ (Fig. 13.3). Calculation of isotopic patterns can be challenging, particularly for high-molecular-weight compounds. Various methods used to meet this challenge are discussed in the literature.

Molecules can be isotopically labeled by substituting atoms of a selected isotope for the naturally occurring isotopic atoms at specified portions of the molecule. A labeled molecule can be separated from an unlabeled molecule by its mass difference. The relative abundances of labeled versus unlabeled molecules can be measured. **Isotope dilution mass spectrometry** uses this feature by adding a known amount of stable isotope-labeled **internal standard** into a sample and basing quantitative measurements on the ratio of abundance of the analyte to the abundance of the internal standard. This technique compensates for variation during sample preparation and instrumental analysis (e.g., incomplete recovery of the target analyte during sample preparation, or day-to-day variations in instrument response) and enables use of MS for quantitative analysis.

INSTRUMENTATION

All mass spectrometers include the following: (1) ion source, (2) vacuum system, (3) mass analyzer, (4) detector, and (5) computer (Fig. 13.4). In most cases, a chromatograph is used as a sample introduction device.

Ion Source

All MS techniques require an **ionization** step, in which an ion is produced from a neutral atom or molecule, or ions present in solution get transferred to the gas phase. Electron ionization (EI) and chemical ionization (CI) are techniques used when gas phase molecules are introduced directly into the mass analyzer, usually from a gas chromatograph. The ionization methods used most frequently when a high-performance liquid chromatograph is interfaced with a mass spectrometer (HPLC-MS) include (1) **electrospray ionization** (ESI) and (2) atmospheric pressure chemical ionization (APCI); other less frequently used ionization techniques are (1) atmospheric pressure photoionization (APPI), (2) inductively coupled plasma (ICP), and (3) **matrix-assisted laser desorption/ionization (MALDI)**.

Electron Ionization

In EI, gas phase molecules are bombarded by electrons emitted from a heated filament (Fig. 13.5). This process must occur in a vacuum to prevent filament oxidation and collisional attenuation of the electron beam. Electrons emitted from a hot filament by thermionic emission are accelerated through a potential difference of ~70 V and directed orthogonal to a flow of vaporized sample. Electrons have enough kinetic energy to eject an electron from the analyte molecule, producing, in nearly every case, a radical cation. In most cases, these radical ions then undergo unimolecular rearrangements and fragmentation to produce other cations and radicals:

$$A - e^- \rightarrow A^{+\bullet}$$

Positive ions are drawn out of the ionization chamber by an electrical field and introduced into the mass analyzer. The relative intensity of the molecular ion and the fragment ions is reasonably reproducible. In EI the mass spectral pattern in most cases is dominated by fragment ions and can be matched to the entries in a mass spectral library.

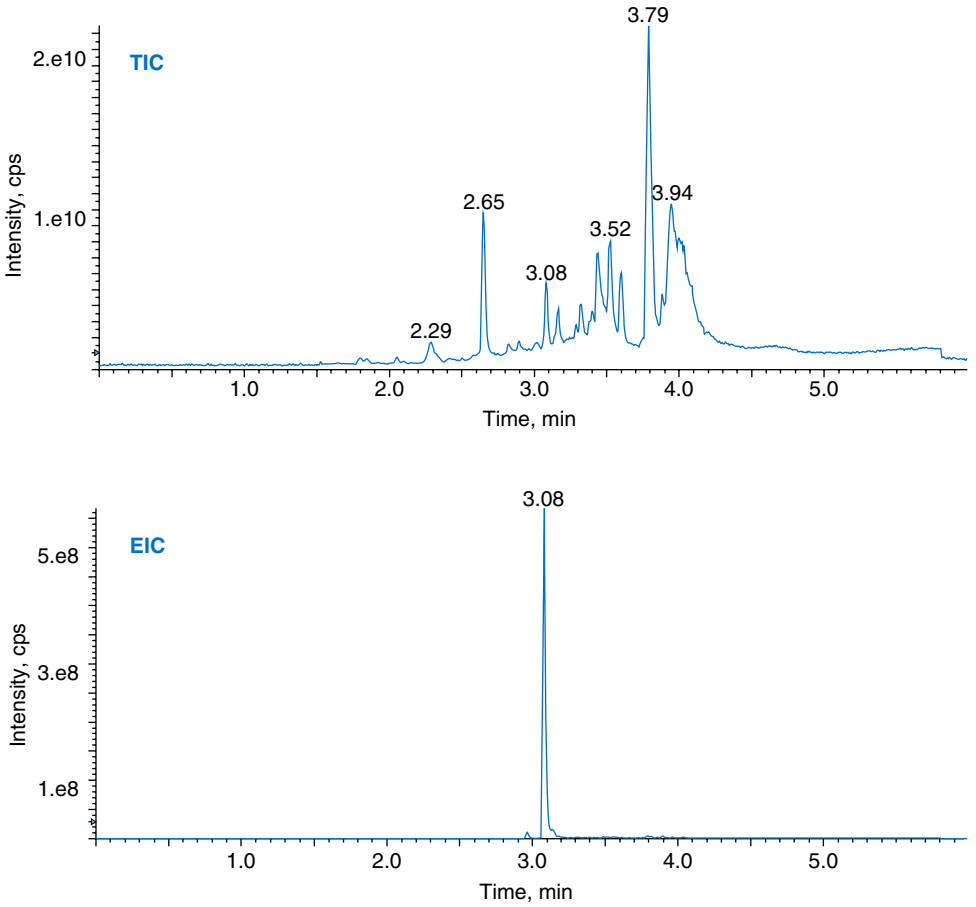


Fig. 13.2 A total ion chromatogram *(TIC)*, scanning *m/z* between 100 and 500, is shown in the top panel, and an extracted ion chromatogram *(EIC)* of human serum with *m/z* between 415.0 and 416.0 is shown in the bottom panel. *CPS*, Counts per second.

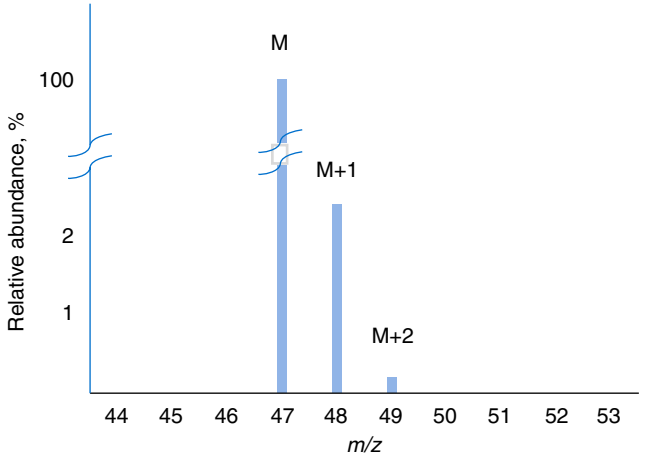


Fig. 13.3 Monoisotopic and isotopic peaks of the ion $C_2H_7O^+$ and their composition. *M* is the monoisotopic ion $C_2H_7O^+$ composed of $^{12}C_2{}^1H_7{}^{16}O^+$. *M+1* is the first isotopic ion of $C_2H_7O^+$ that could be composed of $^{13}C_1{}^{12}C_2{}^1H_6{}^2H^{16}O^+$, $^{12}C_2{}^1H_6{}^2H^{16}O^+$, or $^{12}C_2{}^1H_7{}^{17}O^+$. *M+2* is the second isotopic ion of $C_2H_7O^+$ that could be composed of $^{13}C_2{}^1H_7{}^{16}O^+$, $^{12}C_1{}^{13}C_1H_7{}^{17}O^+$, $^{12}C_2{}^1H_6{}^2H_1{}^{17}O^+$, $^{12}C_2{}^1H_6{}^2H_2{}^{16}O^+$, or $^{13}C_1{}^{12}C_1{}^1H_7{}^{17}O^+$.

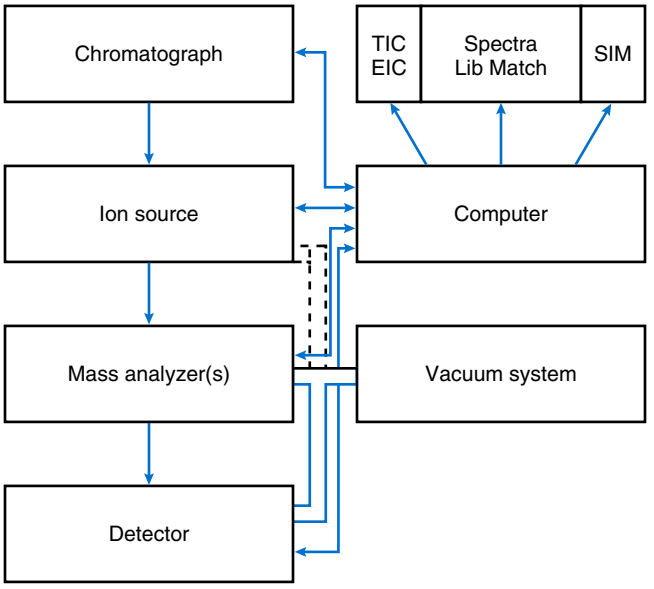


Fig. 13.4 Block diagram of the components of a chromatograph-mass spectrometer system. *EIC*, Extracted ion chromatogram; *SIM*, selected ion monitoring chromatogram; *TIC*, total ion chromatogram.

Chemical Ionization

CI is a "soft" ionization technique, meaning that relatively little fragmentation occurs during the ionization process. A proton is transferred to (or abstracted) from a gas phase analyte by a reagent gas molecule; commonly used reagent gases are (1) methane, (2) ammonia, (3) isobutane, and (4) water. An electron beam ionizes the reagent gas and produces reactive species (e.g., CH_5^+ in the case of methane reagent gas), followed

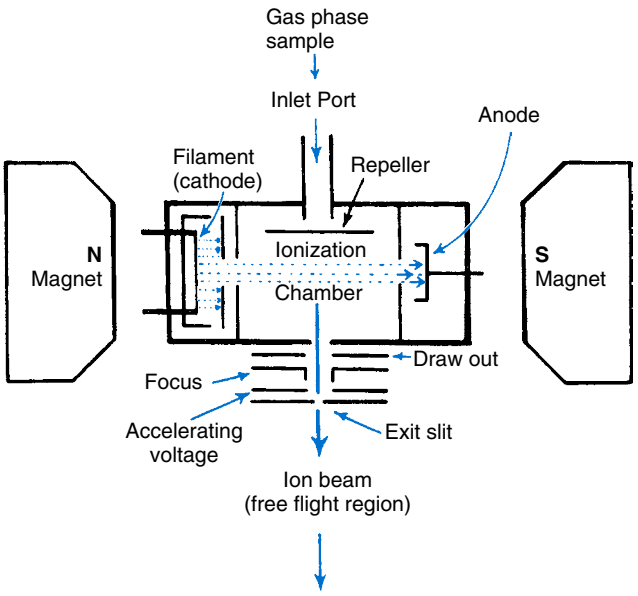


Fig. 13.5 Electron impact ion source. The magnets are used to collimate a dense electron beam, which is drawn from a heated filament placed at a negative potential. The electron beam is positioned in front of a repeller, which is at a slight positive potential compared with the ion source. The repeller sends any positively charged fragment ions toward the opening at the front of the ion source. The accelerating plates strongly attract the positively charged fragment ions.

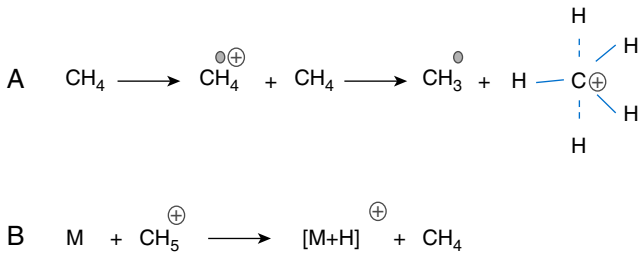


Fig. 13.6 Proposed mechanism of a chemical ionization process. In (A), an electron beam ionizes a reagent gas (CH_4) and produces reactive species (CH_5^+). In (B), positive ions are produced by proton transfer to an analyte molecule.

by proton transfer from CH_5^+ to the analyte molecule. CI is a low-energy process that typically results in a mass spectrum dominated by the (unfragmented) molecular ion (Fig. 13.6). Ions produced by accepting a proton are typically closed-shell ions (i.e., not radical ions). Because of the lack of fragmentation, CI is advantageous for the determination of molecular mass of the constituents of samples, and for quantification.

Another variation of CI, negative ion electron capture CI, has become popular for quantification of molecules containing electronegative functional groups (e.g., chlorine, fluorine, nitro), with formation of negative ions occurring when electrons get captured by the electronegative functional group. Since the number of compounds capable of negative ionization is relatively small, the background signal in the mass spectra is typically low, and because the entire signal corresponds to a molecular ion, negative ion CI often has lower limits of detection than EI and positive ion CI.

Electrospray Ionization

ESI is a technique in which a sample is ionized at atmospheric pressure before it is introduced into the mass analyzer. Effluent from a chromatographic column is passed through a capillary to which 1.5 to 5 kV is applied (Fig. 13.7A). Electrostatic forces applied to the effluent result in formation of charged droplets. Often a coaxially applied nebulizing gas helps nebulize the liquid and direct the charged droplets toward a counter electrode. While droplets evaporate and shrink, the field strength on the surface increases and solvated ions get expelled from the droplets. In positive ion mode, the proton adducts of the molecule (typically clustered with solvent molecules) are "desolvated" to form "bare" ions, which then pass through an aperture inside the vacuum region of the mass analyzer.

ESI depends mainly on solution phase chemistry. Basic compounds (because they are easily protonated) tend to be efficiently detected in positive ion mode. Less commonly a positive ion is formed by adduction of some other positively charged ion, such as an ammonium ion or an alkali metal ion. In some cases, the ion may arise from a permanently charged functional group within the targeted molecule, such as a quaternary ammonium group (e.g., acylcarnitines, choline). Occasionally ionization occurs via electrochemical reactions in the ESI ion source. Acidic compounds may lose a proton to become negatively charged ions in solution, and in many applications can be efficiently detected in negative ion mode.

ESI may produce multiple charged ions from some compounds, particularly from peptides and proteins. It is common to observe approximately one charge for every ~10 amino acid residues in a protein. For example, because a molecule of mass 20,000 yields up to ~20 charges, it is detected at m/z ~1000 (20,000/20), which greatly extends the accessible molecular weight range of an instrument. In most cases where multiple charging occurs, a series of charge states are observed appearing at a series of m/z values (e.g., for a molecule with mass 20,000, m/z 1053.6, 1001.0, 953.4 will be observed for ions of charges +19, +20, +21, respectively). In most cases, ions are produced by proton adduction, and peaks are observed at $m/z = (M + n)/n$, where M is the molecular weight and n is the charge (equal to the number of attached protons).

Although Fig. 13.7A shows the sample introduction probe as it is being directed toward the sampling orifice of the MS, in some designs the ESI probe is offset relative to the sampling cone to enhance performance and minimize contamination of the mass analyzer.

Atmospheric Pressure Chemical Ionization

APCI (Fig. 13.7B) is similar to ESI, as it (1) takes place at atmospheric pressure, (2) involves nebulization and desolvation of ions, and (3) uses the same sampling and extraction cones as ESI. However, no voltage is applied to the inlet capillary; instead, a corona discharge needle is positioned on the path of the evaporated HPLC stream, to initiate a cascade of gas phase reactions that ionize compounds through a series of ion molecule reactions, similar to the reactions occurring in CI (Fig. 13.6), but with molecules of the mobile phase constituents (e.g., water, methanol) serving as reagent molecules. Products of these secondary reactions typically contain clusters of solvent and analyte molecules, so a heated transfer

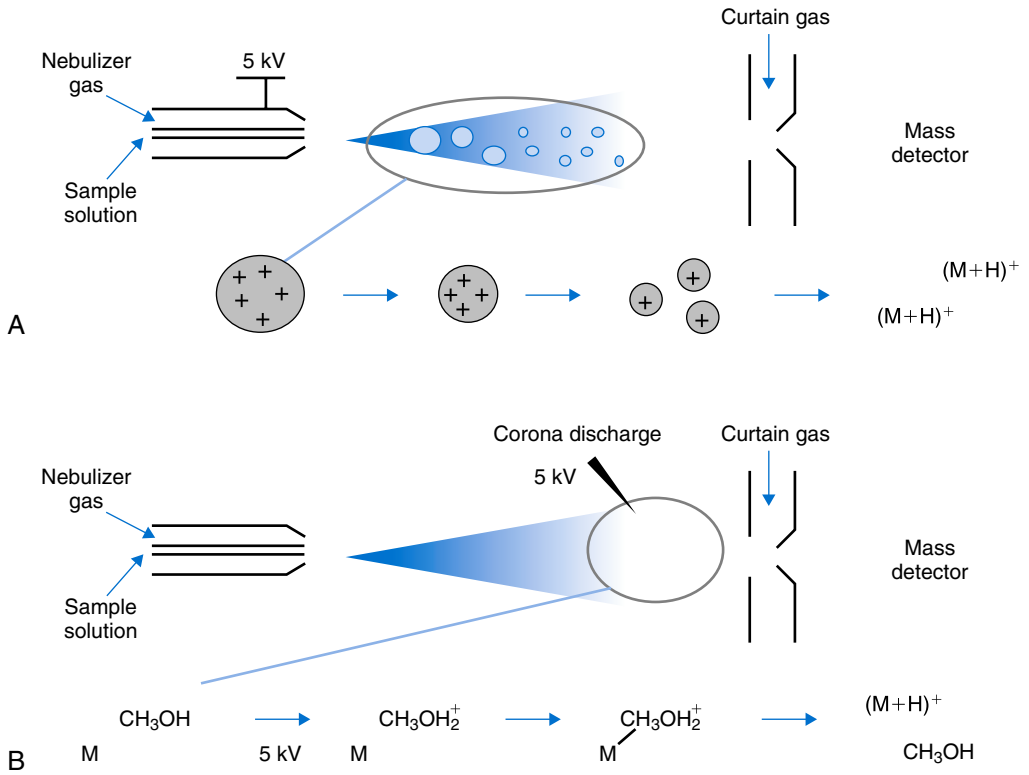


Fig. 13.7 Schematics of (A) electrospray, and (B) atmospheric pressure chemical ionization ion sources. Note the different points where ionization occurs as described in the text.

tube or a countercurrent flow of an inert gas (e.g., nitrogen) and potential differences are used to decluster the ions. APCI is the second most commonly used ion source in methods for quantitative analysis using **liquid chromatography–mass spectrometry (LC-MS)** and LC-MS/MS instruments.

Inductively Coupled Plasma

Similar to ESI and APCI, ICP is an atmospheric pressure ionization method. However, unlike most atmospheric pressure ionization methods, which are "soft" and produce little fragmentation, ICP is the ultimate in "hard" ionization, typically leading to complete atomization of the sample. Consequently its primary use is for elemental analysis. In clinical laboratories it is useful for trace metal and heavy metal analysis in tissue or body fluids. ICP based instruments are extremely sensitive (e.g., parts per trillion) and are capable of a wide dynamic range. The sample is typically prepared by acid digestion, and is introduced into the ion source via a nebulizer fed by a peristaltic pump. The nebulized sample is transmitted into hot plasma maintained at atmospheric pressure by inductively coupling power into the plasma using a high-power radio frequency (RF) generator. A small orifice samples the plasma, and ions are transmitted to the vacuum region of a mass analyzer through a series of differential pumping stages.

ICP-MS is comparatively free from interferences. However, some interferences, such as polyatomic species formed in the torch via ion-molecule reactions, could be of the same molecular mass as the metals of interest; for example, $^{40}Ar^{35}Cl^+$, $^{59}Co^{16}O^+$, $^{36}Ar^{38}Ar^1H^+$, $^{38}Ar^{37}Cl^+$, and $^{36}Ar^{39}K$ are all isobaric to ^{75}As. One approach to resolve these interferences is to use a dynamic reaction cell where a reactant gas (e.g., ammonia) is introduced after the ion source, but before the mass analyzer, causing breakdown of the polyatomic interferences.

Matrix-Assisted Laser Desorption/Ionization

MALDI is a soft ionization technique using a laser energy absorbing matrix to create ions. In order to prepare a sample for MALDI ionization, it should be mixed with a solution of matrix, typically a low-molecular-weight UV-absorbing compound, placed on a target, and dried. As the liquid evaporates, the matrix crystallizes and incorporates analyte molecules in the crystals. A pulsed laser irradiates the sample, triggering rapid vaporization and desorption of the sample and matrix material; the analyte molecules are ionized in the plume of ablated gases, and are introduced into a mass analyzer (Fig. 13.8). Because MALDI produces discrete, pulsed ion packets, it is often coupled with a time-of-flight (TOF) mass analyzer. During the last 5 years, MALDI–TOF instruments have become widely adopted in clinical laboratories for bacterial identification.

POINTS TO REMEMBER

Ionization is required for all mass spectrometry techniques. The term *ionization* describes the production of ions from neutral atoms or molecules using various techniques, including but not limited to EI, CI, and APCI, ESI, or MALDI.

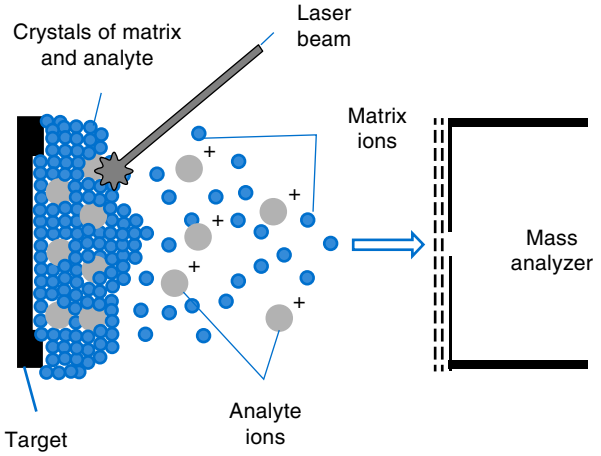


Fig. 13.8 Matrix-assisted laser desorption ionization. Co-crystallized matrix and analyte are irradiated with an ultraviolet laser resulting in formation of a plume of matrix ions and analyte ions. Gas phase ions are directed into a mass analyzer.

Vacuum System

With the exception of certain **ion trap** mass spectrometers, ion separation in a mass analyzer requires that the ions do not collide with other ions and molecules during *m/z* analysis, requiring the use of a vacuum ranging from 10^{-3} to 10^{-9} **torr**, depending on the mass analyzer type. To reach this level of vacuum, mass spectrometers use a combination of mechanical and high vacuum pumps. The mechanical pump evacuates the system to a vacuum at which the high-vacuum pump is effective. The mechanical pump also acts as a backing pump to the high-vacuum pump. Turbomolecular pumps are the most commonly used type of high vacuum pumps in mass spectrometers.

Mass Analyzers, Ion Detectors, and Tandem Mass Spectrometers

Mass spectrometers measure mass-to-charge ratio and not mass. This has a fundamental impact on the physical operating principles of mass spectrometers and influences all aspects of (1) instrument design, (2) operation, and (3) interpretation of results.

Types of Mass Spectrometers

Mass spectrometers are broadly classified as (1) beam and (2) trapping type instruments. In beam-type instruments, ions make one trip through the instrument and are destructively detected when striking a detector. The entire process, from the time an ion enters the analyzer until the time it is detected, generally takes microseconds to milliseconds.

In a trapping-type analyzer, ions are held in a spatially confined region by a combination of magnetic, electrostatic, and/or RF electrical fields. The trapping fields are manipulated to allow *m/z* measurements to be performed. Trapping times vary from a fraction of a second to minutes; most applications use trapping time at the low end of this range.

Beam-type designs. Beam-type mass spectrometers include (1) quadrupole, (2) magnetic sector, and (3) TOF instruments. It is convenient to categorize beam-type instruments into two broad categories: those that produce a mass spectrum by scanning the *m/z* range over a period of time (quadrupole and magnetic sector) and those that acquire successive instantaneous snapshots of the mass spectrum (TOF). This categorization is not definitive, as certain instrument designs have been adapted to either scanning or nonscanning operation. Nevertheless, the categorization is useful because it covers most of the currently available instruments, with both types of instruments used for various types of applications.

Quadrupole. Quadrupole mass spectrometers, sometimes known as *quadrupole mass filters* (QMFs), are the most widely used type of mass analyzers. Although these instruments lag behind magnetic sector instruments in terms of (1) sensitivity, (2) mass range capabilities, (3) resolution, and (4) mass accuracy, they offer a number of attractive and practical features, including (1) ease of use, (2) flexibility, (3) adequate performance for most applications, (4) relatively low cost, (5) smaller size, and (6) less critical site requirements.

A quadrupole mass spectrometer consists of four parallel electrically conductive rods arranged in a square array (Fig. 13.9). The beam enters near the axis at one end of the array, passes through the array in a direction parallel to the axis, and exits at the far end of the array. The ion beam entering the quadrupole array may contain a mixture of ions of various *m/z* values; by adjusting voltages applied to the quadrupoles and the lenses, it is possible to allow ions of a narrow *m/z* range (typically $\Delta m/z < 1$) to be transported through the device to reach the detector. Ions with *m/z* outside this narrow range are ejected radially. The $\Delta m/z$ range represents a passband, analogous to the bandwidth of an interference filter in optics. This is why quadrupole mass spectrometers in the past were referred to as "mass filters."

The principle of operation of quadrupole mass spectrometers is based on a superposition of RF and direct current (DC) potentials applied to the rods. DC voltages are applied to the electrodes in a quadrupolar pattern. For example, a positive DC potential is applied to electrodes 1 and 3, as indicated in Fig. 13.9, and an equivalent negative DC potential is applied to electrodes 2 and 4. The DC potentials are relatively small, on the order of a few volts. Superimposed on the DC potentials are the RF potentials, also applied in a quadrupolar fashion. The RF potentials range up to the kilovolt range and a frequency of ~1 MHz, which is typically fixed to a constant value.

The device may be operated in a selected ion monitoring mode (SIM) or in a scanning mode. In an SIM mode, both DC and RF voltages are set to allow the center of the passband to be transmitted through the ion path. For example, the mass spectrometer may be set to pass ions of *m/z* 363 ± 0.5; both the center *m/z* and the $\Delta m/z$ are adjusted by the appropriate choice of DC and RF and establishes the passband ($\Delta m/z$) and the resolution [$(m/z)/(\Delta m/z)$].

In general, quadrupole instruments are limited to a resolution of up to several thousand. Because of the relatively small

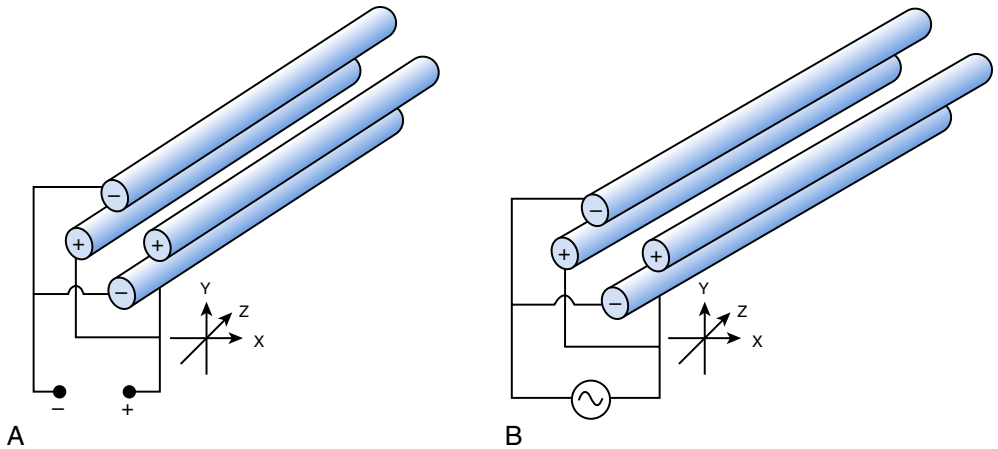


Fig. 13.9 Diagram of quadrupole mass filter, showing (A) direct current and (B) radio frequency voltages applied to a quadrupole rod assembly.

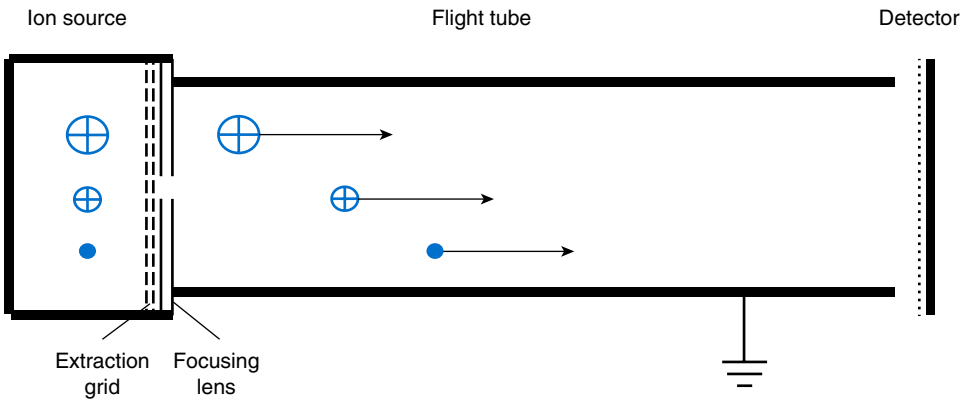


Fig. 13.10 Diagram of time of flight mass analyzer.

size and simplicity of operation, quadrupole mass analyzers are commonly interfaced with gas and liquid chromatographs.

Time of flight. TOF mass spectrometers are widely used and offer a number of advantages, including (1) nearly unlimited *m/z* range, (2) high acquisition speed, (3) high mass accuracy, (4) moderately high resolution, (5) high sensitivity, (6) absence of peak skew in the mass spectrum, and (7) reasonable cost. They are well adapted to pulsed ionization sources, which is an advantage in some applications, particularly with MALDI and related techniques.

Modern TOF mass spectrometers (TOF-MS) are capable of single-digit parts per million (ppm) mass accuracy, which in some cases may allow to confirm the compound's molecular formula. TOF-MSs are conceptually simple to understand, as they are based on a principle that a lighter ion travels faster than a heavier ion, provided that both have the same kinetic energy. A TOF-MS resembles a long pipe (Fig. 13.10); ions are created or injected at the source end of the pipe and accelerated by a potential of several kilovolts. The ions travel down the flight tube and strike the detector at the far end of the flight tube. The time it takes to traverse the tube is known as the *flight time*, which is related to the *m/z* of the ion, and is proportional to the square root of the *m/z* ratio.

To accurately capture such transitory signals, the data recording system must operate on a nanosecond time scale. Advances in signal processing electronics have made TOF-MS practical at relatively modest costs, and this has been a major factor in the rise in popularity of TOF mass analyzers.

TOF-MS is often used with continuous flow ion sources (ESI), but because it is inherently a pulsed technique, it couples readily to pulsed ionization sources, with MALDI being the most common type. MALDI-TOF makes its biggest impact in the area of protein, peptide, and bacterial identification applications.

Another area where TOF-MS excels is high-mass analysis (e.g., intact protein) because its mass range is nearly unlimited. In MALDI-TOF, for example, it is not unusual to detect proteins with molecular weights exceeding 100,000 Da. TOF is also used with ESI and EI ion sources.

Absence of peak skew is another advantage of TOF-MS when it is used as a chromatographic detector. Peak skew is a distortion of the relative abundances of peaks in different parts of a mass spectrum, and this in turn distorts the shape of chromatographic peaks derived from the mass spectral data. Peak skew is related to interaction between the scanning function of the instrument and the changing concentrations of analytes over the elution profile of a chromatographic peak. TOF-MS instruments are essentially free of this artifact.

Trapping mass spectrometers. These mass spectrometers are based on trapping and holding ions for an extended time in a confined space, with the trapping times varying from a fraction of a second to minutes. The division between scanning and nonscanning instruments has less meaning for ion-trapping instruments than for beam-type instruments. The main practical difference between scanning and nonscanning instruments is related to the peak skew, as discussed in the above section on TOF mass analyzers. In terms of producing skewed mass spectra, trapping devices are more similar to nonscanning instruments, such as TOF, than to scanning instruments. This can be explained by capturing ions in an instant and the mass analysis at leisure. Because the sample is captured in an instant, no skewing of the spectra occurs, regardless of whether the *m/z* analysis is performed by a scanning procedure or by a nonscanning procedure.

Classes of ion traps include (1) quadrupole ion traps (QITs), which act by applying an RF field to provide ion trapping; (2) linear ion traps, which are closely related to QITs in their operating principles; (3) ion cyclotron resonance (ICR) mass spectrometers, which rely on a combination of magnetic fields and electrostatic fields for trapping; and (4) Orbitrap mass spectrometers, which use a static electrical field and orbital dynamics for trapping.

Quadrupole ion trap. QITs are used primarily as gas chromatography (GC) or HPLC detectors. They are (1) relatively compact, (2) relatively inexpensive, and (3) versatile. QITs are useful for exploratory studies, structural characterization, and identification of unknown components of samples.

Operation of the QIT is based on the same physical principle as the quadrupole mass spectrometer because both types of the devices use RF fields to confine ions. However, the RF field of an ion trap is designed to trap ions in three dimensions rather than to allow the ions to pass through as in a QMF, which confines ions in two dimensions. A diagram of an ion trap mass spectrometer is shown in Fig. 13.11. In QIT ions are trapped and selectively ejected to generate a mass spectrum.

Although QITs and QMFs were first described at approximately the same time, the QMF initially achieved greater popularity as an analytical device, especially for quantitative analysis. Later two major discoveries allowed wider implementation of QIT: First, addition into the trap of low-molecular-weight gas (e.g., helium at 10^{-3} torr) improved mass resolution and lowered detection limits; second, the development of mass-selective ejection, which improved QIT scanning performance. When no DC voltage is applied and low RF voltage is used, ions of all *m/z* are stored in the QIT field. When the RF voltage is increased, ions of increasing *m/z* become axially unstable and leave the QIT sequentially according to the *m/z*. Ions leaving the QIT through the end cap are detected by an electron multiplier. In addition, a lower limit of detection is obtained with QIT when an axial modulation waveform is applied to the end cap electrodes. This oscillating voltage improves the efficiency of ion ejection from the trap and improves mass resolution.

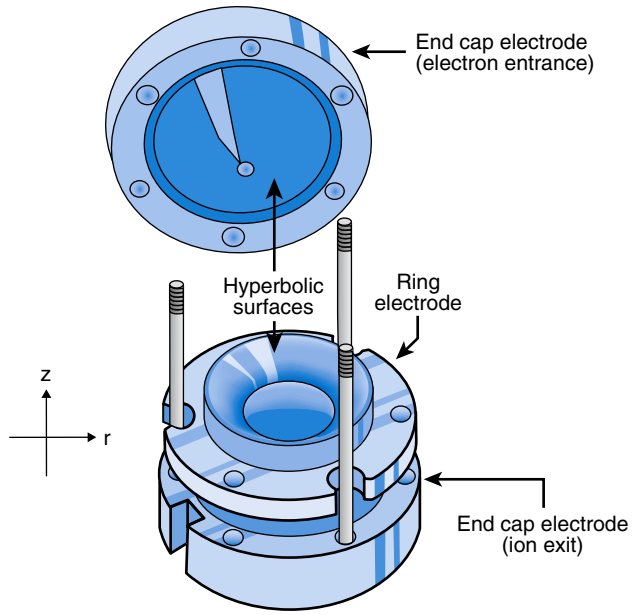


Fig. 13.11 Diagram of a quadrupole ion trap mass analyzer.

In addition to the oscillating voltage mode of operation, the QIT is capable of operation in other modes. For example, the QIT may be operated in a mass-selective storage mode that involves selecting a combination of RF and DC that only allows ions of one *m/z* to be stored in the QIT.

The ability to apply customized waveforms to the QIT makes ion trap one of the most versatile of mass spectrometers, rivaled only by the ICR mass spectrometer. This is most strongly evident in tandem mass spectrometry (MS/MS) and related techniques, which will be discussed later in this chapter.

The QIT also shares some advantages with TOF-MS. In particular, ion trap mass spectrometry is very sensitive. As stated above, because sampling of the ions is decoupled from scanning, no mass spectral peak skewing is seen in GC-MS and HPLC-MS instruments using QIT. One unique capability of QIT, not available in most other types of mass analyzers, is that multiple-stage MS/MS experiments (MS/MS/MS, or MS^n) can be performed in this type of instruments.

Linear ion trap. The linear ion trap is an RF ion trap that is based on a modified linear QMF. Rather than serving as a pass-through device like in a normal linear QMF, in a linear ion trap electrostatic fields are applied to the ends to prevent ions from exiting the device. While trapped, ions are manipulated in the same ways as in a QIT. An advantage of the linear quadrupole trap is that it allows more ions to be stored and consequently has a higher dynamic range than a QIT. Commercial triple quadrupole mass spectrometers are available, in which the third quadrupole is modified so it can function either as a QMF or as a linear trap.

Tandem Mass Spectrometers

Tandem mass spectrometry (MS/MS) has become an important technique in clinical and analytical laboratories. In addition to quantitative analysis, MS/MS instruments are useful for compound identification and structural characterization. When

coupled with chromatographic techniques, well-designed MS/MS assay are characterized by (1) high specificity, (2) low consumable cost, and (3) high sample throughput.

The physical principle of **tandem mass spectrometers** is best understood for beam-type instruments. Two mass spectrometers are arranged sequentially, with a "collision cell" placed between the two mass analyzers. The first analyzer is used to select ions of a particular *m/z*, called *precursor ions*. The precursor ion is directed into the collision cell, where ions collide with background gas molecules and are fragmented into smaller ions, called **product ions**. The second analyzer acquires the mass spectrum of the product ions.

The key to the high selectivity of MS/MS is that it characterizes compounds by two physical properties: precursor ion mass and product ion mass. Chromatographic separation adds a third dimension of selectivity. Use of these three unique properties allows high-specificity analysis and eliminates most of the interferences, which could be encountered using techniques with less selective detection.

As illustrated in Fig. 13.12, a variety of scan functions are possible with tandem mass spectrometers. A product ion scan involves setting the first mass spectrometer, MS1, to select a given *m/z* and scanning through the full mass spectrum of product ions. This mode of instrument operation is often used for structural characterization of unknown compounds. In **precursor ion scan** mode, the second mass spectrometer (MS2) is set to select a specific product ion, and MS1 is scanned through the spectrum. Peaks in the precursor ion scan are indicative of parent ions producing a specific product ion—a capability that is often used to analyze for specific classes of compounds. A **constant neutral loss** scan consists of scanning both mass spectrometers synchronously, with a constant offset between the two mass analyzers. This scan function is selective for ions that lose a neutral fragment of a specific mass corresponding to the constant offset between the two mass spectrometers. A multiple reaction monitoring (MRM) scan is not actually a scan function but is rather a method of cycling through a list of precursor/product ion pairs.

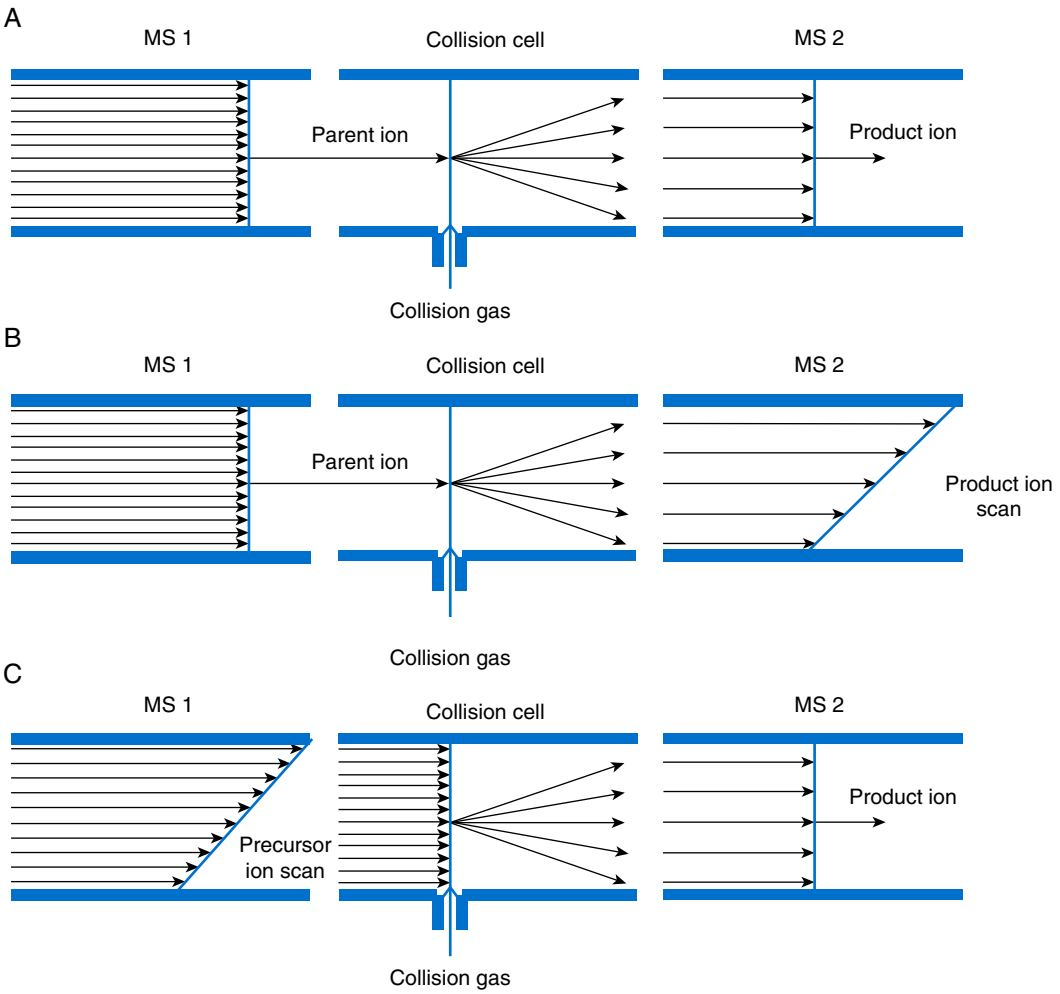


Fig. 13.12 Scan modes in triple quadrupole mass spectrometry (MS/MS). (A) Shows single reaction monitoring *(SRM)*. The first mass analyzer *(MS1)* is fixed to monitor a single *m/z*, and the second mass analyzer *(MS2)* is fixed to monitor a single *m/z*. In cases when multiple mass transitions are monitored, this scan type is named "multiple reaction monitoring". (B) Shows product ion scan mode. MS1 is fixed to monitor a single *m/z*, and MS2 scans through a range of *m/z*. (C) Shows precursor ion scan mode. MS2 is fixed to monitor a single *m/z*, and MS1 is scanned through a range of *m/z* values.

As with single-stage mass spectrometers, tandem mass spectrometers are categorized as beam-type and trapping instruments. The most popular beam-type instrument is the triple quadrupole. In this instrument, the first quadrupole (Q1) functions as MS1, and the third quadrupole (Q3) functions as MS2. Between these two quadrupoles is another quadrupole, Q2, which functions as the collision cell rather than as a QMF.

So-called hybrid mass spectrometers include a combination of two different types of mass spectrometers in a tandem arrangement. One popular type of instrument is the combination of a quadrupole for MS1 and a TOF for MS2. These instruments are mainly used in proteomics and metabolomics research.

POINTS TO REMEMBER

Multiple reaction monitoring (MRM) is a type of mass spectral acquisition in triple quadrupole mass spectrometers that involves cycling through a list of precursor/product ion pairs. With MRM acquisition, both mass analyzers MS1 and MS2 are set in a static mode, whereby only precursor ions specific for the targeted compounds and the internal standards are passed through MS1. These preselected precursor ions are then fragmented in the collision cell, and the fragment ions derived from the compounds of interest are passed by MS2 to the detector.

In triple quadrupole mass spectrometers the first quadrupole (Q1) functions as the first mass analyzer, and the third quadrupole (Q3) functions as the second mass analyzer. Between these two mass analyzers is quadrupole Q2, which functions as the collision cell.

Detectors

With the exception of Fourier-transform ICR MS and Orbitrap type instruments, nearly all mass spectrometers use electron multipliers for ion detection. Three main types of electron multipliers are (1) discrete dynode multipliers; (2) continuous dynode electron multipliers, also known as *channel electron multipliers;* and (3) microchannel plate electron multipliers. Although they are different in design, the principle of operation of all three is based on a multiplication process that is repeated through a chain of dynodes. Most designs of electron multipliers have 12 to 24 discrete dynodes. Both channel electron multipliers and continuous dynode electron multipliers use a continuous semiconducting layer. The multiplication process typically produces a gain of 10^4 to 10^8, where the generation of one electron at the first dynode produces a pulse of 10^4 to 10^8 electrons at the end of the cascade. The duration of the pulse is typically less than 10 nanoseconds.

Another type of detector used in mass spectrometers is the Faraday cup. This detector collects the ion current directly without going through a multiplication step. The Faraday cup (commonly used with ICP-MS instruments) is used when the ion abundance is so high that it would saturate the output of an electron multiplier.

Data Acquisition and Data Analysis Software

The signal produced by an instrument is digitized, and the digital signal is recorded and processed by computers. Modern MS instruments generate immense quantities of raw data, requiring high computing power. For example, in toxicology laboratories, one important function of the data analysis is library searching, which assists in compound identification. Several commercial libraries are available, including (1) Wiley Registry of Mass Spectral Data (http://www.wiley.com/WileyCDA/WileyTitle/productCd-1119284228.html; accessed on October 18, 2017), (2) the US National Institute of Standards and Technology (NIST) Mass Spectral Database (http://www.nist.gov/srd/nist1a.cfm/; accessed on October 18, 2017), and the (3) Pfleger, Maurer, Weber drug libraries. In addition, many laboratories generate custom libraries. Important factors in the utility of such libraries include (1) the quality of available spectra, (2) the search algorithm, and (3) the number of entries in the library. Several library search algorithms are available; the most popular are probability-based matching and the dot product matching approaches. All approaches allow assessment of match quality between the observed spectra and the library spectra.

The computer and software are used for (1) controlling the instrument and data acquisition, (2) signal processing, (3) quantitative calibration, (4) calculation of concentrations in quantitative analysis, (5) database searches, and (6) generating reports.

CLINICAL APPLICATIONS

Mass spectrometers coupled with gas and liquid chromatographs (GC-MS and LC-MS) are versatile analytical instruments that allow development of methods of high specificity, high sensitivity, and high throughput analysis of complex samples.

Gas Chromatography-Mass Spectrometry

Gas chromatography–mass spectrometry (GC-MS) has been used for decades in the analysis of biological samples. For example, it is used by the NIST as the definitive method of qualifying standard reference materials and assigning certified values of many clinically relevant analytes. One of the most common applications of GC-MS is in drug testing for clinical or forensic purposes. Because many drugs have relatively low molecular weight and are sufficiently volatile, they are suitable for analysis by GC. EI with full-scan mass acquisition is a widely used approach for comprehensive drug screening. Identification of unknown compounds is achieved by comparison of the acquired mass spectra with entries in a mass spectral library. In addition to identification of unknown constituents in the samples, mass spectrometry based techniques are the standard methodology for confirming presence of drugs in samples presumptively found to be positive by immunochemical analyses.

Some examples of the clinical use of GC-MS based methods beyond drug testing are in the quantitative analysis of (1) xenobiotic compounds, (2) steroids, (3) pesticides, (4) pollutants, and (5) inborn errors of metabolism.

An important limitation of GC-MS is the requirement that compounds be sufficiently volatile to allow transfer from the solid or liquid phase to the gas phase, and transfer from the chromatographic column to mass spectrometer. Considering that volatility is required in order for a compound to be amenable to GC analysis, many biological compounds need to be derivatized before they can be analyzed using GC-MS.

Derivatization is a technique that transforms a chemical compound into a different compound of similar, though not identical, chemical structure. Derivatization involves the reaction of chemically active functional groups within the targeted molecule and results in a product that may have improved volatility and stability at elevated temperatures, different boiling and melting temperatures, and more favorable fragmentation patterns during mass analysis. The resulting new chemical and physical properties can be beneficial for making compounds amenable to GC-MS analysis, and may enhance sensitivity and specificity.

Liquid Chromatography-Mass Spectrometry

Compared with gas chromatographs, it is more difficult to interface liquid chromatographs with mass spectrometers because the mobile phase is a liquid rather than a gas. The necessity to evaporate liquids often causes problems for the vacuum pumping system of the mass spectrometer. As discussed above, several interface techniques for coupling liquid chromatographs to mass spectrometers now allow HPLC-MS and HPLC-MS/MS to be successfully applied to a wide variety of applications. In theory, as long as a compound is dissolved in a liquid and can be transferred in a gas phase and ionized, it can be introduced into an MS system.

LC-MS/MS methods are widely used in the areas of (1) newborn screening for metabolic disorders, (2) endocrinology and metabolism, (3) toxicological drug screening and confirmation, (4) therapeutic drug monitoring, (5) measurement of cancer biomarkers, and (6) analysis of protein and peptide biomarkers. The majority of current methods for targeted analysis use MRM acquisition, using mass transitions specific to the analyte of interest (Fig. 13.12A).

An important area in which tandem mass spectrometry is used clinically is newborn screening and confirmation of genetic and metabolic disorders. The ability to analyze multiple compounds in a single analysis makes this technique a highly efficient tool for screening purposes. For example, electrospray tandem MS is the most commonly used methodology for analysis of acylcarnitines and amino acids, which are used to screen for disorders of: fatty acid oxidation, organic acid metabolism, the urea cycle, and enzymes in the steroid pathway.

One difficulty in measuring acylcarnitines and amino acids relates to the different sensitivities required to measure the different analytes. To address this issue, a butyl ester derivatization of carboxyl groups can be used to make the molecules more ionizable in positive ion mode, thus yielding more uniform ionization efficiency and sensitivity. Typical amino acid analysis by MS/MS includes esterification to form butyl esters of amino acids, followed by LC-MS/MS analysis using collision-induced dissociation, resulting in the loss by most of the derivatized amino acids of a neutral fragment, butyl formate (m/z 102 Da) to enable sensitive analysis. The method uses a neutral loss scan of m/z 102 Da to selectively detect amino acids. The same derivatization strategy based on formation of butyl derivatives can be used for analysis of acylcarnitines, with detection (using precursor ion scan mode) of the molecules resulting in a loss of the m/z 85 Da. Because of rapid analysis, these methods are routinely used for screening of large number of samples.

A large number of quantitative LC-MS and LC-MS/MS assays have been developed and introduced for routine use to measure clinically relevant compounds. In addition to toxicology-related LC-MS applications, quantitative analysis of many endogenous compounds has been achieved, including (1) sex hormones, (2) adrenal steroids, (3) biogenic amines, (4) vitamin D metabolites [25(OH)-vitamin D and 1,25$(OH)_2$-vitamin D], (5) bile acids, (6) methylmalonic acid, (7) homocysteine, and (8) thyroid hormones.

For quantification of targeted compounds, the most effective approach is MRM analysis (Fig. 13.12). With MRM acquisition both mass analyzers, MS1 and MS2 are set in a static mode, whereby only precursor ions specific for the targeted compounds and the internal standards are passed through MS1. These preselected precursor ions are then fragmented in the collision cell, and the fragment ions derived from the compounds of interest are passed by MS2 to the detector. Because only molecule-specific ions are transmitted through MS1 and typically two molecule specific fragment ions are transmitted through MS2, the MRM approach allows much greater specificity of analyses, as well as lower limits of quantification.

Matrix-Assisted Laser Desorption/Ionization–Time of Flight Mass Spectrometry

MALDI is typically coupled with TOF mass analyzers (Fig. 13.13). During the last few years, identification of microorganisms using MALDI-TOF has become widely accepted for routine use in clinical laboratories. Identification of the microorganisms by MALDI-TOF is performed by analysis of extracts of cultured microorganisms. The fact that different bacteria express unique mixture of proteins and peptides serves as the basis of bacteria identification. When culture samples are analyzed using MALDI-TOF, the bacteria specific mass spectra are observed in the mass range of m/z 2000 to 20,000 Da, allowing the use of mass spectral libraries with fingerprints of the peaks corresponding to the bacteria-specific proteins and peptides. Some of the disadvantages of the technique are related to the lack of information on the identity of most of the observed mass spectral peaks, inability to distinguish some strains of the bacteria, and that the mass spectral libraries are typically specific to the method of sample preparation.

Some limitations of MALDI inherent of the ionization process are (1) MALDI cannot be directly interfaced with

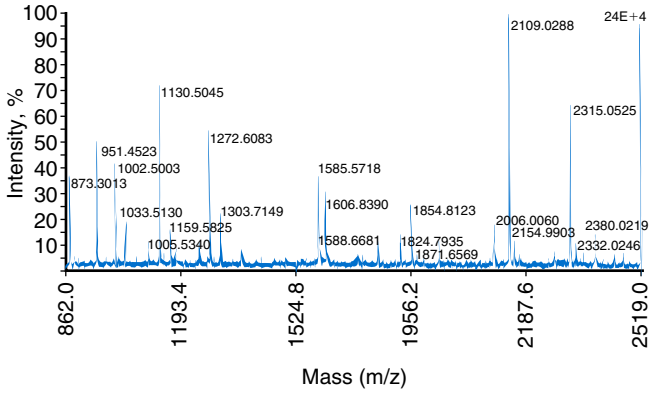


Fig. 13.13 An example of a matrix-assisted laser desorption/ionization–time of flight spectrum showing peptides generated in a tryptic digest of a spot from a two-dimensional sodium dodecyl sulfate-polyacrylamide gel.

online separations (e.g., HPLC, capillary electrophoresis), (2) there is a high background noise in MALDI mass spectra, and (3) poor reproducibility among individual scans. Because of the above, MALDI-TOF mass spectrometry is predominantly a qualitative technique. Progress has been made toward MALDI-TOF use for quantitative analysis; however, such applications often require off-line sample preparation and fractionation prior to the instrumental analysis.

Inductively Coupled Plasma Mass Spectrometry

In clinical laboratories, ICP-MS is used for quantification of trace elements in whole blood, urine, plasma, serum, and tissue biopsy samples. Clinical reasons for analyzing biological samples for trace elements include screening and confirmation of suspected heavy metal poisoning, and diagnosing/monitoring patients with metabolic disorders (e.g., Wilson disease). In some cases, toxicity depends on the presence the metal within organic or inorganic molecules. In these cases it could be important to determine concentration of the element in specific molecules, rather than the total concentration of the element in the sample. These applications use sample fractionation with HPLC or capillary electrophoreses to separate various metal-containing molecules before sample introduction in an ICP-MS instrument.

Proteomics and Metabolomics

In the mid-1990s MS came to the forefront of analytical techniques used to study proteins, and the term *proteomics* was coined. Currently MS is routinely used to accomplish many tasks in proteomics. Traditionally proteomics referred to the identification and quantification of proteins and their posttranslational modifications. A common approach to identify proteins is known as the bottom-up analysis, whereby proteins are enriched and then digested. The resulting digests are then analyzed, and the detected peptides are matched to the theoretical digests using mass spectral databases. Partial sequencing of intact proteins may be achieved without prior digestion of proteins to peptides using top-down techniques, or using partially digested proteins using a middle-down approach. Approaches used for protein top-down characterization include extraction of the proteins from samples, followed by fractionation and analysis of the samples using high resolution MS/MS with collision induced dissociation, higher energy collision dissociation (HCD), electron capture dissociation (ECD), or electron-transfer dissociation (ETD) fragmentation. The main benefits of the top-down and middle-down analyses are in the ability to detect proteins containing posttranslational modifications and sequence variants.

Most often the term *proteomics* is used in the context of biomarker discovery. In clinical diagnostic testing, it also includes qualitative and quantitative analysis of known protein and peptide biomarkers. One of methods recently introduced in routine testing is quantitative analysis of thyroglobulin (Tg), a marker for the recurrence of thyroid cancer. Measurement of Tg is commonly used for the follow-up of patients treated for differentiated thyroid carcinoma. Because thyroid tissue is the only source of Tg, after total thyroidectomy and radioactive ablation, serum concentrations of Tg should decrease to very low or undetectable concentrations. A rise in the serum concentration of Tg is indicative of cancer recurrence. The presence of endogenous anti-Tg autoantibodies (Tg-AAb) can mask the epitopes used by reagent antibodies in immunoassays for Tg, and this may lead to false negative results. The LC-MS/MS based methods for measurement of Tg overcome this interference of Tg-AAb. The published methods use enzymatic digestion of proteins present in samples. During the digestion, Tg and Tg-AAb get digested, allowing accurate measurement of Tg based on the concentration of Tg-specific peptides present in the digested samples. Currently LC-MS/MS methods for thyroglobulin are offered in several commercial laboratories; the method provides an important tool for physicians monitoring patients for the recurrence of thyroid cancer. Introduction of these methods for the measurement of Tg in clinical practice is one example of the application of mass spectrometry for routine quantitative measurement of proteins. LC-MS/MS methods have also been used for hemoglobinopathies, insulin analogs, quantitative measurement of insulin-like growth factor, parathyroid hormone (PTH), parathyroid hormone related protein (PTHrP), and other clinically relevant protein biomarkers. Another area where MS plays an important role is the field of *metabolomics*. This scientific area investigates and characterizes intermediates and products of metabolism that are present in biological fluids under various conditions. Some particular areas of interest in metabolomics include (1) normal homeostasis, (2) disease states, (3) stress, (4) dietary modification, (5) treatment protocols, and (6) aging. In a fashion similar to a mass spectrum providing a fingerprint signature for a specific molecule, patterns of compounds identified in metabolomic studies may provide a fingerprint signature for different physiologic states.

REVIEW QUESTIONS

1. Which of the following types of mass spectrometers involve a single path for ions through the instrument, followed by the ions striking a detector where they are destructively detected?
   a. Trapping type
   b. Beam type
   c. Mass-selective ejection type
   d. Ion-cyclotron resonance type
2. In mass spectrometry, the sum of all ions produced displayed as a function of time represents:
   a. product ion.
   b. extracted ion profile.
   c. mass spectrum.
   d. total ion chromatogram.
3. Which type of ionization technique uses an electron beam to ionize the reagent gas and produce a reactive species as a result of ion-molecule reactions?
   a. Electron
   b. Chemical
   c. Electrospray
   d. Inductively coupled plasma
4. Inductively coupled plasma mass spectrometry is typically used for:
   a. determination of trace elements.
   b. identification of a protein.
   c. drug identification.
   d. isotopic labeling of molecules.
5. The ionization technique most commonly used in mass analyzers coupled with gas chromatography is:
   a. electrospray ionization.
   b. chemical ionization.
   c. matrix-assisted laser desorption/ionization.
   d. electron ionization.
6. The mass spectrometer most commonly used for quantitative analysis is the:
   a. quadrupole mass spectrometer.
   b. time-of-flight instrument.
   c. magnetic sector mass spectrometer.
   d. linear ion trap mass spectrometer.

SUGGESTED READINGS

Anderson, N. L., Anderson, N. G., Pearson, T. W., Borchers, C. H., Paulovich, A. G., Patterson, S. D., et al. (2009). A human proteome detection and quantitation project. *Molecular and Cellular Proteomics*, *8*, 883–886.

Ceglarek, U., Lembcke, J., Fiedler, G. M., Werner, M., Witzigmann, H., Hauss, J. P., & Thiery, J. (2004). Rapid simultaneous quantification of immunosuppressants in transplant patients by turbulent flow chromatography combined with tandem mass spectrometry. *Clinical Chimica Acta*, *346*, 181–190.

Cherkaoui, A., Hibbs, J., Emonet, S., Tangomo, M., Girard, M., Francois, P., et al. (2010). Comparison of two matrix-assisted laser desorption ionization-time of flight mass spectrometry methods with conventional phenotypic identification for routine identification of bacteria to the species level. *Journal of Clinical Microbiology*, *48*, 1169–1175.

Clarke, N. J., & Hoofnagle, A. N. (2011). Mass spectrometry continues its march into the clinical laboratory. *Clinical Laboratory Medicine*, *31*, IX–XI.

Clarke, N. J., Zhang, Y., & Reitz, R. E. (2012). A novel mass spectrometry-based assay for the accurate measurement of thyroglobulin from patient samples containing antithyroglobulin autoantibodies. *Journal of Investigative Medicine*, *60*, 1157–1163.

Fenn, J. B. (2003). Electrospray wings for molecular elephants (nobel lecture). *Angewandte Chemie International Edition England*, *42*, 3871–3894.

Hoofnagle, A. N., & Roth, M. Y. (2013). Clinical review: Improving the measurement of serum thyroglobulin with mass spectrometry. *Journal of Clinical Endocrinology and Metabolism*, *98*, 1343–1352.

Kuhara, T. (2005). Gas chromatographic-mass spectrometric urinary metabolome analysis to study mutations of inborn errors of metabolism. *Mass Spectrometry Reviews*, *24* (6), 814–827.

Kushnir, M. M., Komaromy-Hiller, G., Shushan, B., Urry, F. M., & Roberts, W. L. (2001). Analysis of dicarboxylic acids by tandem mass spectrometry. High-throughput quantitative measurement of methylmalonic acid in serum, plasma, and urine. *Clinical Chemistry*, *47*, 1993–2002.

Kushnir, M. M., Rockwood, A. L., & Bergquist, J. (2010). Liquid chromatography tandem mass spectrometry applications in endocrinology. *Mass Spectrometry Reviews*, *29*, 480–502.

Kushnir, M. M., Rockwood, A. L., Roberts, W. L., Abraham, D., Hoofnagle, A. N., & Meikle, A. W. (2013). Measurement of thyroglobulin by liquid chromatography-tandem mass spectrometry in serum and plasma in the presence of antithyroglobulin autoantibodies. *Clinical Chemistry*, *59*, 982–990.

Mandal, B. K., Ogra, Y., Anzai, K., & Suzuki, K. T. (2004). Speciation of arsenic in biological samples. *Toxicology and Applied Pharmacology*, *198*, 307–318.

Mbughuni, M. M., Jannetto, P. J., & Langman, L. J. (2016). Mass spectrometry applications for toxicology. *EJIFCC*, *27*, 272–287.

Rockwood, A. L., & Palmblad, M. (2013). Isotopic distributions. *Methods in Molecular Biology*, *1007*, 65–99.

Rockwood, A. L., & Van Orden, S. L. (1996). Ultrahigh-speed calculation of isotope distributions. *Analytical Chemistry*, *68*, 2027–2030.

Sandalakis, V., Goniotakis, I., Vranakis, I., Chochlakis, D., & Psaroulaki, A. (2017). Use of MALDI-TOF mass spectrometry in the battle against bacterial infectious diseases: Recent achievements and future perspectives. *Expert Review of Proteomics*, *14*, 253–267.

Villoria, J. G., Pajares, S., Lopez, R. M., Marin, J. L., & Ribes, A. (2016). Neonatal screening for inherited metabolic diseases in 2016. *Seminars in Pediatric Neurology*, *23*, 257–272.

Whitehouse, C. M., Dreyer, R. N., Yamashita, M., & Fenn, J. B. (1985). Electrospray interface for liquid chromatographs and mass spectrometers. *Analytical Chemistry*, *57*, 675–679.

Enzyme and Rate Analyses

Renze Bais, Mauro Panteghini

OBJECTIVES

1. Understand how enzymes are used in diagnosis and monitoring disease conditions.
2. Delineate how enzymes are classified in the Enzyme Commission (EC) system and give an example.
3. Give examples of primary, secondary, tertiary, and quaternary structures of enzymes.
4. Understand the importance of the relationship of structure to enzyme function.
5. Give an example of how enzymes carry out their reactions at the active site.
6. Write an equation for the formation of an enzyme-substrate complex.
7. Describe how enzymatic reactions are influenced by changes in:
   a. Enzyme concentration
   b. Substrate concentration
   c. Ionic concentration
   e. Temperature
   f. Inhibitors and activators
8. Give an example of an enzyme assay and describe how it works.
9. Write an equation that delineates how the velocity of an enzyme reaction depends on substrate concentration.
10. Describe three different kinds of enzyme inhibition.
11. Understand the use of enzymes as analytical reagents.

KEY WORDS AND DEFINITIONS

Activator Small molecule or ion that increases the rate of an enzyme-catalyzed reaction by promoting formation of the most active state of the enzyme, substrate, or other reactant.

Active center/active site That part of an enzyme formed by the tertiary structure at which the noncovalent binding of substrate occurs to form the intermediate enzyme-substrate complex.

Catalyst A substance that modifies and increases the rate of a chemical reaction without being permanently changed or consumed; an enzyme is a protein catalyst of biological origin.

Clinical enzymology The branch of medical science that deals with the biochemical nature and activity of enzymes of clinical relevance.

Coenzyme A small molecule (some contain structures derived from vitamins) that is required for some enzyme-catalyzed reactions to occur. Coenzymes are temporarily or permanently bound to the enzyme.

Enzyme A protein with catalytic properties.

First-order kinetics A phase in an enzymatic reaction where substrate concentration is limiting and the reaction rate is proportional to the concentration of substrate.

Inhibitor A substance that reduces the rate of an enzymatic reaction and is classified as reversible or irreversible.

International unit The quantity of enzyme that catalyzes one micromole of substrate per minute.

Isoenzyme A form of an enzyme that originates at the level of the gene that encodes the structure of the enzyme protein and is distinguished by certain physical properties.

Isoform A form of an enzyme that is modified by posttranslational processing.

Katal The quantity of an enzyme that catalyzes a reaction rate of one mole of substrate per second.

Lineweaver-Burk plot A plot of the reciprocal of the velocity of an enzyme catalyzed reaction (ordinate; y-axis) against the reciprocal of the substrate concentration (abscissa; x-axis).

Michaelis-Menten constant (K_m) A constant for a given enzyme acting under given conditions; the experimentally determined substrate concentration at which the enzymatic reaction velocity equals one-half of the maximum velocity of the enzymatic reaction.

Product The substance produced by an enzyme catalyzed conversion of a substrate.

Substrate A reactant in an enzyme-catalyzed reaction that binds to the active center of an enzyme.

Zero-order kinetics A phase in an enzymatic reaction when the substrate concentration is saturating and the reaction rate is maximal; the reaction rate depends only on the enzyme concentration and is independent of substrate concentration.

Enzymes are biological catalysts that can be used in the monitoring and diagnosis of disease. Enzymes have remarkable properties that make them sensitive indicators of pathologic change. Metabolism can be regarded as an integrated series of enzymatic reactions, and some diseases as a derangement of the physiologic pattern of metabolism. Many enzymes exist in multiple forms and differences in their properties help in differentiating them and understanding organ-specific pathophysiology. Genetically determined variations in enzyme structure between individuals are used to account for such characteristics as differences in sensitivity to drugs and differences in metabolism, which manifest themselves as hereditary metabolic diseases.

ENZYME NOMENCLATURE

Historically, individual enzymes were identified using the name of the **substrate** or group on which they act and then adding the suffix *-ase*. For example, the enzyme hydrolyzing urea was ure*ase*. Later, the type of reaction involved was identified, as in carbonic anhydrase, D-amino acid oxidase, and succinate dehydrogenase. In addition, some enzymes were given empirical names such as trypsin, diastase, ptyalin, pepsin, and emulsin.

Because this combination of trivial common names and semisystematic names was found to be inadequate, in 1955 the International Union of Biochemistry (IUB) appointed an Enzyme Commission (EC) to study the problem of enzyme nomenclature. Its subsequent recommendations, with periodic updating, provide a rational and practical basis for identifying all enzymes now known and enzymes that will be discovered in the future.

With the IUB system, a systematic and trivial name is provided for each enzyme. The systematic name describes the nature of the reaction catalyzed and is associated with a unique numeric code. The trivial or practical name, which may be identical to the systematic name but is often a simplification of it, is suitable for everyday use. The unique numeric designation for each enzyme consists of four numbers, separated by periods (e.g., 2.2.8.11), and is prefixed by the letters *EC,* denoting *Enzyme Commission.* The first number defines the class to which the enzyme belongs. All enzymes are assigned to one of six classes, characterized by the type of reaction they catalyze: (1) oxidoreductases, (2) transferases, (3) hydrolases, (4) lyases, (5) isomerases, and (6) ligases. The next two numbers indicate the subclass and the sub-subclass to which the enzyme is assigned. For example, these may differentiate the amino-transferring subclass from the phosphate-transferring subclass, or the sub-subclass accepting ethanol from those accepting acyl groups. The last number is the specific serial number given to each enzyme within its sub-subclass.

Table 14.1 lists selected enzymes of clinical interest, identified by trivial, abbreviated, and systematic names and by their code numbers.

Although not recommended by the EC, it is a common and convenient practice to use capital letter abbreviations for the names of certain enzymes, such as ALT for alanine aminotransferase, AST for aspartate aminotransferase, LD for lactate dehydrogenase, and CK for creatine kinase (see Table 14.1).

TABLE 14.1 Enzyme Commission Numbers, Systematic and Trivial Names, and Frequently Adopted Abbreviations of Enzymes of Major Clinical Importance

EC Number	Systematic Name	Trivial Name	Abbreviation
1.1.1.27	L-Lactate: NAD^+ oxidoreductase	Lactate dehydrogenase	LD
1.4.1.3	L-Glutamate: $NAD(P)^+$ oxidoreductase (deaminating)	Glutamate dehydrogenase	GLD
2.3.2.2	(γ-Glutamyl)-peptide: amino acid γ-glutamyltransferase	γ-Glutamyltransferase	GGT
2.6.1.1	L-Aspartate: 2-oxoglutarate aminotransferase	Aspartate aminotransferase (transaminase)	AST
2.6.1.2	L-Alanine: 2-oxoglutarate aminotransferase	Alanine aminotransferase (transaminase)	ALT
2.7.3.2	ATP: creatine *N*-phosphotransferase	Creatine kinase	CK
3.1.1.3	Triacylglycerol acylhydrolase	Lipase	LPS
3.1.1.7	Acetylcholine acetylhydrolase	Acetylcholinesterase, true cholinesterase, choline esterase I	—
3.1.1.8	Acetylcholine acylhydrolase	Pseudocholinesterase, butyryl cholinesterase, choline esterase II (serum cholinesterase)	CHE
3.1.3.1	Orthophosphoric-monoester phosphohydrolase (alkaline optimum)	Alkaline phosphatase	ALP
3.1.3.2	Orthophosphoric-monoester phosphohydrolase (acid optimum)	Acid phosphatase	ACP
3.1.3.5	5′-Ribonucleotide phosphohydrolase	5′-Nucleotidase	NTP
3.2.1.1	1,4-α-D-Glucan glucanohydrolase	Amylase	AMY
3.4.21.4		Trypsin	TRY
4.1.2.13	D-Fructose-1,6-bisphosphate D-glyceraldehyde-3-phosphate-lyase	Aldolase	ALD

EC, Enzyme Commission; *NAD*, nicotinamide adenine dinucleotide; *NAD(P)*, NAD phosphate.

ENZYMES AS PROTEINS

Enzymes are proteins and thus possess the primary, secondary, and tertiary structural characteristics of proteins (see Chapter 18). Many enzymes also exhibit quaternary structure. The *primary* structure, the linear sequence of amino acids linked through their α-carboxyl and α-amino groups by peptide bonds, is specific for each type of enzyme molecule. Each polypeptide chain is coiled into three-dimensional (3D) secondary and tertiary structures. Secondary structure refers to the conformation of limited segments of the polypeptide chain, namely α-helices, β-pleated sheets, random coils, and β-turns. The combination of secondary structural elements and amino acid side chain interactions that define the 3D structure of the folded protein is referred to as its tertiary structure. In many cases, biological activity, such as the catalytic activity of enzymes, requires two or more folded polypeptide chains (subunits) to associate to form a functional molecule. The specific combination of these subunits defines the quaternary structure. The subunits may be multiples of the same polypeptide chain (e.g., the MM isoenzyme of CK, the H_4 isoenzyme of LD), or they may be combinations of distinct polypeptides (e.g., the MB isoenzyme of CK).

No feature of primary structure is common to all enzyme molecules. However, considerable homologies of amino acid sequences are found between enzymes that appear to share a common evolutionary origin, such as the proteases trypsin and chymotrypsin. Similarities of sequence are even more marked among the members of a family of isoenzymes. The amino acid sequence in the immediate neighborhood of the **active center** of the enzyme (discussed later) is often similar in enzymes with related function (e.g., the *serine proteases* are so called because they all have this amino acid in the active center).

Enzyme molecules differ in the proportion of secondary structures—such as α-helices—that they contain, although no enzyme molecule studied thus far approaches the large proportion of α-helices found in myoglobin and hemoglobin. The tertiary structures of different types of enzyme molecules are as individually characteristic as their primary structures; nevertheless, some common features exist. Enzyme molecules are roughly globular in overall shape, with a preponderance of polar amino acid side chains on the outside of the molecule and nonpolar side chains in the interior. The ionizable residues in contact with the surrounding medium are responsible for many of the properties of the enzyme molecules in solution, such as their migration in an electric field and their solubility. Covalent disulfide bridges may link different parts of the polypeptide chains in some enzyme molecules, but the 3D structure is mainly stabilized by the large number of hydrophobic interactions formed between the nonpolar side chains in the interior of the molecule.

The application of physical methods, such as x-ray crystallography and multidimensional nuclear magnetic resonance, has provided structural insights from which enzyme mechanisms have been proposed. Furthermore, the tools of molecular biology, such as molecular cloning, have enabled the purification and characterization of enzymes that previously were available only in minute amounts. Molecular biology also enables the manipulation of the amino acid sequence of enzymes. Site-directed mutagenesis (substituting one amino acid residue for another) and deletion mutagenesis (eliminating sections of the primary structure) have enabled the identification of chemical groups that participate in ligand binding and specific chemical steps during catalysis. The biological activity of a protein molecule depends generally on the integrity of its structure, and any disruption of its structure will likely be accompanied by a loss of activity, a process known as *denaturation.* If the process of denaturation is minimal, it may be reversed with the recovery of enzyme activity on removal of the denaturing agent. However, prolonged or severe denaturing conditions result in an irreversible loss of activity. Denaturing conditions include increased temperatures, extremes of pH, and exposure to certain chemicals. Heat inactivation of most enzymes takes place at an appreciable rate at room temperature, and in most cases becomes very fast above 60°C. However, certain enzymes, most notably some polymerases, can retain activity at temperatures as high as 95°C. This property has been exploited in the polymerase chain reaction (see Chapter 48). Extremes of pH also cause unfolding of enzymes and, with few exceptions, should be avoided when enzyme samples are preserved. Addition of chemicals, such as urea and detergents, disrupts hydrogen bonding and hydrophobic interactions, also resulting in enzyme inactivation.

SPECIFICITY AND THE ACTIVE CENTER

With the exception of enzymes such as proteases, nucleases, and amylases, which act on macromolecular substrates, enzyme molecules are considerably larger than their substrate molecules. Consideration of the structure of an enzyme's active site and its substrate(s) in their ground and transition states is necessary to understand the rate enhancement and specificity of chemical reactions catalyzed by an enzyme. The active site of an enzyme will vary among enzymes, but in general, the following statements hold:

1. The active site of an enzyme is relatively small compared with the total size of the enzyme molecule and may involve less than 5% of its amino acids.
2. The active sites of enzymes are 3D structures that are formed as a result of the overall tertiary structure of the protein. This results in the amino acids and cofactors in the active site of an enzyme being spatially structured in an exact, 3D relationship with respect to one another and to the structure of the substrate molecule.
3. Typically, the binding of substrate molecules to the active site of an enzyme is noncovalent, and is stabilized by hydrogen bonding, electrostatic, hydrophobic, and van der Waals interactions.
4. Active sites of enzymes typically occur in clefts and crevices in the protein's tertiary structure.

The specificity of substrate binding is a function of the exact spatial arrangement of atoms in the enzyme active site that complements the structure of the substrate molecule.

ISOENZYMES AND OTHER MULTIPLE FORMS OF ENZYMES

Isoenzymes are enzymes that can catalyze the same reaction, but that differ in some aspect of their structure. They may occur within a single organ or even within a single type of cell. The forms can be distinguished on the basis of differences in various properties, such as electrophoretic mobility, immunological reactivity, and resistance to chemical or thermal inactivation. They often have significant quantifiable differences in catalytic properties, but all forms of a particular enzyme retain the ability to catalyze its characteristic reaction.

The existence of multiple forms of enzymes in human tissue has important implications in the study of human disease. The presence in different organs of isoenzymes with distinctive properties helps in understanding organ-specific patterns of metabolism. Genetically determined variations in enzyme structure among individuals account for such characteristics as differences in sensitivity to drugs and differences in metabolism, which manifest themselves as hereditary metabolic diseases. For diagnostic enzymology, the existence of multiple forms of enzymes, whether the result of genetic or nongenetic causes, provides opportunities to increase the diagnostic specificity and sensitivity of enzyme assays in body fluid samples.

Similar to other proteins, enzymes usually elicit the production of antibodies when they are injected into animals of a species other than those in which they originate. Even small structural differences among closely similar molecules, such as the members of a family of isoenzymes, are often sufficient to render them antigenically distinct, allowing antibodies to be produced specific to a single isoenzyme. The availability of immunochemical methods has been particularly important in the analysis of isoenzyme mixtures with many commercial immunoassays using monoclonal antibodies to increase specificity.

Genetic Origins of Enzyme Variants

True isoenzymes result from the existence of more than one gene coding for the structure of the enzyme protein. Many human enzymes (more than one-third) are known to be determined by more than one structural gene. Similar but distinct genes have undergone differential modifications during the course of evolution, so that the enzyme proteins coded by them no longer have identical structures, although they can catalyze the same reaction; in other words, they are isoenzymes.

The multiple genes that determine a particular group of isoenzymes are not necessarily closely linked on one chromosome. For example, the structural genes that code for human salivary and pancreatic amylases both are located on chromosome 1, whereas the genes that code for cytoplasmic and mitochondrial malate dehydrogenase are on chromosomes 2 and 7, respectively. Among the enzymes of clinical importance that exist as isoenzymes because of the presence of multiple-gene loci are LD, CK, α-amylase, and some forms of alkaline phosphatase (ALP).

An enzyme can exist in molecular forms that differ from one individual to another because of the existence of alternative alleles that are inherited according to Mendelian laws. These alleles give rise to gene **products**, enzymes, with the same function. Isoenzymes that result from the existence of different alleles are termed *allozymes.* Many human genes have allelic variation and the probability that individuals will differ in their isoenzyme patterns is high.

The number of allelic variants and the frequency with which particular variants occur within the population vary considerably across enzymes. For example, mutations at either of the two principal loci that determine human LD are extremely rare, but a high incidence of mutant alleles occurs at the single locus that determines the structure of placental ALP. More than 400 mutations in the glucose-6-phosphate dehydrogenase gene have now been identified on the X chromosome. When isoenzymes, because of variation at a single locus, occur with appreciable frequency in a human population, the population is said to be *polymorphic* with respect to the isoenzymes in question.

Another category of multiple molecular forms can arise when enzymes are oligomeric, and possess quaternary structure. The association of different subunits in various combinations gives rise to a range of active enzyme molecules. When the subunits are derived from different structural genes—multiple genes or multiple alleles—the hybrid molecules so formed are called *hybrid isoenzymes.* The ability to form hybrid isoenzymes is evidence of considerable structural similarities among the different subunits. Hybrid isoenzymes can be formed in vitro, but they are also formed in vivo when different types of constituent subunits are present in the same subcellular compartment.

The number of different hybrid isoenzymes that can form from two nonidentical protomers depends on the number of subunits in the complete enzyme molecule. For a dimeric enzyme, one mixed dimer (hybrid isoenzyme) can be formed. If the enzyme is a tetramer, three heteropolymeric isoenzymes may be formed (Fig. 14.1). Examples of hybrid isoenzymes include the mixed MB dimer of CK and the three hybrid isoenzymes, LD-2, LD-3, and LD-4 of LD.

Nongenetic Causes of Multiple Forms of Enzymes

Posttranslational modifications of enzyme molecules give rise to multiple forms known as **isoforms** (Fig. 14.2). For example, removal of amide groups accounts for some of the heterogeneity of amylase and carbonic anhydrase (each of these enzymes also exists as a true isoenzyme). Modification also can take place as a result of extraction procedures. Many erythrocyte enzymes, including adenosine deaminase, acid phosphatase, and some forms of phosphoglucomutase, contain sulfhydryl groups that are susceptible to oxidation. In hemolysates, oxidation may be brought about by the action of oxidized glutathione, although in intact cells, this compound is present in its reduced form.

Serum isoforms of CK are formed as part of the normal clearance process of the cell. Human myocardial and skeletal muscle tissues have the CK-MM and CK-MB isoenzymes, which

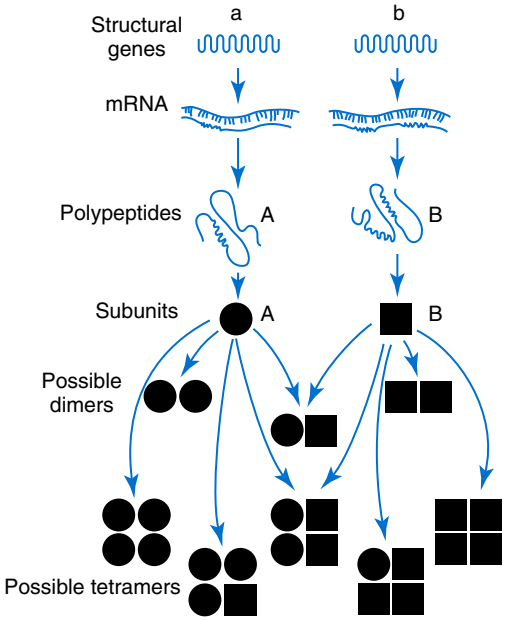


Fig. 14.1 Diagram showing the origin of isoenzymes, assuming the existence of two distinct genes. When the active enzymes are polymers containing more than one subunit, hybrid isoenzymes consisting of mixtures of different subunits may be formed. One such isoenzyme can be formed in the case of a dimeric enzyme, such as creatine kinase and three if the enzyme is a tetramer (e.g., lactate dehydrogenase). In both cases, two homopolymeric isoenzymes can also exist. (From Moss, D. W. [1979]. *Isoenzyme analysis*. London: The Chemical Society. Reproduced by permission of The Royal Society of Chemistry.)

are modified on release into the circulation by removal of the C-terminal amino acid lysine from the M subunits by the action of carboxypeptidase.

Modifications of enzyme molecules also may include nonprotein components that lead to molecular heterogeneity. Many enzymes are glycoproteins, and variations in carbohydrate side chains are a common cause of nonhomogeneity. Some carbohydrate moieties, notably *N*-acetylneuraminic acid (sialic acid), are strongly ionized and consequently have a profound effect on some enzyme molecules. For example, removal of terminal sialic acid groups from human liver and/or bone ALP with neuraminidase greatly reduces the electrophoretic heterogeneity of the enzyme.

Aggregation of enzyme molecules with each other or with nonenzymatic proteins may give rise to multiple forms that can be separated by techniques that depend on differences in molecular size. For example, four catalytically active cholinesterase components with molecular weights ranging from approximately 80,000 to 340,000 are found in most sera, with the heaviest component, C_4, contributing most of the enzyme activity. Other enzyme forms are also occasionally present, but it appears that the principal serum cholinesterase fractions can be attributed to different states of aggregation of a single monomer.

A specific form of interaction between enzymatic and nonenzymatic proteins, such as immunoglobulins, results in the formation of an enzyme-protein complex (macrocomplex). Enzyme–immunoglobulin complexes have been observed involving amylase, LD, CK, ALP, amylase, and other enzymes.

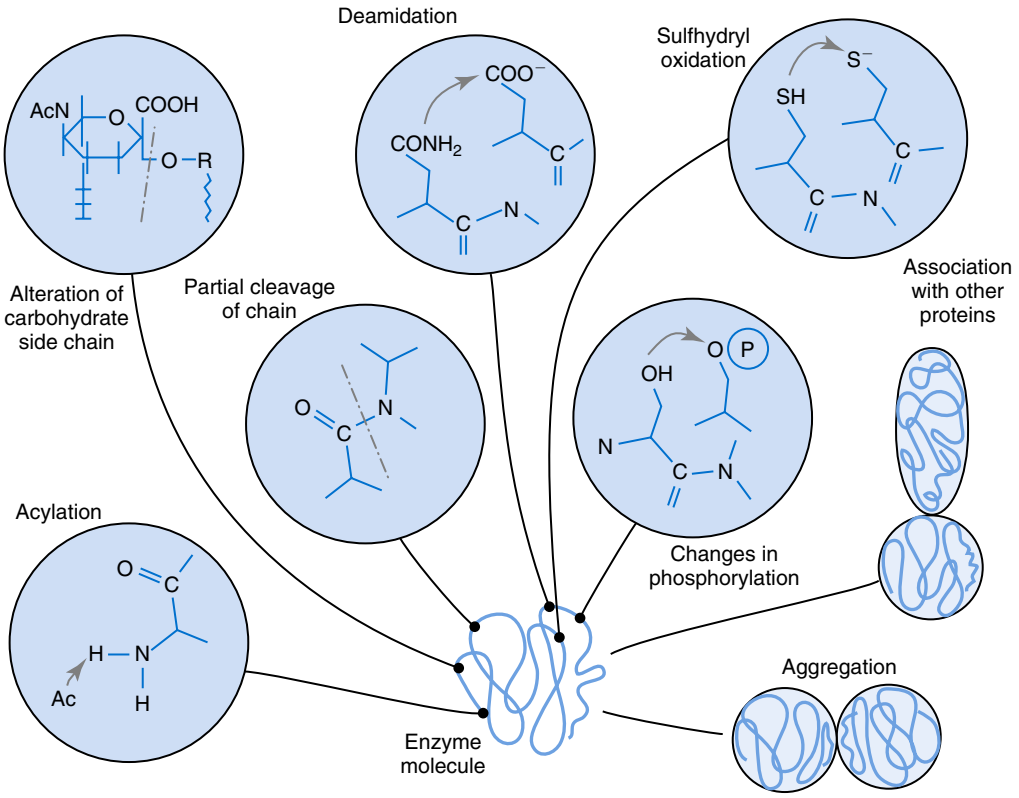


Fig. 14.2 Nongenetic modifications that may give rise to multiple forms of enzymes. (From Moss, D. W. [1982]. *Isoenzymes*. London: Chapman & Hall.)

A single polypeptide chain, in theory, exists in an infinite number of different conformations. However, one specific conformation generally appears to be the most stable for any given sequence of amino acids, and this conformation is assumed by the chain as it is synthesized within the cell. Thus the primary structure of the polypeptide chain also determines its 3D secondary and tertiary structures. It is conceivable that in some cases, several alternative conformations (conformers) of a single chain that are almost equally stable may be present, and therefore these alternative forms may coexist. However, multiple-enzyme forms that demonstrate conformational isomerism have not been shown unequivocally.

Changes in Isoenzyme Distribution During Development and Disease

The patterns of many isoenzymes change during physiologic development in tissues from many species. For example, during the embryonic development of skeletal muscle, the proportions of more cationic isoenzymes—both LD and CK—progressively increase until approximately the sixth month of intrauterine life when the pattern resembles that of differentiated muscle.

The liver also shows characteristic changes in the patterns of several isoenzymes during embryogenesis. In early fetal development, three aldolase isoenzymes—A, B, and C—together with various hybrid tetramers, can be detected in extracts of liver. However, at birth, as in the adult liver, aldolase B is the predominant isoenzyme. Striking changes in the distribution of isoenzymes of alcohol dehydrogenase also occur in human liver during prenatal development.

Changes in isoenzyme patterns during development result from changes in the relative activities of different genes within developing cells of a particular type (e.g., muscle cells). Other alterations in the balance of isoenzymes within the whole organism may derive from changes in the number or activity of cells that contain large amounts of a characteristic isoenzyme. An example of this is the increased number and activity of the osteoblasts, which are responsible for mineralization of the skeleton between the early postnatal period and the beginning of the third decade of life. An excess of ALP from active osteoblasts enters the circulation, where its presence can be recognized by its characteristic properties and where it increases the total serum ALP activity of young individuals to values higher than in skeletally mature adults. An ALP from the liver also contributes to the total activity of this enzyme in the plasma of healthy individuals, and the amount of this form in plasma shows a small, progressive increase with age. The reason for the latter age-dependent change is not known, but it may result from increased synthesis of the enzyme by hepatocytes in response to continuing exposure to inducing factors.

The distributions of isoenzymes of aldolase, LD, and CK in the muscles of patients with progressive muscular dystrophy are similar to those in the early development of fetal muscle. Isoenzyme abnormalities in dystrophic muscle have been interpreted as failure to reach or maintain a normal degree of differentiation. Isoenzyme patterns seen in regenerating tissues also may show some tendency to approach fetal distributions. This tendency may result from relaxation or modification of control systems in rapidly dividing cells and may account for some of the isoenzyme changes noted (e.g., in muscle in acute polymyositis).

Cancer cells show progressive loss of the structure and metabolism of the healthy cells from which they arise. Therefore the pattern of isoenzymes of mature, differentiated tissue may be lost or modified if normal differentiation is arrested or reversed; and many examples of isoenzyme changes accompanying cancer have been reported. Reemergence of fetal patterns of isoenzyme distribution is a feature of malignant transformation in many tissues. This phenomenon was first studied extensively in the case of LD isoenzymes. Malignant tumors in general show a significant shift in the balance of isoenzymes toward LD-4 and LD-5. The decline in activity of the LD-1 and LD-2 isoenzymes results in patterns that are reminiscent of those occurring in embryonic tissues. Tumors of prostate, cervix, breast, brain, stomach, colon, rectum, bronchus, and lymph nodes are among those that show this transformation. In contrast, comparatively benign gliomas show a relative increase in LD-1 and LD-2. A relative increase in the proportion of LD-4 and LD-5 has been observed in tissue adjacent to malignant tumors (e.g., the colon), although the cells in these regions are morphologically normal.

Differences in Properties Among Multiple Forms of Enzymes

Structural differences among the multiple forms of an enzyme give rise to differences in physicochemical properties, such as electrophoretic mobility, resistance to inactivation, and solubility, or in catalytic characteristics, such as preferential reactions with substrate analogs or natural substrates, or relative susceptibility to inhibitors. Methods of isoenzyme analysis have therefore been designed to investigate a wide range of catalytic and structural properties of enzyme molecules. However, it is usually possible to make only limited deductions about the nature of the underlying structural differences between isoenzymes that are responsible for the dissimilar properties. Equally, the changes in properties that may result from specific structural alterations in enzyme molecules are difficult to predict.

Techniques of molecular biology, such as gene cloning and sequencing, have revolutionized investigation of the primary structures of isoenzymes. Differences in primary structures among isoenzymes, whether derived from multiple genes or from different alleles, are now known to exist in a growing number of cases. Furthermore, many questions have been answered about whether multiple-enzyme forms represent true (genetically determined) isoenzymes or result from post-translational modification.

Isoenzymes caused by the existence of multiple genes usually differ quantitatively in catalytic properties. These differences may be manifested in such characteristics as molecular activity, K_m values for substrate(s), sensitivity to various inhibitors, and relative rates of activity with substrate analogs

(when the specificity of the isoenzymes allows the substrate to be varied), underscoring the biological importance of isoenzymatic variation. In contrast, multiple-enzyme forms that arise by posttranslational modifications such as aggregation usually have similar catalytic properties.

Multigene isoenzymes also usually differ in terms of antigenic specificity, although these differences may be less pronounced among isoenzymes that have emerged relatively recently in evolutionary history and are closely related in structure. Immunologic cross-reaction is not uncommon among multigene isoenzymes. Multiple-enzyme forms caused by posttranslational modification frequently have common antigenic determinants. Isoenzymes derived from different alleles of the same gene (allozymes) are also often antigenically similar, even to the extent that they may cross-react with antisera to the common isoenzyme, despite a mutation having abolished enzyme activity altogether. The capacity for detecting differences among antigenically similar isoenzyme molecules depends on the extent of monoclonal antibody specificity.

Differences in resistance to denaturation (e.g., by heat, concentrated urea solutions, detergents) are commonly found among true isoenzymes, whether these are the products of multiple genes or multiple alleles. Other multiple forms of enzymes often do not differ or differ only slightly in this respect. The most commonly exploited difference among isoenzymes is the difference in net molecular charge that results from the altered amino acid compositions of the molecules; this forms the basis of separation by zone electrophoresis, ion-exchange chromatography, or isoelectric focusing. Separation methods that depend on differences in molecular size, such as gel filtration, do not distinguish among the small size differences that often exist among true isoenzyme molecules but are important in the detection of multiple forms that involve aggregation or association of enzyme molecules with other proteins, such as immunoglobulins.

ENZYMES AS CATALYSTS

A **catalyst** is a substance that modifies and increases the rate of a particular chemical reaction without being consumed or permanently altered. Enzymes are protein catalysts of biological origin. Metabolism is a coordinated series of chemical reactions that occur within a living cell to provide energy and accomplish biosynthesis. The process can be regarded as an integrated series of enzymatic reactions, whereas some diseases can be viewed as a derangement of the normal pattern of metabolism. The remarkable properties of enzymes make them sensitive indicators of pathologic change.

Because of their catalytic activity, enzymes convert an enormous number of substrate molecules into products within a short time. This property is used to measure increased amounts of enzymes in the bloodstream, although the amount of enzyme protein released from damaged cells is small compared with the total quantity of nonenzymatic proteins in blood.

UNITS FOR EXPRESSING ENZYME ACTIVITY

When enzymes are measured by their catalytic activities, the results are expressed as *units of activity*. The unit of activity is defined by the rate at which the reaction proceeds (e.g., the quantity of substrate consumed or product formed in a chosen unit of time). In **clinical enzymology**, enzyme activity is generally reported as units of activity per unit of volume, such as units per 100 mL, per liter of serum, or per 1.0 mL of packed erythrocytes. Because the rate of the reaction depends on experimental parameters, such as pH, type of buffer, temperature, nature of substrate, ionic strength, concentration of activators, and other variables, these parameters must be specified in the definition of the unit.

To standardize how enzyme activities are expressed, the EC proposed that the unit of enzyme activity should be defined as the quantity of enzyme that catalyzes the reaction of 1 μmol of substrate per minute and that this unit should be termed the **international unit** (U). Catalytic concentration is expressed in terms of units per liter (U/L) or kilounits (kU/L), whichever gives the more convenient numeric value. In this chapter, the symbol U is used to denote the international unit. In those instances in which there is some uncertainty about the exact nature of the substrate, or when difficulty is encountered in calculating the number of micromoles reacting (as with macromolecules such as starch, protein, and complex lipids), the unit is expressed in terms of the chemical group or residue produced in the reaction (e.g., μmol/min glucose or amino acid formed).

Contrary to the previously noted EC recommendation, the International System of Units (SI) recommends that enzyme activity be expressed in moles per second or **katals**. Enzyme concentration can then be expressed in terms of katals per liter (kat/L). The relationship between the EC and SI measures of enzymatic activity is:

$$\begin{aligned} 1\ \text{SI Unit (katal)} &= 1\ \text{mol/s} = 10^6\,\mu\text{mol/s} \\ &= 6\times10^7\,\mu\text{mol/min} = 6\times10^7\,\text{EC Units} \end{aligned}$$

The katal is so large that it is difficult to use and therefore, the EC units of enzyme activity (μmol/min) are commonly employed.

POINTS TO REMEMBER

- Enzymes are biological catalysts that can be used in the monitoring and diagnosis of disease.
- Enzymes are proteins with the ability to accelerate or catalyze chemical reactions.
- The characteristics of the reactions catalyzed by an enzyme depend on the structure of its active site and its relationship to substrates.
- Enzymes exist in multiple forms, isoenzymes, the properties and distribution of which are important in diagnosis.

ENZYME KINETICS

Enzyme-Substrate Complex

Enzymes act through the formation of an enzyme–substrate (ES) complex, in which a molecule of substrate is bound to

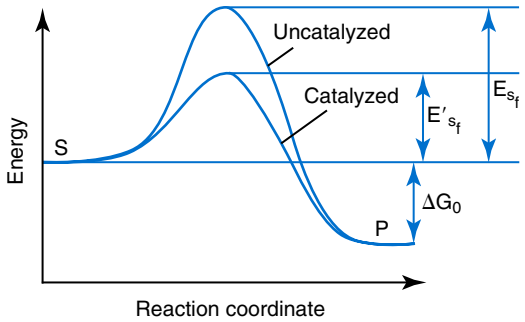


Fig. 14.3 Activation energy barrier and reaction course, with and without enzyme catalysis. E_{S_f} is the activation energy for the forward reaction (S → P) in the absence of a catalyst, and E'_{S_f} is the activation energy in the presence of a catalyst. ΔG_0 is the change in free energy for the reaction.

the active center of the enzyme molecule. The binding process transforms the substrate molecule to its activated state. The energy required for this transformation is provided by the free energy of binding of S to E. Therefore activation takes place without the addition of external energy, so that the energy barrier to the reaction is lowered and the conversion to products is accelerated (Fig. 14.3). The ES complex dissociates to give the reaction products (P) and free enzyme (E):

$$E + S \leftrightarrow ES \rightarrow P + E \qquad \textbf{(14.1)}$$

All reactions catalyzed by enzymes are in theory reversible. However, under the conditions found in the human body in vivo, or in the clinical lab in vitro, the reaction is usually more rapid in one direction than in the other, so that an equilibrium is reached in which the product of the forward or the backward reaction predominates, sometimes so markedly that the reaction is virtually irreversible.

If the product of the reaction in one direction is removed as it is formed (e.g., because it is the substrate of a second enzyme present in the reaction mixture), the equilibrium of the first enzymatic process will be displaced so that the reaction will proceed to completion in that direction. Reaction sequences in which the product of one enzyme-catalyzed reaction becomes the substrate of the next enzyme and so on, often through many stages, are characteristic of biological processes. In the laboratory also, several enzymatic reactions may be linked together to provide a means of measuring the activity of the first enzyme or the concentration of the initial substrate in the chain. For example, the activity of CK is usually measured by coupling the adenosine triphosphate (ATP) formed in the CK-catalyzed reaction to the phosphorylation of glucose catalyzed by hexokinase, and then coupling the glucose-6-phosphate product to oxidation by glucose-6-phosphate dehydrogenase, with the formation of a photometrically measurable product, nicotinamide adenine dinucleotide phosphate (NADPH).

When a secondary enzyme-catalyzed reaction, known as an indicator reaction, is used to determine the activity of a different enzyme, the primary reaction must be the rate-limiting step. Reaction conditions must ensure that the rate of reaction catalyzed by the indicator enzyme is directly proportional to the rate of product formation in the first reaction.

FACTORS GOVERNING THE RATE OF ENZYME-CATALYZED REACTIONS

Factors that affect the rate of enzyme-catalyzed reactions include enzyme and substrate concentration, pH, temperature, and the presence of inhibitors, activators, **coenzymes**, and prosthetic groups.

Enzyme Concentration

The simplest enzymatically catalyzed reaction for converting substrate S into product P with the intermediate formation of an ES complex is as follows:

$$E_f + S \underset{k_1}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_2} E_f + P \qquad \textbf{(14.2)}$$

where E_f = free enzyme; k_1 = rate constant for the association of the complex; k_1 = rate constant for the dissociation of the complex; ES = enzyme–substrate complex; k_2 = rate constant for breakdown of ES to E_f and P; and P = product.

Michaelis and Menten assumed that equilibrium is attained rapidly among E, S, and ES, with the effect of product formation (ES → P) on the concentration of ES being negligible. In addition, the formation of product is written as an irreversible process because there is no product in the solution under initial conditions. Therefore the overall rate of the reaction under otherwise constant conditions is proportional to the concentration of the ES complex.

Provided that an excess of free substrate molecules is present, addition of more enzyme molecules to the reaction system increases the concentration of the ES complex and thus the overall rate of reaction. This accounts for the observation that the rate of reaction is generally proportional to the concentration of enzyme present in the system and is the basis for the quantitative determination of enzymes by measurement of reaction rates. Reaction conditions are selected to ensure that the observed reaction rate is proportional to enzyme concentration over as wide a range as possible.

Substrate Concentration

In addition to explaining the dependence of reaction rate on enzyme concentration under conditions in which excess substrate is present, the formation of an ES complex accounts for the hyperbolic relationship between reaction velocity and substrate concentration (Fig. 14.4A).

Single-Substrate Reactions

If the enzyme concentration is fixed and the substrate concentration is varied, the rate of reaction is first order and proportional to substrate concentration when the substrate concentration is low. Under these conditions, only a fraction of the enzyme is associated with substrate, and the rate observed reflects the low concentration of the ES complex. At high substrate concentrations, variation in substrate concentration has no effect on rate, and the reaction is zero order with respect to substrate concentration. Under these conditions, all the enzyme is bound to the substrate and a much higher rate of reaction is obtained. Moreover, because all the enzyme is now present in the form of the complex, no further

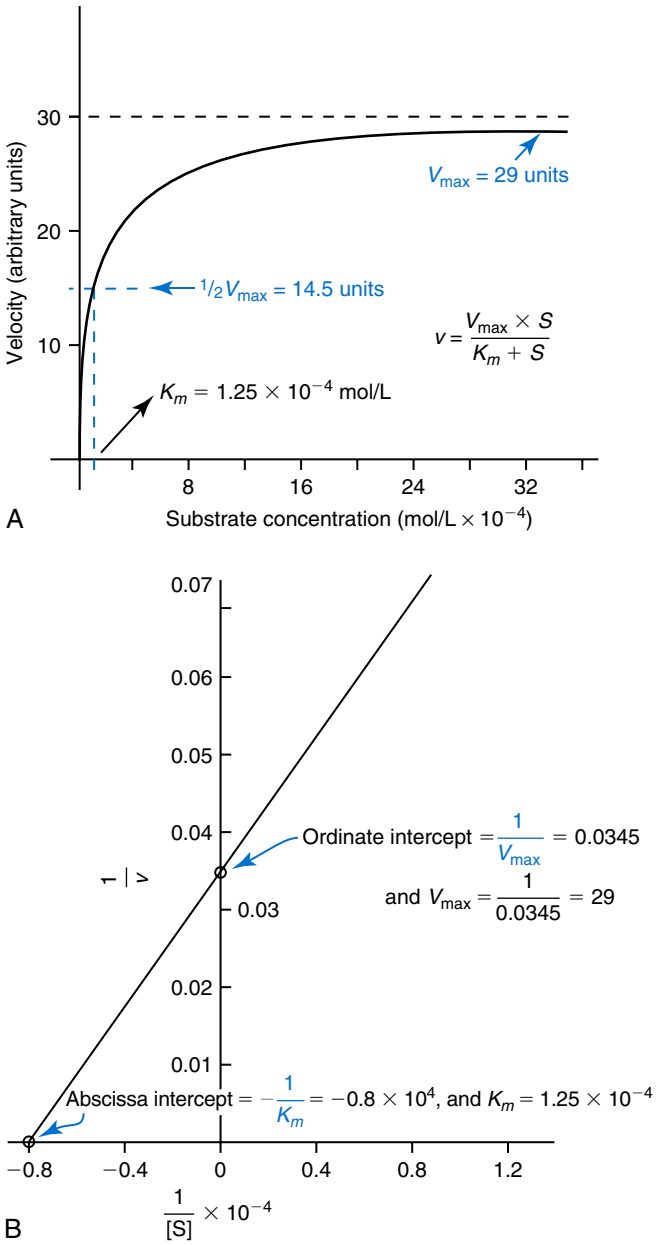


Fig. 14.4 (A) Michaelis-Menten curve relating the velocity (rate) of an enzyme-catalyzed reaction to substrate concentration. The value of K_m is given by the substrate concentration at which one-half of the maximum velocity is obtained. (B) Lineweaver-Burk transformation of the curve in (A), with $1/v$ plotted on the ordinate (y-axis), and 1/[S] on the abscissa (x-axis). The indicated intercepts permit calculation of V_{max} and K_m.

increase in complex concentration and no further increment in reaction rate are possible. The maximum possible velocity for the reaction has been reached. The significance of substrate-rate curves was first emphasized by Michaelis and Menten, and such curves are referred to as Michaelis-Menten plots.

Referring again to Eq. 14.2, the overall rate of the reaction (v) is determined by the rate at which product is formed:

$$v = \frac{d[P]}{dt} = k_2 [ES] \quad \textbf{(14.3)}$$

The formation of ES will depend on the rate constant k_1 and the availability of enzyme and substrate. If it is assumed that the system is in a steady state with the ES complex being formed and broken down at the same rate so that overall (ES) is constant, then the steady-state equation is

$$k_1 [E][S] = k_{-1} [ES] + k_2 [ES] \quad \textbf{(14.4)}$$

This equation can be rearranged to:

$$[ES] = \frac{k_1 [E][S]}{k_{-1} + k_2} \quad \textbf{(14.5)}$$

when these rate constants are combined into a single term; writing the **Michaelis-Menten constant (K_m)** as:

$$K_m = \frac{k_{-1} + k_2}{k_1} \quad \textbf{(14.6)}$$

and then substituting this into Eq. 14.5 gives:

$$[ES] = \frac{[E][S]}{K_m} \quad \textbf{(14.7)}$$

Because the total amount of enzyme in the system does not change,

$$[E_t] = [ES] + [E] \quad \textbf{(14.8)}$$

and on substituting Eq. 14.3 into Eq. 14.7 and eliminating [ES] using Eq. 14.8,

$$v = k_2 \times \frac{[E_t] \times [S]}{K_m + [S]} \quad \textbf{(14.9)}$$

For a given amount of enzyme, the maximum reaction velocity (V_{max}) is reached when all of the enzyme is saturated with substrate (i.e., [ES] = [Et]), and therefore $V_{max} = k_2 \times$ [Et]. Substituting this in Eq. 14.9 gives

$$v = \frac{V_{max} [S]}{K_m + [S]} \quad \textbf{(14.10)}$$

A plot of v against [S] gives a section of a rectangular hyperbola (see Fig. 14.4A), and this is the shape of the curve that is found experimentally for most enzyme reactions. When [S] = K_m, K_m is the substrate concentration at which the reaction proceeds at half of its maximum velocity. In practice, it is now customary to restrict K_m to the experimentally determined substrate concentration at which $v = 0.5\ V_{max}$ and to use the symbol K_s to represent the true ES association constant, where this is known.

Although it is simple to set up an experiment to determine the variation of v with [S], the exact value of V_{max} is not easily determined from hyperbolic curves. Furthermore, many enzymes deviate from ideal behavior at high substrate concentrations and may even be inhibited by the excess of substrate, so the calculated value of V_{max} cannot be achieved in practice. In the past, it was common practice to transform the Michaelis-Menten equation (Eq. 14.10) into one of several reciprocal forms (Eqs. 14.11 and 14.12), and either $1/v$ was plotted against 1/[S], or [S]/v was plotted against [S].

$$\frac{1}{v} = \left(\frac{K_m}{V_{max}} \times \frac{1}{[S]}\right) + \frac{1}{V_{max}} \quad \textbf{(14.11)}$$

$$\frac{[S]}{v} = \left(\frac{1}{V_{max}} \times [S]\right) + \frac{K_m}{V_{max}} \quad \textbf{(14.12)}$$

Eq. 14.11, for example, when plotted, results in a **Lineweaver-Burk plot** that gives a straight line with intercepts at $1/V_{max}$ on the ordinate and $-1/K_m$ on the abscissa. The data for Fig. 14.4A are replotted in Lineweaver-Burk form in Fig. 14.4B.

It is now routine practice to determine kinetic constants such as K_m and V_{max} using software packages. Free and commercial packages are available; these vary from specialized routines for kinetic simulations or for data fitting, to general mathematical, statistical, or graphical packages.

The value of K_m, is to compare the binding of homologous or related substrates to the same enzyme. Also, if measured against the same substrate under defined conditions, the K_m value can be used to compare the properties of similar enzymes from different sources. Isoenzymes determined by distinct genes typically differ in their K_m (e.g., for the isoenzymes of LD).

When setting up methods of enzyme assay, it is necessary to (1) explore the relationship between reaction velocity and substrate concentration over a wide range, (2) determine K_m, and (3) detect any inhibition at high substrate concentrations. **Zero-order kinetics** is maintained if the substrate is present in large excess (i.e., concentrations at least 10 and preferably 100 times that of the value of K_m). When $[S] = 10 \times K_m$, v is approximately 91% of the theoretical V_{max}. The K_m values for the majority of enzymes are on the order of 10^{-5} to 10^{-3} mol/L; therefore substrate concentrations are usually chosen to be in the range of 0.001 to 0.10 mol/L. On occasion, the optimal concentrations of substrate cannot be used, for example, when the substrate has limited solubility or when the concentration of a given substrate inhibits the activity of another enzyme needed in a coupled reaction system.

Two-Substrate Reactions

Most enzymes catalyze reactions with two or more interacting substrates symbolized by the following equation:

$$\underset{S_1}{\text{Substrate 1}} + \underset{S_2}{\text{Substrate 2}} \overset{E}{\leftrightarrow} \underset{P_1}{\text{Product 1}} + \underset{P_2}{\text{Product 2}} \qquad (14.13)$$

When one of the substrates is water (i.e., when the process is one of hydrolysis), with the reaction taking place in aqueous solution, only a fraction of the total number of water molecules present participates in the reaction. The small change in the concentration of water has no effect on the rate of reaction and these pseudo–one substrate reactions are described by one-substrate kinetics. More generally, the concentrations of both substrates may be variable and both may affect the rate of reaction. Among the bisubstrate reactions important in clinical enzymology are the reactions catalyzed by dehydrogenases, in which the second substrate is a specific coenzyme, such as the oxidized or reduced forms of nicotinamide adenine dinucleotide (NADH), or NADPH, and the amino-group transfers catalyzed by the aminotransferases.

If a bisubstrate reaction proceeds by way of intermediate ES complexes, so that

$$E + S_1 \leftrightarrow ES_1 \qquad (14.14)$$

followed by

$$ES_1 + S_2 \leftrightarrow ES_1S_2 \rightarrow P_1 + P_2 + E \qquad (14.15)$$

and if S_1 and S_2 bind to separate sites on the enzyme molecule, the rate of reaction is given by

$$v = \frac{V_{max} \times [S_1][S_2]}{[S_1][S_2] + [S_2]K_m^1 + [S_1]K_m^2 + K_S^1 K_m^2} \qquad (14.16)$$

K_m^1 and K_m^2 are the K_m values for the two substrates, and $[S_1]$ and $[S_2]$ are their concentrations. K_s^1 is the equilibrium constant for the reversible reaction between the enzyme and S_1. If the equation is rearranged into the double reciprocal form,

$$\frac{1}{v} = \frac{1}{[S_1]}\left(\frac{K_m^1}{V_{max}} + \frac{K_m^2 K_S^1}{[S_2]V_{max}}\right) + \frac{1}{V_{max}}\left(1 + \frac{K_m^2}{[S_2]}\right) \qquad (14.17)$$

a plot of $1/v$ against $1/[S_1]$ gives a straight line, but both the slope of the line and its intercept on the ordinate are dependent on $[S_2]$, the concentration of the second substrate. A plot of $1/v$ against $1/[S_2]$ is rectilinear but with the slope and intercept dependent on $[S_1]$.

In some bisubstrate reactions, no ternary complex ES_1S_2 is formed, because the binding of the first substrate is followed by release of the first product before the second substrate is bound and the second product is released. This sequence is described as a *ping-pong bi-bi* type of reaction. This occurs in reactions catalyzed by aminotransferases.

The relationship between reaction velocity and the concentrations of the two substrates in *ping-pong bi-bi* reactions reduces to the form

$$v = \frac{V_{max} \times [S_1][S_2]}{[S_1][S_2] + [S_2]K_m^1 + [S_1]K_m^2} \qquad (14.18)$$

The reciprocals of v and $[S_1]$ are related by the equation

$$\frac{1}{v} = \frac{1}{[S_1]} \times \frac{K_m^1}{V_{max}} + \frac{1}{V_{max}}\left(1 + \frac{K_m^2}{[S_2]}\right) \qquad (14.19)$$

so that a plot of $1/v$ against $1/[S_1]$ is unchanged in slope by variation in $[S_2]$, but by the intercept on the ordinate, and therefore the value of V_{max} changes as $[S_2]$ is varied. Similar equations describe the variation of V_{max} with $[S_1]$ when $1/v$ is plotted as a function of $1/[S_2]$.

Values of K_m and V_{max} for each substrate are derived from experiments in which the concentration of the first substrate is held constant at saturating quantities while the concentration of the second substrate is varied, and vice versa. The selection of reaction conditions for the measurement of enzymatic activity involving two substrates is approached similarly by varying the concentration of the first substrate and keeping the concentration of the second substrate constant until maximum activity is reached. The process is then repeated with the concentration of the first substrate held at the value thus determined, whereas the concentration of the second substrate is varied.

In practice, the choice of substrate concentrations is limited by such considerations as the solubility of the substrates, the

viscosity and initial absorbance of concentrated solutions, and the relative costs of the reagents. Furthermore, the selection of appropriate substrate concentrations is only one of the factors to consider in formulating an optimal assay system for the measurement of a certain enzyme activity. Additional, frequently interdependent factors that affect the reaction rate also include the concentrations of activators, inhibitors, the pH, and the nature of the buffer system.

Consecutive Enzymatic Reactions

As discussed previously, an enzymatic reaction is usually found to be more rapid in one direction than the other, so that the reaction is essentially irreversible. If the product of the reaction in one direction is removed as it is formed (i.e., because it is the substrate of a second enzyme present in the reaction mixture), the equilibrium of the first enzymatic process is displaced so that the reaction may continue to completion in that direction. Analytically, several enzymatic reactions may be linked together to provide a means of measuring the activity of the first enzyme or the concentration of the initial substrate in the chain.

When a linked enzyme assay, known as an *indicator reaction,* is used to determine the activity of a different enzyme, it is essential that the primary reaction be the rate-limiting step. For example, in the determination of AST activity, the indicator reaction is the reduction of the 2-oxoglutarate formed in the aminotransferase reaction to malate by malate dehydrogenase and NADH. The activity of the indicator enzyme must be sufficient to ensure the immediate removal of the product of the first reaction so as to prevent significant reversal of the first reaction. The measured enzyme typically acts under conditions of saturation with respect to its substrate; however, the concentration of the substrate of the indicator enzyme (i.e., the product of the first reaction) remains in the region of the Michaelis-Menten curve, in which v is directly proportional to [S]. Therefore the rate of reaction catalyzed by the indicator enzyme is directly proportional to the rate of product formation in the first reaction.

During a lag period that occurs after the start of the first reaction, the concentration of its product reaches a steady state. Because the rate of the second reaction depends on the activity of the indicator enzyme and on the concentration of its substrate (the product of the primary reaction), the duration of the lag period is reduced by increasing the concentration of the indicator enzyme, thus lowering the steady-state concentration of the product of the primary reaction.

The rate of the indicator reaction, v_i, is related to substrate concentration and therefore to the product concentration [P] by the Michaelis-Menten equation

$$v_i = \frac{V^i_{max} \times [P]}{[P] + K^i_m} \quad \textbf{(14.20)}$$

in which V^i_{max} and K^i_m are the maximum velocity and K_m of the indicator enzyme, respectively. For the rate of the indicator reaction not to be the rate limiting factor, v_i must at least equal the limiting velocity of the primary reaction, v_t, which the assay system is expected to measure. Therefore, the minimum activity of the indicator enzyme needed is given by

$$v_t = \frac{V^i_{max} \times [P]}{[P] + K^i_m} \quad \textbf{(14.21)}$$

which can be rearranged to

$$V^i_{max} = v_t \left(1 + \frac{K^i_m}{[P]}\right) \quad \textbf{(14.22)}$$

The ratio of activities of the indicator and primary enzymes varies from one assay method to another, depending on (1) the range of activity measured, (2) the K_m of the indicator enzyme, and (3) the lag period that is considered acceptable. Nevertheless, the catalytic concentration of the indicator enzyme in the reaction mixture must always be much greater than that of the enzyme being determined.

Effect of pH

The rate of an enzyme-catalyzed reaction typically shows a marked dependence on pH (Fig. 14.5). Many of the enzymes in blood plasma show maximum activity in vitro in the pH range from 7 to 8. However, activity has been observed at pH values as low as 1.5 (pepsin) and as high as 10.5 (ALP). The optimal pH for a given forward reaction may be different from the optimal pH found for the corresponding reverse reaction. The form of the pH-dependence curve is the result of a number of separate effects, including ionization of the substrate and the extent of dissociation of certain key amino acid side chains in the protein molecule, both at the active center and elsewhere in the molecule. The pH and ionic environment will also have an effect on the 3D conformation of the protein and therefore on enzyme activity to such an extent that enzymes may be irreversibly denatured at extreme pH values.

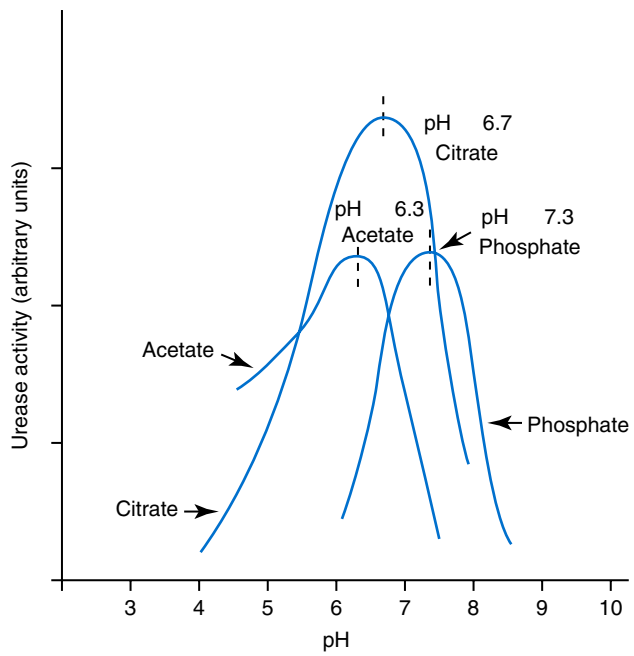


Fig. 14.5 The pH activity curves for urease show the effects of buffer species on pH optimum. (Modified from Howell, S. F., & Sumner, J. B. [1934]. The specific effects of buffers upon urease activity. *Journal of Biological Chemistry, 104,* 619.)

The pronounced effects of pH on enzyme reactions emphasize the need to control this variable by means of adequate buffer solutions. Enzyme assays should be carried out at the pH of optimal activity, because the pH-activity curve has its minimum slope near this pH, and a small variation in pH will cause a minimal change in enzyme activity. The buffer system must be capable of counteracting the effect of adding the specimen (e.g., serum itself is a powerful buffer) to the assay system, as well as the effects of acids or bases formed during the reaction (e.g., formation of fatty acids by the action of lipase). Because buffers have their maximum buffering capacity close to their pK_a values, whenever possible a buffer system should be chosen with a pK_a value within 1 pH unit of the desired pH of the assay. Interaction between buffer ions and other components of the assay system (e.g., the chelation of activating metal ions) may eliminate certain buffers from consideration.

Temperature

The rate of an enzymatic reaction is proportional to its reaction temperature, but only up to a certain point. For most enzymatic reactions, values of Q_{10} (the relative reaction rates at two temperatures differing by 10°C) vary from 1.7 to 2.5. However, an increase in the rate of the catalyzed reaction is not the only effect of increasing temperature on an enzymatic reaction. Enzyme inactivation, due to protein denaturation as the temperature increases, accounts for the existence of an apparent optimal temperature for enzyme activity (Fig. 14.6).

At some critical temperature, an enzyme will undergo thermal inactivation influenced by the presence of substrate and its concentration, the pH, and the nature and ionic strength of the buffer. The presence of other proteins, as in serum samples, may help to stabilize enzymes. Storage of serum samples at low temperatures is necessary to minimize loss of enzyme activity while awaiting analysis, although repeated freezing and thawing should be avoided. However, individual enzymes vary in their stability characteristics, and appropriate storage conditions vary correspondingly. Serum amylase, for example, is stable at room temperature (22°C to 25°C) for 24 hours, whereas acid phosphatase is exceedingly unstable, even when refrigerated, unless kept at a pH below 6.0. ALP exhibits an unusual property; in sera stored at refrigerated temperatures, ALP increases slowly (2% per day), which is thought to be due to the reincorporation of cations required for full activity. Frozen specimens should be thawed and kept at room temperature for 18 to 24 hours before measurement to achieve full enzyme reactivation. This effect is shared by reconstituted, lyophilized preparations of the enzyme and affects their use for quality control purposes. A few enzymes are inactivated at refrigerator temperatures; a clinically important example is the liver isoenzyme of LD, LD-5, which appears to be less stable at lower temperatures. As a result, sera for LD determinations should be kept at room temperature and not refrigerated.

Most analytical systems operate at 37°C, and reference methods for several clinically relevant enzymes have now been developed at this temperature. Accurate and precise temperature control to within ±0.1°C during the enzymatic reaction is essential, and is achievable by modern instruments.

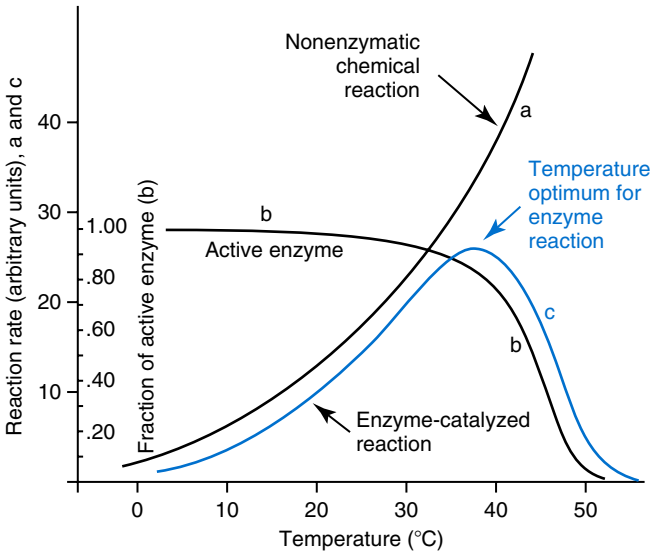


Fig. 14.6 Schematic diagram showing the effects of temperature on the rate of non–enzyme-catalyzed and enzyme-catalyzed reactions.

Inhibitors and Activators

The rates of enzymatic reactions are often affected by substances other than the enzyme or substrate. These modifiers may be **inhibitors** that reduce the reaction rate, or **activators** that increase the rate of reaction. Activators and inhibitors are usually small molecules (compared with the enzyme itself) or even ions. They vary in specificity from modifiers that exert similar effects on a wide range of different enzymatic reactions, to substances that affect only a single reaction. The activity of some enzymes depends on the presence of particular chemical groups, such as reduced sulfhydryl (–SH) groups, in the active center. Reagents that alter these groups (e.g., oxidizing agents) therefore act as general inhibitors of such enzymes.

Some phenomena of enzyme activation or inhibition are caused by interactions between the modifier and a nonenzymatic component of the reaction system, such as the substrate (e.g., magnesium [Mg^{2+}] combining with ATP to form MgATP, a required substrate for the CK reaction). In most cases, however, the modifier binds to the enzyme itself, either to the active site, or to some other part of the enzyme's structure.

Inhibition of Enzyme Activity

Inhibitors are classified as reversible or irreversible. *Reversible inhibition* implies that the activity of the enzyme is fully restored when the inhibitor is removed from the system by some physical separative process, such as dialysis, gel filtration, or chromatography. An *irreversible inhibitor* combines covalently with the enzyme so that physical methods are ineffective in separating the two. For example, organophosphorus compounds are extremely potent irreversible inhibitors of esterases, including

acetylcholinesterase, which is the mechanism for organophosphate poisoning. The enzyme breaks one of the bonds in the inhibitor, but part of the molecule is left bound to the active center of the enzyme, preventing further activity. In some cases, enzymes that have combined with irreversible inhibitors can be reactivated by a chemical reaction that removes the blocking group (e.g., the phosphoryl enzymes formed with organophosphorus compounds can sometimes be reactivated by treatment with oximes or hydroxamic acids).

Reversible inhibition. Reversible inhibition is characterized by the existence of equilibrium between enzyme, E, and inhibitor, I:

$$E + I \leftrightarrow EI \tag{14.23}$$

The equilibrium constant of the reaction, K_i (the *inhibitor constant*), is a measure of the affinity of the inhibitor for the enzyme, just as K_m generally reflects the affinity of the enzyme for its substrate.

A *competitive* inhibitor is usually a structural analog of the substrate that can combine with the free enzyme in such a way that it competes with the normal substrate for binding at the active site. The rate of the reaction in the presence of a reversible inhibitor is strictly dependent on the relative concentrations of substrate and inhibitor. Two equilibriums are therefore possible:

$$E + S \leftrightarrow ES \rightarrow E + \text{Products} \tag{14.24}$$

and

$$E + I \leftrightarrow EI \tag{14.25}$$

The equation that relates the observed reaction velocity to the concentrations of substrate, [S], and inhibitor, [I], is as follows:

$$v = \frac{V_{max}[S]}{[S] + K_m\left(1 + \frac{[I]}{K_i}\right)} \tag{14.26}$$

This is the Michaelis-Menten equation, but with K_m modified by a term including the inhibitor concentration and the inhibitor constant, while V_{max} remains unaltered. Therefore curves of v against [S] in the presence and absence of the inhibitor reach the same limiting value at high substrate concentrations, but when the inhibitor is present, K_m is accordingly greater. Plots of $1/v$ against $1/[S]$ with and without inhibitor intersect the ordinate at the same point but have different slopes and intercepts on the abscissa (Fig. 14.7).

Competitive inhibition is responsible for the inhibition of some enzymes by excess substrate because of competition between substrate molecules for a single binding site. In two-substrate reactions, high concentrations of the second substrate may compete with binding of the first substrate. For example, AST is inhibited by excess concentrations of the substrate 2-oxoglutarate, and this inhibition is competitive with respect to L-aspartate. Therefore to maintain a given velocity at high 2-oxoglutarate concentrations, the concentration of L-aspartate has to be increased above the value needed at lower concentrations of 2-oxoglutarate.

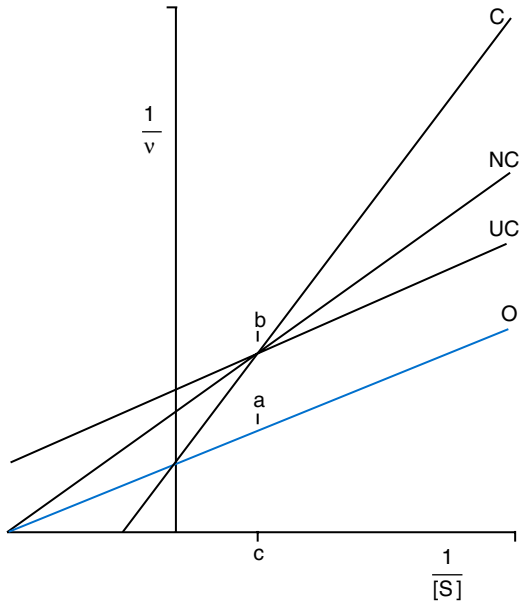


Fig. 14.7 Effects of different types of inhibitors on the double-reciprocal plot of $1/v$ against $1/[S]$. Each of the inhibitors has been assumed to reduce the activity of the enzyme by the same amount, represented by the change in $1/v$ from *a* to *b* at a substrate concentration of *c*. Line *O* is the plot for enzyme without inhibitor, *C* with a competitive inhibitor, *NC* with a noncompetitive inhibitor, and *UC* with an uncompetitive inhibitor. (From Moss, D. W. [1979]. Measurement of enzymes. In: D. J. Hearse & J. de Leiris [Eds.], *Enzymes in cardiology: diagnosis and research.* New York: John Wiley & Sons. Reprinted by permission of John Wiley & Sons.)

Competitive inhibition also contributes to the reduction in the rate of an enzymatic reaction over time. For example, rate reduction can occur because increasing concentrations of reaction products tend to drive the reaction backward, if it is freely reversible. A product may be an inhibitor of the forward reaction, so even if the reaction is not readily reversible, it slows down in the presence of a rising concentration of the inhibitor. A familiar example of product inhibition is the release of the competitive inhibitor inorganic phosphate by the action of ALP on its substrates. In this case, both organic phosphates and inorganic phosphates bind to the active center of the enzyme with similar affinities (i.e., K_m and K_i are of the same order of magnitude).

Product inhibition is a cause of nonlinearity of reaction progress curves during fixed-time methods of enzyme assay. For example, oxaloacetate produced by the action of AST inhibits the enzyme, particularly the mitochondrial isoenzyme. The inhibitory product may be removed as it is formed by a coupled enzymatic reaction; malate dehydrogenase converts the oxaloacetate to malate and at the same time oxidizes NADH to NAD^+.

Competitive inhibition by metal ions occurs when two metal ions compete for the same binding site on the enzyme. Sodium and lithium are potent inhibitors of pyruvate kinase, for which potassium is an obligatory activator.

A noncompetitive inhibitor is usually structurally different from the substrate. It is assumed to bind at a site on the enzyme molecule that is different from the substrate-binding

site; thus there is no competition between inhibitor and substrate, and a ternary enzyme substrate–inhibitor (ESI) complex is formed. Attachment of the inhibitor to the enzyme does not alter the affinity of the enzyme for its substrate (i.e., K_m is unaltered), but the ESI complex does not break down to give products. Because the substrate does not compete with the inhibitor for binding sites on the enzyme molecule, increasing the substrate concentration does not overcome the effect of a noncompetitive inhibitor. Thus V_{max} is reduced in the presence of such an inhibitor, whereas K_m is not altered, as the Lineweaver-Burk plot shows (see Fig. 14.7).

Uncompetitive inhibition is produced by a combination of the inhibitor with the ES complex. It is more common in two-substrate reactions, in which a ternary ESI complex forms after the first substrate has combined with the enzyme. In uncompetitive inhibition, parallel lines are obtained when plots of $1/v$ against $1/[S]$ with and without the inhibitor are compared (see Fig. 14.7); that is, both K_m and V_{max} are decreased.

Irreversible inhibition. Irreversible inhibitors render the enzyme molecule inactive by covalently and permanently modifying a functional group required for catalysis. An irreversible inhibitor is not in equilibrium with the enzyme. Its effect is progressive with time, becoming complete if the amount of inhibitor present exceeds the total amount of enzyme. The rate of the reaction between the enzyme and the inhibitor is expressed as the fraction of the enzyme activity that is inhibited in a fixed time by a given concentration of inhibitor. The velocity constant of the reaction of the inhibitor with the enzyme is a measure of the effectiveness of the inhibitor.

When the inhibitor is added to the enzyme in the presence of its substrate, the reaction between the enzyme and the inhibitor may be delayed because some of the enzyme molecules are combined with the substrate and are therefore protected from reacting with the inhibitor. However, when the substrate molecules react chemically, the active centers become available, and inhibition will eventually become complete, even though an excess of substrate may have been present initially. Furthermore, the addition of more substrate is ineffective in reversing the inhibition in contrast to its effect on reversible competitive inhibition.

Irreversible inhibitors have been useful in mapping active sites by covalently modifying different types of functional groups in the enzyme molecule to establish whether such groups are necessary for catalytic activity.

A physiologically important category of irreversible enzyme inhibition is exemplified by various trypsin inhibitors. These are proteins that bind to trypsin irreversibly, nullifying its proteolytic activity. One such inhibitor is present in the α_1-globulin fraction of serum proteins; others are found in soybeans and lima beans. Similar proteolysis inhibitors present in plasma prevent the accumulation of excess thrombin and other coagulation enzymes, thus keeping the coagulation process under control.

Inhibition by antibodies. The combination of enzyme molecules with specific antibodies often has no effect on catalytic activity, which is retained by the enzyme-antibody complex. However, in some cases, the reaction of the enzyme and antibody reduces or even abolishes enzymatic activity. The most probable explanation for this type of inhibition is that the antibody molecule restricts access of the substrate molecules to the active center by steric hindrance or, in extreme cases, completely masks the substrate-binding site. However, it appears that some examples of enzyme inhibition by combination with antibodies are caused by a conformational change induced in the enzyme molecule.

Inhibition of the activity of an enzyme molecule labeled with a hapten (e.g., morphine) as a result of combination with a specific antibody forms the basis for a homogeneous enzyme immunoassay (see Chapter 15).

Enzyme Activation

Activators increase the rates of enzyme-catalyzed reactions by promoting formation of the most active state of the enzyme or of other reactants, such as the substrate. This generalization covers a wide variety of mechanisms of activation.

Many enzymes contain metal ions as an integral part of their structures (e.g., zinc in ALP and carboxypeptidase A). The function of the metal may be to stabilize tertiary and quaternary protein structures. Removal of divalent metal ions by treatment with an appropriate quantity of ethylenediaminetetraacetic acid (EDTA) solution is accompanied by conformational changes and inactivation of the enzyme. The enzyme can often be reactivated by dialysis against a solution of the appropriate metal ion or simply by addition of the ion to the reaction mixture. Reactivation may take some time because rearrangement of polypeptide chains into the active conformation is not instantaneous.

When the activator ion is an essential part of the functional enzyme molecule, whether as a purely structural element or with an additional catalytic role, it is usually incorporated quite firmly into the enzyme molecule. Therefore it is not usually necessary to add the activator to reaction mixtures and an excess of the ion may even have an inhibitory effect. However, in some cases, the activating ion is attached only weakly or transiently to the enzyme (or its substrate) during catalysis. Enzyme samples may therefore be deficient in the ion, so that addition of the ion increases the reaction rate or indeed may be essential for the reaction to take place. For example, all phosphate transfer enzymes (kinases), such as CK, require the presence of Mg^{2+} ions. Other common activating cations are manganese (Mn^{2+}), iron (Fe^{2+}), calcium (Ca^{2+}), zinc (Zn^{2+}), and potassium (K^+). More rarely, anions may act as activators. Amylase functions at its maximal rate only if chloride (Cl^-) or other monovalent anions, such as bromide (Br^-) or nitrate (NO_3^-), are present. Addition of 5 mmol/L of chloride increases amylase activity almost threefold, at the same time shifting the pH optimum from 6.5 to 7.0. The Cl^- ion may combine with a positively charged group in the enzyme, changing the ionization constant of a group important in catalysis. Some enzymes require the presence of two activating ions. K^+ and Mg^{2+} are essential for the activity of pyruvate kinase, and both Mg^{2+} and Zn^{2+} are required for ALP activity.

The velocity of the reaction depends on the concentration of a reversible activator in a fashion similar to its dependence on substrate concentration. An activator constant, K_a, analogous to K_m, can be determined from data relating enzyme activity to increasing activator concentration in the presence of excess substrate. The simplest interpretation of K_a is that it is the dissociation constant of the equilibrium between E and the activator, A. However, this is true only when the combination of enzyme and activator is independent of the reaction between E and S, and the same value for K_a is obtained at all concentrations of the substrate. If the free enzyme and the ES complex have different affinities for the activator, the value for K_a varies with [S].

Apparent activation of an enzyme may be observed whenever a substance that can counteract the presence of some inhibiting agent is added.

Coenzymes and Prosthetic Groups

Coenzymes are usually more complex molecules than activators, although they are smaller molecules than the enzyme proteins. Some compounds, such as the dinucleotides NAD and NADP, are classified as coenzymes and are specific substrates in two-substrate reactions. Their effects on the rate of reaction follow the Michaelis-Menten pattern of dependence on substrate concentration. The structures of these two coenzymes are identical except for the presence of an additional phosphate group in NADP; nevertheless, individual dehydrogenases, for which these coenzymes are substrates, are predominantly or even absolutely specific for one or the other form.

Coenzymes such as NAD and NADP are bound only momentarily to the enzyme during the course of the reaction, as is the case for substrates in general. Therefore no reaction takes place unless the appropriate coenzyme is present in solution (e.g., by adding it to the reaction mixture in the assay of dehydrogenase activity). In contrast to these entirely soluble coenzymes, some coenzymes are more or less permanently bound to the enzyme molecules, where they form part of the active center and undergo cycles of chemical change during the reaction. In this case, the coenzyme is known as a prosthetic group. Prosthetic groups, such as activators with a structural role, do not usually have to be added to elicit full catalytic activity of the enzyme unless previous treatment has caused the prosthetic group to be lost from some enzyme molecules.

The active *holoenzyme* results from the combination of the inactive *apoenzyme* with the prosthetic group. An example of a prosthetic group is pyridoxal-5′-phosphate (P-5′-P), a component of AST and ALT. The P-5′-P prosthetic group undergoes a cycle of conversion of the pyridoxal moiety to pyridoxamine and back again during the transfer of an amino group from an amino acid to an oxoacid. However, both normal and pathologic serum samples contain appreciable quantities of apoaminotransferases, which are converted to active holoenzymes through a suitable period of incubation with P-5′-P.

A study of the formulas of coenzyme and prosthetic groups shows that many contain structures derived from vitamins (see Chapter 27). Thus the nicotinamide portion of NAD and NADP derives from the vitamin niacin, whereas the P-5′-P prosthetic group of the aminotransferases is a derivative of pyridoxine, vitamin B_6.

POINTS TO REMEMBER

- The catalytic properties of enzymes can be described by equations based on the formation of an ES complex.
- Enzyme reactions are affected by temperature, pH, enzyme concentration, substrate concentration, and the presence of any inhibitors or activators.

ANALYTICAL ENZYMOLOGY

Analytically, the clinical laboratorian is concerned with measuring the activity or mass of enzymes in serum or plasma that are predominantly intracellular, and that are physiologically present in the serum in low concentrations only. By measuring changes in the concentrations of these enzymes in disease, it is possible to infer the location and nature of pathologic changes in body tissues.

Measurement of Reaction Rates

Both *fixed-time* and *continuous-monitoring* methods are used to measure reaction rates. In the fixed-time method, the amount of product produced from substrate by the enzyme is measured after the reaction is stopped at the end of a fixed-time interval. In the continuous-monitoring method, the rate of product formation is monitored as the reaction is proceeding. These two methods have different advantages and limitations. To appreciate them, it is necessary to consider the way in which the rate of an enzymatic reaction varies with time.

The progress of conversion of the substrate into products in the presence of an enzyme is monitored by measuring the decreasing concentration of the substrate or the increasing concentration of the products. Measurement of product formation is preferable because determination of the increase in concentration of a substance above an initially zero or low concentration is analytically more reliable than measurement of a decline from an initially high concentration.

At the moment when the enzyme and the substrate are mixed, the rate of the reaction is zero. The rate then typically rises rapidly to a maximum value, which remains constant for a period of time (Fig. 14.8). During the period of constant reaction rate, the rate depends only on enzyme concentration and is completely independent of substrate concentration. The reaction is said to follow zero-order kinetics, because its rate is proportional to the zero power of the substrate concentration. Ultimately, however, as more substrate is consumed, the reaction rate declines and enters a phase of first-order dependence on substrate concentration. Other factors that contribute to the decline in reaction rate include accumulation of products that may be inhibitory, the growing importance of the reverse reaction, and even enzyme denaturation. Enzyme assays are

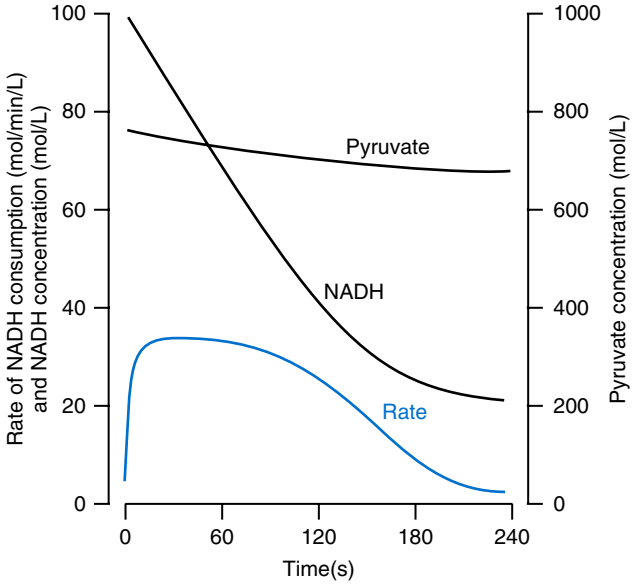


Fig. 14.8 Changes in substrate concentrations and rates of reaction during an assay of lactate dehydrogenase activity at 37°C in phosphate buffer, with pyruvate and nicotinamide adenine dinucleotide *(NADH)* as substrates. The reaction is followed by observing the fall in absorbance at 340 nm as NADH is oxidized to NAD^+. The rate of reaction rises rapidly to a maximum value, from which it declines only slightly until about half the NADH has been used up. During this phase of the reaction, the rate is essentially zero order with respect to substrate concentration. At the point at which the rate falls below about 90% of its maximum value, NADH concentration is approximately 10 × K_m. The K_m for NADH is on the order of 5 × 10^{-6} mol/L, whereas for pyruvate it is 9 × 10^{-5} mol/L. Thus an initial pyruvate concentration approximately 10 times that of NADH is used. (Concentrations are per liter of reaction mixture.) (From Moss, D. W. [1979]. Measurement of enzymes. In: D. J. Hearse & J. de Leiris [Eds.], *Enzymes in cardiology: diagnosis and research.* New York: John Wiley & Sons. Reprinted by permission of John Wiley & Sons.)

usually made under conditions that are initially saturating with respect to substrate concentration. The rate of reaction during the zero-order phase is determined by measuring the product formed during a fixed period of incubation in which the rate remains constant. This is illustrated in Fig. 14.9. Measurement of reaction rates at any portion of *curve A* gives results that are identical to the true *initial rate*. However, *curve B* deviates from linearity over its entire course, and rates fall off with time. From *curve C,* correct results are obtained only if the rate is measured along segment II. Incorrect results are obtained if the rate is measured during the lag phase (I) or during the phase of substrate depletion (III).

Careful selection of reaction conditions, such as the concentrations of substrates and cofactors, improves the reaction progress curves, eliminates lag phases, and prolongs the period of zero-order kinetics, so that fixed-time methods of analysis become feasible. Improvements in optical techniques, leading to more reliable and sensitive measurement of product formation, also have allowed the duration of incubation to be shortened compared with that of older assays. Nevertheless, an upper limit of activity exists in all fixed-time methods, above which progress curves will no longer be linear. In this case, the amount of change measured over the fixed-time interval no longer represents true zero-order rate conditions.

The upper limit of activity must be chosen so samples with activities below it are presumed, with a high degree of certainty, to give linear progress curves; alternatively, if the limit is set too low, many samples will have to be diluted and reanalyzed unnecessarily. Samples that are above the limit should ideally be assayed again after shortening the incubation period until a constant reaction rate is obtained. However, this is difficult or impossible in some automated methods, in

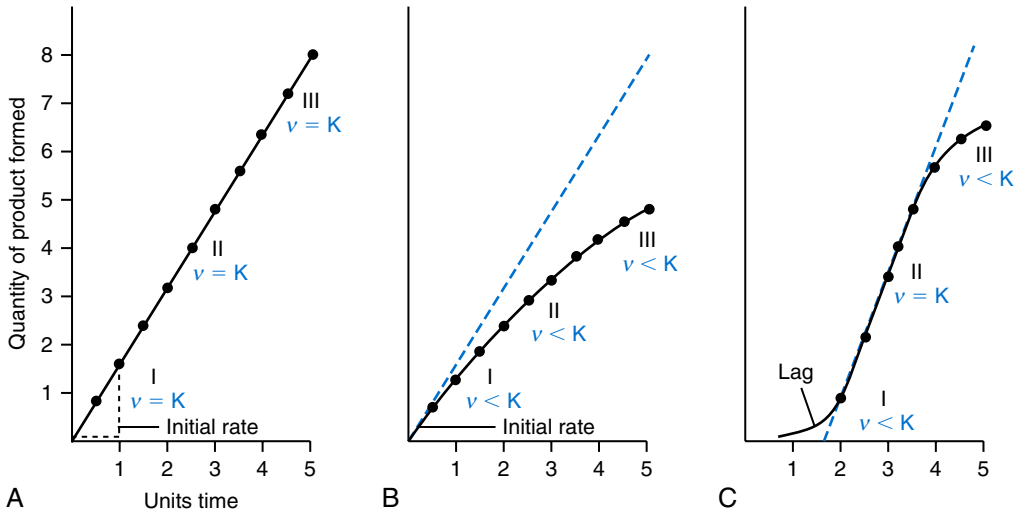


Fig. 14.9 Graphs showing change in enzyme reaction rate as a function of time. (A) The rate is constant during the entire run, and rates calculated as I, II, and III will be identical to the initial rate. (B) The rate falls off continuously; rates calculated at I, II, and III will be different and less than the true initial rate. (C) A measurement at II will be representative of the maximum rate, but at I (lag period) and III (substrate depletion), it will be less than at II.

which the duration of incubation is fixed by the configuration of the apparatus. It then becomes necessary to dilute the specimen to obtain valid results. However, some diluents can cause matrix-related alterations to the reaction kinetics, and may not always result in a proportional change in activity.

The initial rate of reaction theoretically increases without limit as enzyme concentration increases, as long as no other factor, such as substrate concentration, becomes limiting. In practice, the reaction rate becomes so rapid at high enzyme activities that it is impossible to measure the initial rate of reaction, even with continuous monitoring methods. Therefore an upper limit of activity that is measurable without modification of the assay procedure exists even in continuous monitoring methods; but this limit is usually much higher than those observed in fixed-time methods. Therefore fewer samples require special treatment when using continuous monitoring methods. Furthermore, continuous monitoring allows identification of the appropriate zero-order portion of the progress curve for each sample and identification of samples that require special treatment. Continuous-monitoring methods therefore provide a decisive advantage in enzyme assay and should be used whenever possible.

Measurement of Substrates

The amount of substrate transformed into products during an enzyme-catalyzed reaction can be measured with any appropriate analytical method, such as spectrophotometry, fluorometry, or chemiluminescence (see Chapter 9). *Self-indicating* reactions of this type are particularly valuable because they allow continuous monitoring. Important examples of self-indicating reactions are determination of dehydrogenase activity by monitoring the change in absorbance at 339 (340) nm of the coenzyme NADH or NADPH during oxidation or reduction, and measurement of ALP activity by the generation of the yellow *p*-nitrophenolate ion from the substrate *p*-nitrophenyl phosphate in alkaline solution. These indicator reactions are so versatile that dehydrogenase reactions are frequently coupled to the product of a primary reaction, to provide a measurable absorbance change when the primary product does not possess a strong specific absorptivity.

The introduction of prism or diffraction-grating spectrophotometers capable of isolating a narrow beam of monochromatic light in the ultraviolet or visible spectrum, and stable and sensitive photomultipliers as detectors, greatly improved the reproducibility of photometric measurements (see Chapter 9). Consequently, it is customary to make use of the known molar absorptivity of well-defined reaction products (such as NADH) when calculating changes in their concentrations based on measurements made with spectrophotometers.

Optimization and Standardization

To measure enzyme activity reliably, all factors that affect the reaction rate—other than the concentration of active enzyme—need to be optimized and rigidly controlled. When the reaction velocity is at or near its maximum under optimal conditions, a larger analytical signal is obtained, making more accurate and precise results possible than with the smaller signal obtained under suboptimal conditions. Much effort therefore has been devoted to determining optimal conditions for measuring the activities of enzymes of clinical importance.

Optimization

Optimization of reaction conditions for enzyme assays has traditionally been carried out by varying a single factor and studying its effect on the reaction rate, then repeating the experiment with a second factor and so on, until effects of all the variables have been tested. An optimal combination of variables is selected on the basis of these experiments, and the validity of the chosen conditions is verified. Not only is this approach labor intensive, it is also difficult to adapt in situations in which the effects of different variables are interdependent, as is frequently the case in enzyme analysis. This traditional empirical approach to optimization has been replaced by techniques of simplex cooptimization and response surface methodology.

Standardization

Current enzyme standardization efforts are focused on the development of a system that allows comparability of test results, independent of the measurement method. To achieve this, a reference measurement system based on the concepts of metrologic traceability and on the hierarchy of analytical methods has been developed. A reference procedure and certified reference materials form the basis of the metrological traceability chain (Fig. 14.10). As part of this hierarchy, reference procedures at 37°C for the most common enzymes have been developed, and manufacturers have used these as the basis of their commercial methods.

For a reference system to be capable of standardizing the results of different assays of a given enzyme activity, some conditions must be satisfied. First, the reference procedure used to assign the value of a reference material and the field method(s) to be calibrated must have similar, if not identical, specificities for the analyte enzyme, isoenzyme, or isoform under study. Second, the calibrator material must be "commutable" with natural human serum, so that the numerical relationship between enzyme concentration and measured enzyme activity is the same in the calibrators as it is in human serum.

Measurement of Enzyme Mass Concentration

Many immunoassays for human enzymes and isoenzymes that measure protein mass instead of catalytic activity have been described. To develop such assays, purified enzyme protein has to be prepared to (1) act as a calibrator, (2) be labeled, and (3) be used to raise the enzyme-specific antibody. These methods identify all molecules with the antigenic determinants necessary for recognition by the antibody, so that inactive enzyme molecules that are immunologically unaltered are measured along with active molecules. In the majority of cases, however, no degradation or changes in the active enzyme occur in blood so that the clinical equivalence of the

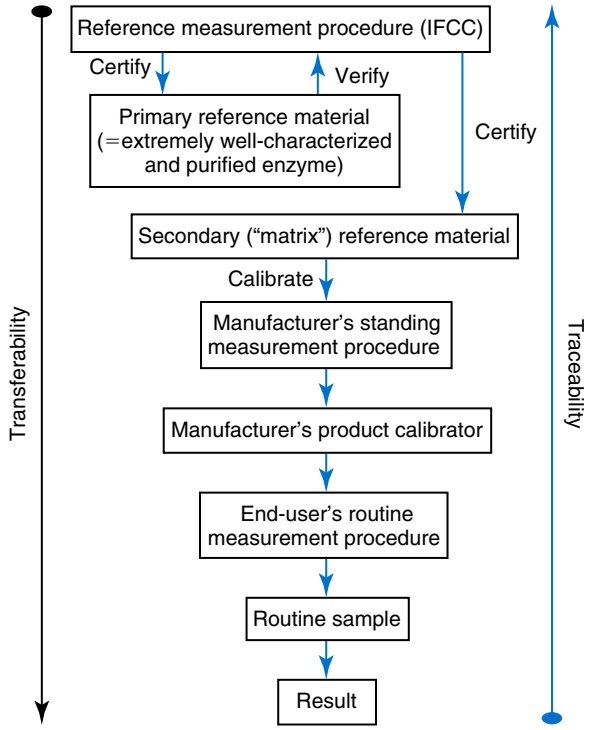


Fig. 14.10 The proposed reference system for enzyme measurement showing the traceability of the laboratory result to the reference measurement procedure. *IFCC,* The International Federation of Clinical Chemistry and Laboratory Medicine. (From Panteghini, M., Ceriotti, F., Schumann, G., & Siekmann, L. [2001]. Establishing a reference system in clinical enzymology. *Clinical Chemistry and Laboratory Medicine, 39,* 795–800. Reprinted with permission of Walter de Gruyter.)

different measurement approaches (i.e., estimation of catalytic activity and mass concentration) is obtained.

Immunoassays typically are not used for determination of total activities for the more important diagnostic enzymes because these assays generally cannot compete in terms of speed, precision, and cost with automated measurements of catalytic activity. Furthermore, several enzyme activities in serum are due to mixtures of immunologically distinct forms, so an assay using a single type of antibody usually determines only one of the enzyme forms. However, this disadvantage in the determination of total enzyme activity becomes a marked advantage in the measurement of specific isoenzymes and isoforms, and immunologic methods have been used routinely in the isoenzyme analysis of CK, amylase, and ALP.

Enzymes as Analytical Reagents

Enzymes are used as analytical reagents for measurement of several metabolites and substrates and in immunoassays to detect and quantify immunologic reactions.

Measurement of Metabolites

The use of enzymes as analytical reagents to measure metabolites frequently offers the advantage of great specificity for the substance being determined. This high specificity typically removes the need for preliminary separation or purification steps, so the analysis can be carried out directly on complex mixtures such as serum. Uricase (urate oxidase), urease, and glucose oxidase are examples of highly specific enzymes used in clinically important assays, such as the measurement of uric acid, urea, and glucose in biological fluids. However, high specificity cannot always be achieved in practice, and knowledge of the substrate specificities of reagent enzymes is essential to identify possible interferences with the assay, and to correct for them if possible. Coupled reactions are often necessary to construct an enzymatic assay for a specific compound. For example, in the enzymatic determination of glucose, hexokinase converts a number of sugars to their 6-phosphate esters. However, the indicator reaction used to monitor this change uses glucose-6-phosphate dehydrogenase, which is highly specific for glucose-6-phosphate, rendering the assay highly specific for glucose.

Equilibrium methods. Most assays used to determine the amount of a substance enzymatically are allowed to continue to completion, so that all substrate has been converted into a measurable product. These methods are called *end point* or, more correctly, *equilibrium* methods, because the product concentration no longer increases when equilibrium is reached. Reactions in which the equilibrium point corresponds virtually to complete conversion of the substrate to product are obviously preferable for this type of analysis. However, an unfavorable equilibrium often can be overcome by additional enzymatic or nonenzymatic reactions that convert or trap a product of the first reaction. Removing one of the products of the reaction keeps the system in disequilibrium, allowing for nearly total conversion of substrate into product. For example, when measuring lactate with LD, the pyruvate formed can be trapped by the addition of hydrazine, with which it forms an irreversible hydrazone, preventing the pyruvate from causing product inhibition of the LD reaction.

Theoretically, the time required to transform a fixed quantity, Q, of substrate into products is inversely proportional to the amount of enzyme, [E], present:

$$Q = k_1 \times [\mathrm{E}] \times t \qquad \textbf{(14.27)}$$

and

$$[\mathrm{E}] = \frac{Q}{k_1} \times \frac{1}{t} \qquad \textbf{(14.28)}$$

where k_1 is the rate constant and t is the elapsed time. Equilibrium methods may therefore require the use of appreciable amounts of enzyme for each sample, to avoid inconveniently long incubation periods. As the substrate concentration falls to low quantities toward the end of the reaction, the K_m of the enzyme becomes important in determining the reaction rate. Enzymes with high affinities for their substrates (low K_m values) are most suitable for equilibrium analysis. Equilibrium methods are largely insensitive to minor changes in reaction conditions, so it is not necessary to have exactly the same amount of enzyme in each reaction mixture or to maintain the pH or temperature absolutely constant, provided that the variations are not so great that the reaction is not completed within the fixed time allowed.

Kinetic methods. First-order or pseudo–first-order reactions are the most important reactions for the kinetic

determination of substrate concentration. For any first-order reaction, the substrate concentration [S] at a given time t after the start of the reaction is given by

$$[S] = [S_0] \times e^{-kt} \quad (14.29)$$

where $[S_0]$ is the initial substrate concentration, e is the base of the natural log, and k is the rate constant.

The change in substrate concentration $\Delta[S]$ over a fixed time interval, t_1 to t_2, is related to $[S_0]$ by the equation

$$[S_0] = \frac{-\Delta[S]}{e^{-kt_1} - e^{kt_2}} \quad (14.30)$$

As this equation indicates, the change in substrate concentration over a fixed time interval is directly proportional to its initial concentration. This is a general property of first-order reactions.

For an enzymatic reaction, **first-order kinetics** is followed when [S] is small compared with K_m. Thus

$$v = \frac{V_{max}}{K_m} \times [S] \quad (14.31)$$

or

$$v = k[S] \quad (14.32)$$

Therefore the first-order rate constant, k, is equal to $\frac{V_{max}}{K_m}$.

Methods in which some property related to substrate concentration (e.g., absorbance, fluorescence, chemiluminescence) is measured at two fixed times during the course of the reaction are known as *two-point* kinetic methods. Theoretically, they are most accurate for the enzymatic determination of substrates. However, these methods are technically more demanding than equilibrium methods and all factors that affect reaction rate, such as pH, temperature, and amount of enzyme, must be kept constant from one assay to the next, as must the timing of the two measurements. These conditions can be readily achieved in automated analyzers. A reference solution of the analyte (substrate) is used for calibration. To ensure first-order reaction conditions, the substrate concentration must be low compared with the K_m (i.e., in the order of less than $0.2 \times K_m$). Enzymes with high K_m values are therefore preferred for kinetic analysis to obtain a wider usable range of substrate concentrations.

Immunoassay

In enzyme immunoassay, enzyme-labeled antibodies or antigens are first allowed to react with a ligand; then an enzyme substrate is added. Enzymes such as ALP have been used as enzyme labels. A modification of this methodology is the enzyme-linked immunosorbent assay (ELISA), in which one of the reaction components is bound to a solid-phase surface. In this technique, an aliquot of sample is allowed to interact with the solid-phase antibody. After washing, a second antibody labeled with enzyme is added to form an Ab–Ag–Ab–enzyme complex. Excess free enzyme-labeled antibody is then washed away, and the substrate is added; the conversion of substrate is proportional to the quantity of antigen. In immunoassays, it is not the specificity of labeled enzymes that is important, but rather their sensitivity.

Analytical Applications of Immobilized Enzymes

For some types of enzymatic analyses, enzyme consumption is reduced by the use of immobilized enzymes that are reused for several analyses. Immobilized enzymes have been chemically bonded to adsorbents such as (1) microcrystalline cellulose, (2) diethylaminoethyl (DEAE) cellulose, (3) carboxymethyl cellulose, and (4) agarose. Diazo, triazine, and azide groups are used to join the enzyme protein to the insoluble matrix, forming particles in contact with the substrate solution or a surface in contact with substrate solution, such as a membrane or a coating on the inner surface of a vessel holding the substrate solution. Among enzymes available in such immobilized form are (1) urease, (2) hexokinase, (3) α-amylase, (4) glucose oxidase, (5) trypsin, and (6) leucine aminopeptidase. Immobilized enzymes are generally more stable than enzymes in solution; but some properties of the enzyme, such as its K_m or its pH optimum, may be altered.

Electrochemical techniques such as (1) potentiometry, (2) polarography, and (3) microcalorimetry have been adapted to exploit the benefits of immobilized enzymes. Enzymes incorporated into membranes form part of enzyme electrodes such as on blood gas analyzers. The surface of an ion-sensitive electrode is coated with a layer of porous gel in which an enzyme has been polymerized. When the electrode is immersed in a solution of the appropriate substrate, the action of the enzyme produces ions to which the electrode is sensitive.

Measurement of Isoenzymes and Isoforms

A number of analytical techniques have been used to measure specific isoenzymes or isoforms including (1) electrophoresis (see Chapter 11), (2) chromatography (see Chapter 12), (3) chemical inactivation, and (4) differences in catalytic properties, but the most widely used methods are now based on immunochemical principles (see Chapter 15).

Electrophoresis

Various forms of electrophoresis have been used to separate isoenzymes. The methods are time-consuming, difficult to quantify, and relatively expensive and thus tend to be restricted to larger, reference laboratories.

Chromatography

Ion-exchange chromatography makes use of differences in net molecular charge at a given pH to separate isoenzymes. A typical ion-exchange material is DEAE cellulose, in which ionizable DEAE groups are attached to an inert cellulose matrix. Ion-exchange chromatography, in general, lacks the resolving power of electrophoresis, but relatively large quantities of proteins can be separated with good recovery of enzymatic activity; so the method is of great value in enzyme purification.

Other forms of chromatography that have been applied to fractionation of isoenzyme mixtures include high-performance liquid chromatography and affinity chromatography. The latter makes use of differences between isoenzymes in their affinities for a specific ligand that is attached to an inert insoluble support used as the stationary phase in a chromatography column or in a batch technique.

Immunochemical Assays

Immunochemical methods of isoenzyme analysis are particularly applicable to isoenzymes derived from multiple genes because these are usually most clearly antigenically distinct. However, the greater discriminating power of monoclonal antibodies has potentially brought all multiple forms of an enzyme within the scope of immunochemical analysis. Some of these methods detect catalytic activity of the isoenzymes. For example, residual activity may be measured after reaction with antiserum. Radioimmunoassays in which isoenzyme labeled with a radioactive tracer competes with unlabeled isoenzyme for antibody-binding sites have been applied in the past to isoenzyme measurement. However, with the development of automated immunoassay systems, the preferred method for measuring isoenzymes in the clinical laboratory is a solid-phase sandwich ELISA assay because of its convenience and relative low cost. These methods do not depend on the catalytic activity of the isoenzyme being determined.

The choice and application of various methods of isoenzyme analysis in clinical enzymology in relation to specific isoenzyme systems are discussed in Chapter 19.

POINTS TO REMEMBER

- Because of their specificity, enzymes can be used as analytical reagents to measure compounds such as metabolites in complex mediums.
- The measurement of isoenzymes increases the diagnostic specificity and sensitivity of enzyme assays in body fluid samples.

REVIEW QUESTIONS

1. The enzyme 2.2.8.11 is a(n):
   a. oxidoreductase.
   b. transferase.
   c. hydrolase.
   d. lyase.
   e. isomerase.
   f. ligase.
2. A disulfide bond in a protein is an example of:
   a. primary structure.
   b. secondary structure.
   c. tertiary structure.
   d. quaternary structure.
3. The active site of an enzyme is:
   a. small compared to the entire enzyme.
   b. usually in the center of the enzyme.
   c. not specific for inhibitors.
   d. typically only involves a few continuous bases along the primary structure.
4. Which of the following can affect the rate of an enzyme-catalyzed reaction?
   a. Enzyme concentration
   b. Substrate concentration
   c. pH
   d. Reaction temperature
   e. All of the above
5. K_m is:
   a. the enzyme concentration that results in 50% of the maximal reaction rate.
   b. a property of an enzyme and substrate under certain conditions.
   c. a property of an enzyme irrespective of the substrate.
   d. a property of a substrate irrespective of the enzyme.
6. Which statement is correct?
   a. Either the enzyme or the coenzyme, but not both, are consumed during an enzymatic reaction.
   b. An enzyme without a prosthetic group attached is referred to as a coenzyme.
   c. Coenzymes may contain vitamins.
   d. All of the above statements are correct.
7. An enzyme assay determined by a single measurement taken at the termination of the enzymatic reaction is referred to as a:
   a. fixed-time assay.
   b. terminator assay.
   c. kinetic assay.
   d. continuous-monitoring assay.
8. A reactant in a catalysis reaction that binds to the enzyme's active site is the:
   a. enzyme.
   b. product.
   c. substrate.
   d. coenzyme.
9. An international unit (U) of enzyme activity is the amount of enzyme that will catalyze the reaction of:
   a. one micromole of substrate per minute.
   b. one mole of substrate per second.
   c. one micromole of enzyme per hour.
   d. one millimole of substrate per minute.
10. In what kind of inhibition is the inhibitor usually a structural analog of the substrate?
    a. Competitive
    b. Noncompetitive
    c. Uncompetitive
    d. Racemic mirroring
11. In an analytical enzymatic measurement of substrate, an enzyme reaction is accompanied by a change in absorbance of one component of the assay system. If this change is continuously monitored while it is occurring, the reaction is referred to as a(n):
    a. fixed-time reaction.
    b. end point assay.
    c. rate-limiting reaction.
    d. self-indicating reaction.

SUGGESTED READINGS

Canalias, F., Camprubí, S., Sánchez, M., & Gella, F. J. (2006). Metrological traceability of values for catalytic concentration of enzymes assigned to a calibration material. *Clinical Chemistry and Laboratory Medicine*, *44*, 333–339.

Copeland, R. A. (2000). *Enzymes: a practical introduction to structure, mechanism, and data analysis* (2nd ed.). New York: Wiley-VCH.

Cook, P. F., & Cleland, W. W. (2007). *Enzyme kinetics and mechanism*. Burlington: Taylor & Francis Inc.

Harris, T. K., & Keshwani, M. M. (2009). Measurement of enzyme activity. *Methods in Enzymology*, *463*, 57–71.

Infusino, I., Bonora, R., & Panteghini, M. (2007). Traceability in clinical enzymology. *Clinical Biochemistry Reviews*, *28*, 155–161.

Infusino, I., Schumann, G., Ceriotti, F., & Pnateghini, M. (2010). Standardization in clinical enzymology: A challenge for the theory of metrological traceability. *Clinical Chemistry and Laboratory Medicine*, *48*, 301–307.

Moss, D. W. (1979). *Isoenzyme analysis*. London: The Chemical Society.

Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). (1992). *Enzyme nomenclature: Recommendations*. San Diego: Academic Press. p. 862. 1992 with Supplement 1(1993), Supplement 2 (1994), Supplement 3 (1995), Supplement 4 (1997), Supplement 5 (in Eur J Biochem 1994;223:1–5; Eur J Biochem 1995;232:1–6; Eur J Biochem 1996;237:1–5; Eur J Biochem 1997;250;1–6, and Eur J Biochem 1999;264;610–650; respectively, and subsequent Supplements 6–19 at http://www.chem.qmul.ac.uk/iupac/bibliog/jcbn.html.

Okotore, R. O. (2015). *Essentials of enzymology*. Bloomington: Xlibris Corporation.

Panteghini, M., Ceriotti, F., Schumann, G., & Siekmann, L. (2001). Establishing a reference system in clinical enzymology. *Clinical Chemistry and Laboratory Medicine*, *39*, 795–800.

Price, N., & Stevens, L. (2000). *Fundamentals of enzymology: Cell and molecular biology of catalytic proteins* (3rd ed.). Oxford: Oxford University Press.

Rautela, G. S., Snee, R. D., & Miller, W. K. (1979). Response-surface co-optimization of reaction conditions in clinical chemical methods. *Clinical Chemistry*, *25*, 1954–1964.

Siekmann, L., Bonora, R., Burtis, C. A., Ceriotti, F., Clerc-Renaud, P., Férard, G., et al. (2002). IFCC primary reference procedures for the measurement of catalytic activity concentrations of enzymes at 37°C. I. the concept of reference procedures for the measurement of catalytic activity concentrations of enzymes. *Clinical Chemistry and Laboratory Medicine*, *40*, 631–634.

Schumann, G., Canalias, F., Joergensen, P. J., Kang, D., Lessinger, J. M., & Klauke, R. (2010). IFCC reference procedures for measurement of the catalytic concentrations of enzymes: corrigendum, notes and useful advice. International Federation of Clinical Chemistry and Laboratory Medicine (IFCC)—IFCC Scientific Division. *Clinical Chemistry and Laboratory Medicine*, *48*, 615–621.

15

Immunochemical Techniques

Larry J. Kricka, Jason Y. Park

OBJECTIVES

1. Define the following:
   - Affinity
   - Antibody
   - Antigen
   - Avidity
   - Hapten
   - Immunoassay
   - Immunogen
   - Monoclonal antibody
   - Polyclonal antiserum
   - Sandwich immunoassay
2. Diagram, label, and state the function of the components of an immunoglobulin G antibody molecule.
3. List three binding forces that act to produce antigen/antibody binding.
4. Explain how the following factors affect antigen/antibody binding:
   - Addition of a linear polymer
   - Ion species
   - Precipitin reaction
5. For each of the following qualitative immunochemical techniques, state the principle and clinical use:
   - Immunofixation
   - Western blotting
6. For each of the following quantitative immunochemical techniques, state the principle and clinical use:
   - Nephelometry
   - Turbidimetry
7. List five types of nonisotopic labels used in a labeled immunochemical assay, and provide one example of each type of label.
8. Compare competitive with noncompetitive immunoassays, including procedure, assay components, and uses in the clinical laboratory. Do the same for heterogeneous and homogeneous immunoassays.
9. For each of the following immunochemical techniques, state the principle, the components of each assay, the type of assay, and the clinical uses:
   - Cloned enzyme donor immunoassay
   - Enzyme-linked immunosorbent assay
   - Enzyme-multiplied immunoassay
   - Fluoroimmunoassay
   - Immunocytochemistry
10. Describe three factors that can adversely affect immunoassay results.

KEY WORDS AND DEFINITIONS

Affinity Energy of interaction of a single antibody-combining site and its corresponding epitope on the antigen.

Antibody Immunoglobulin (Ig) class of molecule (e.g., IgA, IgG, IgM) that binds specifically to an antigen or hapten.

Antigen Any material capable of reacting with an antibody without necessarily being capable of inducing antibody formation.

Avidity Overall strength of binding of antibody and antigen; includes the sum of the binding affinities of all individual combining sites on the antibody.

Competitive immunoassay An immunoassay in which all reactants are simultaneously or sequentially mixed together and unlabeled antigen competes with labeled antigen for binding sites on the antibody.

Hapten A chemically defined determinant that, when conjugated to an immunogenic carrier, stimulates the synthesis of antibody specific for the hapten.

Heterogeneous immunoassay An immunochemical reaction in which it is assumed that the formation of the antigen/antibody complex occurs more quickly than the breakdown of the complex into antigen and antibody; in this assay, the antigen is labeled and separation of the free from the bound labeled antigen is required.

Homogeneous immunoassay An immunochemical reaction in which the activity of the label attached to the antigen is modulated directly by antibody binding; this assay does not require a separation.

Hook effect A phenomenon occurring with certain immunoassays due to very high concentrations of a particular analyte; it results in a false-negative result. The hook effect mostly affects one-step immunometric assays.

Immunoassay An assay based on the reaction of an antigen with an antibody specific for the antigen.

Immunogen A substance capable of inducing an immune response.

Label Any substance with a measurable property attached to an antigen, antibody, or binding substance.

Noncompetitive immunoassay An immunoassay in which a capture antibody is bound to a surface with subsequent antigen binding followed by the addition of a second labeled antibody that reacts with the initial antigen/antibody complex.

Western blotting Membrane-based assay in which proteins are separated by electrophoresis, which is followed by transfer to a membrane and probing with a labeled antibody.

BASIC CONCEPTS AND DEFINITIONS

An **antigen** is any material capable of reacting with an antibody. With **immunoassays**, the antigen is the analyte of clinical interest that is being measured. **Antibodies** are immunoglobulins that bind specifically to a wide array of natural and synthetic antigens, such as (1) proteins, (2) carbohydrates, (3) nucleic acids, (4) lipids, and (5) other molecules. Analytically, immunoglobulin (Ig)G is the most prevalent immunochemical reagent in use. It is a glycoprotein that has a molecular weight (MW) of 158,000 Da and is composed of two duplex chains, with each set composed of a heavy (γ) and a light (λ or κ) chain joined by disulfide bonds (Fig. 15.1). Interchain disulfide bonds hold the duplex chains together and create a symmetrical molecule. The variable amino acid sequence at the amino terminal end of each chain determines the antigenic specificity of the particular antibody. Each unique amino acid sequence is a product of a single plasma cell line or clone, and each plasma cell line produces antibodies with single specificities. A complex antigen elicits a multiplicity of antibodies with different specificities derived from different cell lines. An antibody developed in this manner is termed *polyclonal* and exhibits diverse specificities in its reactivity with the immunogen. Each unique region of the molecular antigen that binds a complementary antibody is termed an *epitope* (antigenic determinant).

An **immunogen** is a protein or a substance coupled to a carrier, usually a protein. Introduction of an immunogen into a foreign host induces the formation of an antibody. A **hapten** is a chemically defined determinant that by itself will not

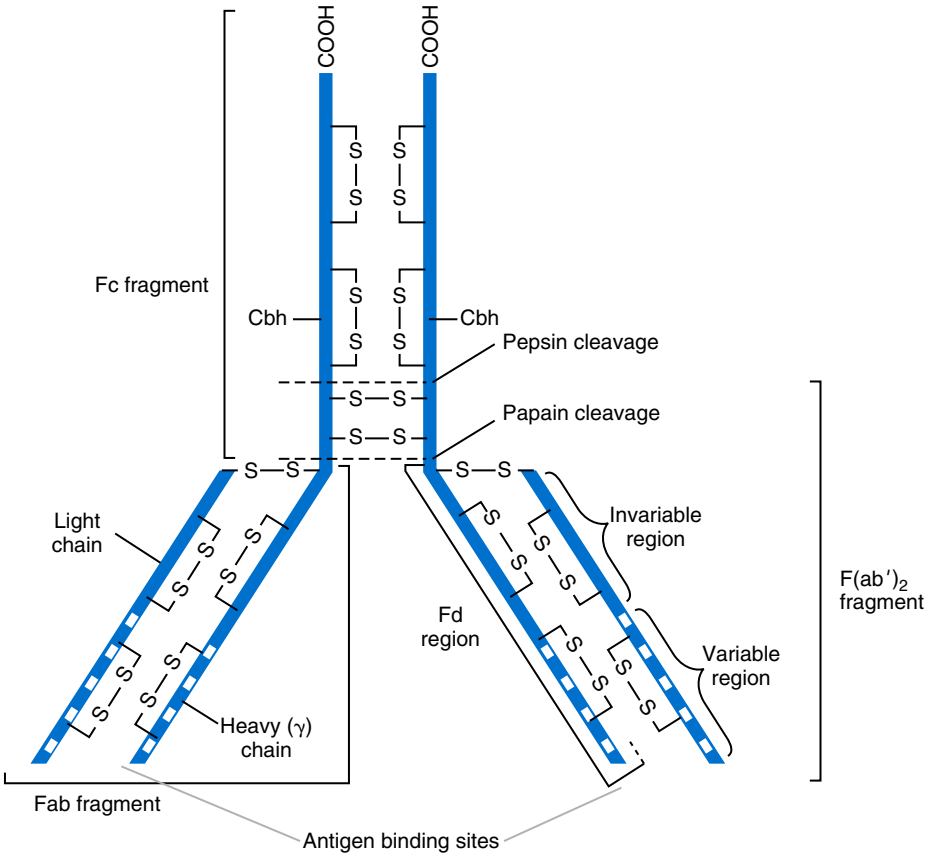


Fig. 15.1 Schematic diagram of immunoglobulin G (IgG) antibody molecule showing carbohydrate *(Cbh)*, disulfide bonds *(SS)*, and major fragments produced by proteolytic enzyme treatment ($F[ab']_2$, Fc, Fab, Fd).

stimulate an immune response. However, when conjugated to an immunogenic carrier, the conjugated molecule stimulates the synthesis of antibody specific for the hapten. Some general properties required for immunogenicity include the following:

1. Areas of structural stability within the molecule
2. Randomness of structure
3. Minimum MW of 4000 to 5000 Da
4. Ability to be metabolized (a necessary but insufficient criterion for some classes of antigens)
5. Accessibility of a particular immunogenic configuration to the antibody-forming mechanism
6. Structurally foreign quality

The strength or energy of interaction between the antibody and the antigen is described in two terms. **Affinity** refers to the thermodynamic quantity defining the energy of interaction of a single antibody-combining site and its corresponding epitope on the antigen. **Avidity** refers to the overall strength of the binding of antibody and antigen and includes the sum of the binding affinities of all individual combining sites on the antibody. Thus affinity is a property of the substance bound (antigen), and avidity is a property of the binder (antibody).

Polyclonal antiserum is produced in a normal animal host in response to immunogen administration. In contrast, a *monoclonal antibody* is the product of a single clone or plasma cell line rather than a heterogeneous mixture of antibodies produced by many cell clones in response to immunization. Monoclonal antibodies now are used widely as reagents in immunoassay techniques.

Monoclonal antibodies provide an analytical advantage in that two different antibody specificities can be used in a single assay. For example, a solid phase antibody specific for a unique epitope and another labeled antibody specific for a different epitope are reacted with antigen in a single incubation step. This combination eliminates (1) the two-step sequential addition of antigen and labeled antibody to the solid phase, (2) one incubation step, and (3) one washing step, which would be necessary when polyclonal antibodies binding to both sites are used.

ANTIGEN-ANTIBODY BINDING

In this section, several of the factors that affect the binding of antigens and antibodies are discussed.

Binding Forces

Several forces act cooperatively to produce antigen-antibody binding. The three major contributing forces are (1) electrostatic van der Waals–London dipole-dipole interactions, (2) hydrophobic interactions, and (3) ionic coulombic bonding (primarily between carboxylate and protonated amino groups on the antigen and the antibody).

Reaction Mechanism

The binding of antigen to antibody is not static but is an equilibrium reaction that proceeds in three phases. The initial reaction (phase 1) of a multivalent antigen (Ag_n) and a bivalent antibody (Ab) occurs very rapidly in comparison with subsequent growth of the complexes (phase 2) and is depicted by the following equation:

$$Ag_n + Ab \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} Ag_nAb \underset{k_{-2}}{\overset{k_1}{\rightleftharpoons}} Ag_aAb_b \qquad \textbf{(15.1)}$$

where $k_1 >>> k_2$, n is the number of epitopes per molecule, and a and b are the numbers of antigen and antibody molecules per complex. Phase 3 of the reaction involves the precipitation of the complex after a critical size is reached. The speed of these reactions depends on electrolyte concentration, pH, and temperature, as well as on antigen and antibody types and the binding affinity of the antibody.

Precipitin Reaction

The curve shown in Fig. 15.2 is a schematic diagram of the classic precipitin curve. Although the concentration of total antibody is constant, the concentration of free antibody, $[Ab]_f$, and of free antigen, $[Ag]_f$, varies for any given Ag/Ab ratio. A low Ag/Ab ratio exists in Fig. 15.2A (zone of antibody excess). Under these conditions, excess Ab is free in solution, but all Ag is bound and immune complexes are small. As total antigen increases, the size of the immune complexes increases up to equivalence (see Fig. 15.2B), and little or no $[Ab]_f$ or $[Ag]_f$ exists. This is the zone of equivalence, and it is the optimal combining ratio for cross-linking. As Ag/Ab further increases (see Fig. 15.2C), the immune complex size decreases and $[Ag]_f$ increases (zone of antigen excess).

Chemical Factors

Chemical factors that influence antibody-antigen binding include ionic species, ionic strength, and polymeric molecules.

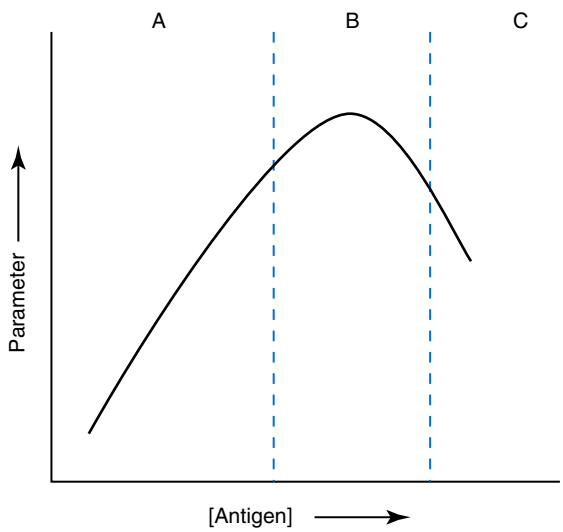


Fig. 15.2 Schematic diagram of precipitin curve illustrating different antigen concentration zones. *A,* Antibody excess. *B,* Equivalence. *C,* Antigen excess. The parameter measured may be the quantity of protein precipitated, light scattering, or another measurable parameter. Antibody concentration is held constant in this example.

Ion Species and Ionic Strength Effects

Cationic salts inhibit antibody binding to a cationic hapten. The order of inhibition by various cations corresponds to their decreasing ionic radius and the increasing radius of hydration (e.g., $Cs^+ > K^+ > Li^+$). For anionic haptens and anionic salts, the order of inhibition of binding again follows the order of decreasing ionic radius and increasing radius of hydration (e.g., SCN^- [thiocyanate] $> I^- > F^-$).

Polymer Effect

The addition of a linear polymer to a mixture of antigen and antibody causes a significant increase in the rate of immune complex growth and enhances the precipitation of immune complex, especially with low-avidity antibody. This is due to the increased probability of contact between the antigen and antibody caused by the presence of large polymer molecules in solution (steric exclusion). The most desirable characteristics of the polymer are high (1) MW, (2) degree of linearity (minimum branching), and (3) aqueous solubility. Polyethylene glycol 6000 has these characteristics and is particularly useful in immunochemical methods at concentrations of 3 to 5 g/dL.

QUALITATIVE METHODS

Immunochemical techniques used for qualitative purposes include (1) immunofixation (IF) and (2) Western blotting.

IF has gained widespread acceptance as an immunochemical method used to identify proteins. With this technique, electrophoresis is first performed in an agarose gel to separate the proteins in the mixture. Subsequently, antiserum spread directly on the gel causes the protein(s) of interest to precipitate. The immune precipitate is trapped within the gel matrix, and all other nonprecipitated proteins are then removed by washing of the gel. The gel then is stained for identification of the proteins. The utility of IF, which now is used widely for the evaluation of myeloma proteins, is illustrated in Fig. 15.3.

Blotting

The previously discussed techniques use direct examination of immunoprecipitation of the protein(s) in the gel. However, certain media, such as polyacrylamide, do not lend themselves to direct immunoprecipitation, nor does sufficient antigen concentration always exist to produce an immunoprecipitate that is retained in the gel during subsequent processing. Under these circumstances, the technique of **Western blotting** is used. This technique involves an electrophoresis step, followed by transfer of the separated proteins onto an overlying strip of nitrocellulose or nylon membrane by a process called *electroblotting*. Once the proteins are fixed to the membrane, they are detected with antibody probes labeled with molecules such as enzymes. With the use of such probes, the limits of detection are 10 to 100 times lower than the values obtained through direct immunoprecipitation and staining of proteins.

An example of blotting analysis for human immunodeficiency virus type 1 (HIV-1) antibodies is shown in Fig. 15.4. When applied to antigen assays, concentrations of antigen as low as 500 ng/mL or 2.5 ng per band in the gel have been detected. The detection limit of the technique is lowered even farther to approximately 100 pg by chemiluminescent detection of the enzyme-labeled antibody.

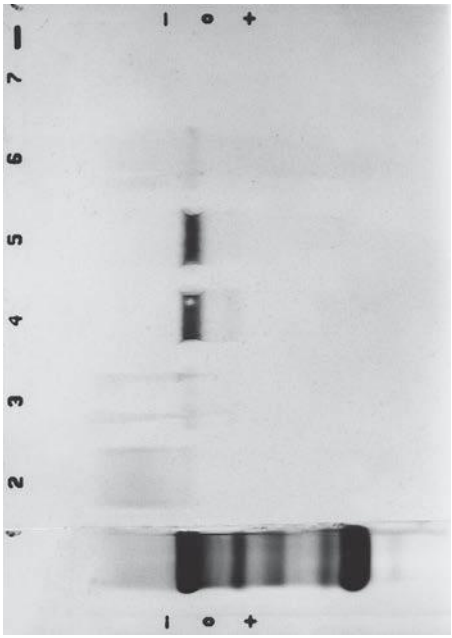


Fig. 15.3 Immunofixation of a serum containing an immunoglobulin (Ig)M kappa paraprotein. *Lane 1,* Serum electrophoresis stained for protein; *lane 2,* anti-IgG, Fc piece specific; *lane 3,* anti-IgA, α-chain specific; *lane 4,* anti-IgM, α-chain specific; *lane 5,* anti-κ light chain; *lane 6,* anti-λ light chain. (Courtesy Katherine Bayer, Philadelphia, PA.)

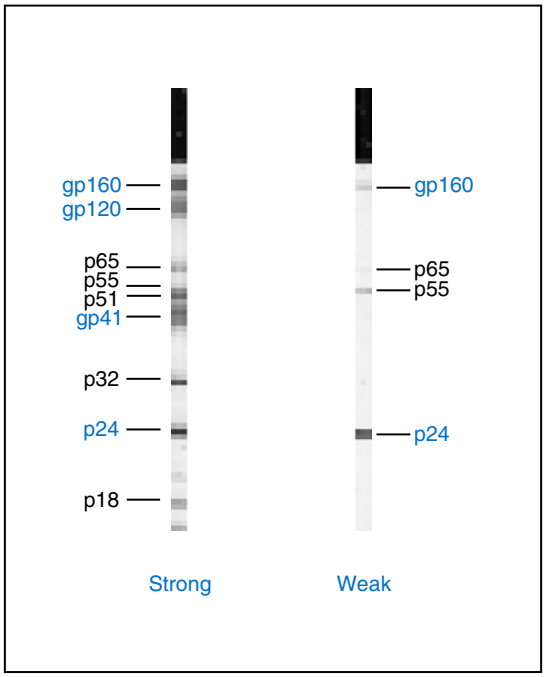


Fig. 15.4 Blot analysis of serum samples strongly positive and weakly positive for human immunodeficiency virus-1 antibody. Core proteins (GAGs, group-specific antigens) p18, p24, and p55; polymerase p32, p51, and p65; and envelope proteins gp41, gp120, and gp160. (Courtesy Bio-Rad Laboratories Diagnostics Group, Hercules, CA.)

A simpler technique that bypasses the electrophoretic separation step is known as *dot blotting*. A protein sample to be analyzed is applied to a membrane surface as a small "dot" and is dried. The membrane then is exposed to a labeled antibody specific for the test antigen contained in the dotted mixture. After the membrane is washed, bound-labeled antibody is detected with a photometric or chemiluminescent detection system.

QUANTITATIVE METHODS

Immunochemical techniques have been used to develop quantitative methods and include turbidimetric and nephelometric assays and labeled immunochemical assays.

Turbidimetric and Nephelometric Assays

Turbidimetry and nephelometry are convenient techniques, based on turbidity and light scattering, respectively, that are used to measure the rate of formation of immune complexes in vitro. Instrumental principles for these methods are described in Chapter 9. Studies have shown that the reaction between antigen and antibody begins within milliseconds and continues for hours. The performance of both types of assays has been improved significantly by increases in the reaction rate attained by the addition of water-soluble linear polymers.

Both turbidimetric and nephelometric immunochemical methods using rate and pseudoequilibrium protocols have been described for (1) proteins, (2) antigens, and (3) haptens. In rate assays, measurements are made early in the reaction because the largest change (dI_s/dt) in intensity of scattered light (I_s) with respect to time is attained during this time interval. For pseudoequilibrium assays, waiting 30 to 60 minutes is necessary so that the dI_s/dt is small relative to the time required to make the necessary measurements. Such assays are termed *pseudoequilibrium* rather than *equilibrium* because true equilibrium is not reached within the time allowed for these assays.

Nephelometric methods are generally more sensitive than turbidimetric assays and have a lower limit of detection of approximately 1 to 10 mg/L for a serum protein. Lower limits of detection are attained in fluids such as cerebrospinal fluid and urine because of their lower lipid and protein concentrations, which result in a higher signal-to-noise ratio. In addition, for low-molecular-weight proteins such as myoglobin (MW 17,800 Da), limits of detection have been lowered through a latex-enhanced procedure based on antibody-coated latex beads.

Nephelometric and turbidimetric assays have also been applied to the measurement of drugs (haptens). An example of this type of assay is the particle-enhanced turbidimetric inhibition immunoassay (PETINIA) that measures decreasing agglutination in the presence of increasing concentrations of analyte (e.g., digoxin). The reagents comprise an antibody to the analyte of interest and a reagent containing the analyte of interest linked to a latex bead. In the absence of analyte in a specimen, reagent antibody binds to the reagent containing the analyte linked to the latex bead, resulting in increased turbidity. When a specimen with high analyte concentrations is added to the reagent mixture, the reagent antibody is bound by the analyte in the specimen and not to the reagent containing the analyte linked to the latex bead. Thus the presence of specimen analyte results in less turbidity.

Surface Plasmon Resonance–based Immunoassay

Surface plasmon resonance (SPR) is a label-free immunoassay based on an optical phenomenon that enables detection of unlabeled reactants in real time based on changes in the index of refraction at the surface where the binding interaction occurs. The assay is conducted on an electrically conducting gold-coated glass slide mounted on a prism. The binding agent is immobilized on the gold surface over which reactant is flowed. Polarized light is directed underneath the glass slide and interacts with the gold surface to produce electron charge density waves called plasmons at the sample and gold surface interface. This results in a reduction in the intensity of the reflected light. Slight changes in the refractive index at the interface lead to a change in the signal, thus facilitating real-time detection of surface molecular interactions (e.g., association and dissociation of biomolecules with the binder immobilized on the gold surface). SPR assays are used to characterize binding reactions (e.g., antibody affinity, kinetics, cross-reactivity).

Labeled Immunochemical Assays

The previously discussed methods rely on examination of immune complex formation as an index of antigen-antibody reaction. As demonstrated in Eq. 15.1, the overall reaction occurs in sequential phases, and only the final phase consists of formation of the immune complex. However, initial binding of antibody and antigen has been used with antigens and antibodies that have **labels** to develop many sensitive and specific immunochemical assays. The reaction describing this initial binding and the kinetic constant for the overall reaction are shown in Eqs. 15.2a and 15.2b, respectively.

$$\mathrm{Ab} + \mathrm{Ab} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{AbAg} \qquad \textbf{(15.2a)}$$

$$K = \frac{[\mathrm{AbAg}]}{[\mathrm{Ab}]\,[\mathrm{Ag}]} \qquad \textbf{(15.2b)}$$

where k_1 = rate constant for the forward reaction; k_{-1} = rate constant for the reverse reaction; and K = equilibrium constant for the overall reaction.

As predicted from the law of mass action, the concentrations of Ab, Ag, and Ab:Ag are dependent on the magnitude of k_1 and k_{-1}. For polyclonal antiserum, the average avidity of the antibody populations determines K, and the magnitude of k_1 in comparison with k_{-1} determines the ultimate limit of detection attainable with a given antibody population.

Types of Labels

In the decade following the pioneering developments of Yalow and Berson in 1960, all immunoassays used radioactive **labels** in competitive assays. Since the introduction of enzyme

immunoassays (EIAs) in the 1970s, sophisticated assays with various nonisotopic labels (Table 15.1) have been developed.

Methodologic Principles

As shown in Fig. 15.5, the two major types of reaction formats used in immunochemical assays are termed *competitive* (limited reagent assays) and *noncompetitive* (excess reagent, two-site, or sandwich assays).

POINTS TO REMEMBER

- Competitive immunoassays are primarily for small molecules.
- Sandwich immunoassays can be designed for both small and large molecules.
- The sandwich type of immunoassay is more sensitive than the competitive immunoassay.
- Circulating human anti-mouse antibodies can interfere in immunoassays.

TABLE 15.1 Labels Used for Nonisotopic Immunoassay

Chemiluminescent	Acridinium ester, sulfonyl acridinium ester, isoluminol
Enzyme	Alkaline phosphatase, marine bacterial luciferase, β-galactosidase, firefly luciferase, glucose oxidase, glucose-6-phosphate dehydrogenase, horseradish peroxidase, lysozyme, malate dehydrogenase, microperoxidase, urease, xanthine oxidase
Fluorophore	Europium chelate, fluorescein, phycoerythrin, terbium chelate
Metal	Gold sol, selenium sol, silver sol
Particle	Bacteriophage, erythrocyte, latex bead, dyed latex bead, liposome, quantum dot, magnetic particle
Phosphor	Upconverting lanthanide-containing nanoparticle
Polynucleotide	DNA

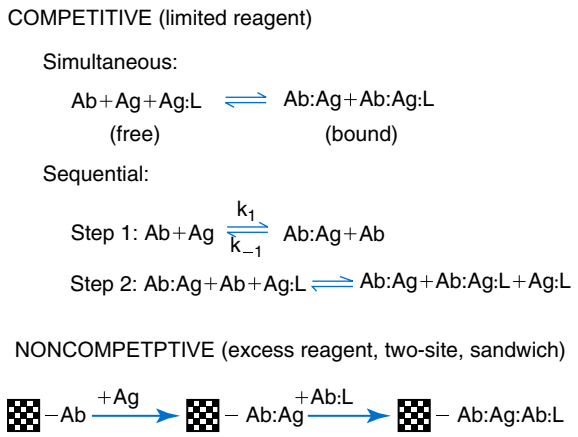


Fig. 15.5 Immunoassay designs. *Ab,* Antibody; *Ag,* antigen; k_1, forward rate constant; k_{-1}, reverse rate constant; *L,* label.

Competitive immunoassays. In a competitive immunochemical assay, all reactants are simultaneously or sequentially mixed together. In the simultaneous approach, labeled antigen (Ag:L) and unlabeled antigen (Ag) compete to bind with the antibody. In such a system, the avidity of the antibody for labeled and unlabeled antigen must be the same. Under these conditions, the probability of the antibody binding the labeled antigen is inversely proportional to the concentration of unlabeled antigen; hence bound label is inversely proportional to the concentration of unlabeled antigen.

In a sequential competitive assay, unlabeled antigen is mixed with excess antibody and binding allowed to reach equilibrium (see Fig. 15.5, Step 1). Labeled antigen is then added sequentially (see Fig. 15.5, Step 2) and allowed to equilibrate. After separation, the bound label is measured and is used to calculate the unlabeled antigen concentration. With this two-step method, a larger fraction of unlabeled antigen is bound by the antibody than in the simultaneous assay, especially at low antigen concentrations. Consequently, a twofold to fourfold lower detection limit is seen in a sequential immunoassay, compared with that in a simultaneous assay, provided $k_1 >> k_{-1}$. This improvement in detection limit results from an increase in AgAb binding (and thus a decrease in Ag:L binding), which is favored by the sequential addition of Ag and Ag:L. If $k_1 \geq k_{-1}$, dissociation of AgAb becomes more probable, resulting in increased competition between Ag:L and Ag. A typical immunochemical binding curve is shown in Fig. 15.6.

Noncompetitive immunoassays. In a typical noncompetitive assay, the "capture" antibody is first passively adsorbed or covalently bound to the surface of a solid phase. Next, the antigen from the sample is allowed to react and is captured by the solid phase antibody. Other proteins then are washed away, and a labeled antibody (conjugate) is added that reacts

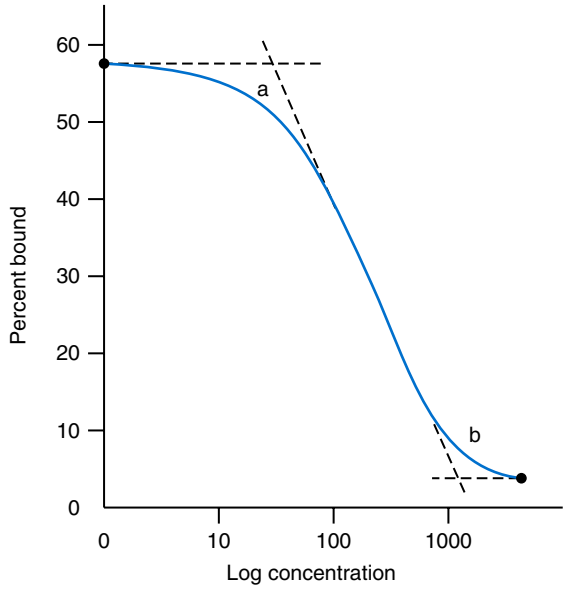


Fig. 15.6 Schematic diagram of the dose-response curve for a typical immunoassay. The analytically useful portion of the curve is bracketed by points *a* and *b*.

with the bound antigen through a second and distinct epitope. After additional washing to remove the excess unbound labeled antibody, the bound label is measured, and its concentration or activity is directly proportional to the concentration of antigen.

In noncompetitive assays, either polyclonal or monoclonal antibodies are used as capture and labeled antibodies. If monoclonal antibodies with specificity for distinct epitopes are used, simultaneous incubation of the sample and the conjugate with the capture antibody is possible, thus simplifying the assay protocol.

Noncompetitive immunoassays are performed in simultaneous or sequential mode. In the simultaneous mode, a high concentration of analyte saturates both capture and labeled antibodies. Under these conditions, the analyte is present in such high concentrations that it reacts simultaneously with the capture and labeled antibodies, reducing the number of complexes formed and producing a falsely low result. Thus the calibration curve of the assay exhibits a **hook effect,** in which the assay response drops off at high analyte concentrations in assays for analytes for which the normal pathologic concentration range is very wide. For example, assays for chorionic gonadotropin (CG) and alpha fetoprotein (AFP) are particularly prone to this problem. Dilutions of a sample usually are reanalyzed to check for this type of analytical interference. In practice, the hook effect is eliminated if a sequential assay format is adopted and the concentrations of capture and labeled antibody are sufficiently high to cover analyte concentrations over the entire analytical range of the assay.

Noncompetitive immunoassays rely on different epitopes on a large analyte molecule so as to provide a binding site for both the capture antibody and for the labeled antibody. In the past, a small molecule could not be measured using the conventional sandwich format because a small molecule had too few epitopes. One strategy developed to expand the scope of the sandwich assay to small molecules is exemplified by a 25-hydroxy vitamin D (25OH-D) assay. In this assay format, 25OH-D binds to an immobilized capture antibody and a second antibody (an antimetatype antibody) specifically recognizes the immunocomplex formed between the 25OH-D molecule and the capture antibody, hence facilitating a sandwich assay design.

Heterogeneous Versus Homogeneous Immunochemical Assays

Immunochemical assays that require separation of the free from the bound label are termed **heterogeneous immunoassays**. These assays use labeled antigen where the measurable property of the label is the same, whether it is bound to antibody or not; thus bound and free label must be physically separated to measure how much is bound to antibody. **Homogeneous immunoassays** do not require separation because the measurable property of the label depends on whether it is bound to antibody or not.

Heterogeneous immunoassays. Heterogeneous immunoassays implicitly assume that $k_1 >> k_{-1}$, and solid phase adsorption is widely used to separate the free, labeled (Ag:L) from the bound, labeled antigen (Ag:L:Ab).

Solid phase adsorption is the separation technique that currently is the most popular and widely used in both manual and automated heterogeneous immunoassays. In this technique, the binding and competition of labeled and unlabeled antigens for the binding sites of the antibody occur on the surface of a solid support onto which the capture antibody has been attached by physical adsorption or covalent bonding. Several different types of solid supports have been used, including the inner surface of plastic tubes or wells of microtiter plates, and the outer surface of insoluble materials such as cellulose or magnetic latex beads or particles. With tubes and microtiter plates, the solid surface containing the attached antibody and the bound antigen is washed "in place," and indicator reagents are subsequently added to complete the assay. When beads or particles are used, they are added directly to the reaction mixture and, after incubation, are removed by centrifugation or magnetic separation. After the supernatant has been removed by siphoning or decanting, the beads or particles are washed and indicator reagents subsequently added to complete the assay.

Homogeneous immunoassays. Homogeneous immunoassays do not require separation of bound and free labeled antibody or antigen. In this type of assay, the activity of the label attached to the antigen is modulated directly by antibody binding. The magnitude of the modulation is proportional to the concentration of the antigen or antibody being measured. Consequently, it is necessary only to incubate the sample containing the analyte antigen with the labeled antigen and antibody and then directly measure the activity of the label "in place," making these assays technically easier and faster (see Points to Remember box).

POINTS TO REMEMBER

Important nonseparation (homogeneous) immunoassays include:

- Enzyme-multiplied immunoassay technique (EMIT)
- Cloned enzyme donor immunoassay (CEDIA)
- Luminescent oxygen channeling immunoassay (LOCI)
- Fluorescence resonance energy transfer (FRET)

Homogeneous immunoassays that use enzyme labels that are performed on spectrophotometric analyzers (see subsequent descriptions of the enzyme multiplied immunoassay technique [EMIT] and the cloned enzyme donor immunoassay [CEDIA]) are in routine use for a range of small molecules. A homogeneous sandwich format chemiluminescent immunoassay has also been developed that expands the scope of homogeneous assays to large molecules (see subsequent description of the luminescent oxygen channeling immunoassay [LOCI]). In addition, a diverse range of homogeneous fluoroimmunoassays (FIAs) based on fluorescence resonance energy transfer (FRET) between fluorescent donor dye– and fluorescent acceptor dye–labeled assay components represent a further expansion of the scope of homogeneous immunoassays (see subsequent descriptions of FRET).

Immunoassay Calibration

Calibration of an immunoassay involves assaying a series of calibrators with known values and fitting a straight line or curve to the resulting data to determine the relationship between the measurable signal and the analyte concentration. This dose-response curve is then used to determine the concentration of analyte in the patients' samples. Joining successive points in a calibration curve is usually achieved by means of an appropriate mathematical equation. Several curve-fitting methods are in use. Interpolation methods join successive points by straight lines (linear interpolation) or curved lines (curvilinear interpolation). When the latter approach is used, a cubic polynomial ($y = a + bx + cx^2 + dx^3$) links the response (y) to the calibrator concentration (x), and the best fit is attained through a series of recalculations (iterations) that smooth the joints between the curves linking successive points on the curve. The resulting equation is called a *spline function*. Empirical curve-fitting methods use different mathematical models, including (1) hyperbolic, (2) polynomial, and (3) log-logit and its variants (e.g., four-parameter log-logistic), to calculate a curve to fit the calibration data.

It should be appreciated that a source of error with all curve-fitting methods is the uncertainty of the shape of the curve between successive calibrators and the imprecision in the measurement of each calibrator. Imprecision may not be constant over the concentration range represented by the calibrators, and in this case the response variable is termed *heteroscedastic*.

Analytical and Functional Sensitivity

The analytical limits of detection of **competitive immunoassays** are determined principally by the affinity of the antibody. A lower limit of detection of 10 fmol/L (i.e., 600,000 molecules of analyte in a typical sample volume of 100 μL) is possible in a competitive assay when an antibody with an affinity of 10^{12} mol/L is used.

For noncompetitive immunoassays, the ability of the detector to measure the label determines the detection limit of an assay. A radioactive label, such as ^{125}I, has low specific activity (7.5 million labels necessary for detection of 1 disintegration/s) compared with enzyme labels and chemiluminescent and fluorescent labels. Enzyme labels provide an amplification (each enzyme label producing many detectable product molecules), and the detection limit for an enzyme is lowered if the conventional photometric detection is replaced with chemiluminescent or bioluminescent detection. The combination of amplification and an ultrasensitive detection reaction makes noncompetitive chemiluminescent EIAs among the most sensitive types of immunoassays. Fluorescent labels also have high specific activity; a single high–quantum-yield fluorophore is capable of producing 100 million photons/s. In practice, several factors degrade the detection limit of an immunoassay. These include (1) background signal from the detector, (2) assay reagents, and (3) nonspecific binding of the labeled reagent.

Secondary labels such as biotin are used to introduce amplification into an immunoassay. The binding constant of the biotin/avidin complex is extremely high (10^{15} mol/L). This high binding allows for the design of immunoassay systems that are even more sensitive than the simple antibody systems. Such a biotin/avidin system uses a biotin-labeled first antibody. Biotin is attached to the antibody in relatively high proportion without loss of immunoreactivity of the antibody. When an avidin-conjugated label is added, a complex of Ag:Ab-biotin:avidin label is formed. Further amplification is achieved by a biotin:avidin:biotin linkage because the binding ratio of biotin:avidin is 4:1 (e.g., Ag:Ab-biotin:avidin:[3 biotin labels]). If the label is an enzyme, large numbers of enzyme molecules in the complete complex provide a large increase in enzymatic activity, coupled with the small amount of antigen being determined, and the antigen assay is correspondingly more sensitive.

Another type of sensitivity is termed "functional sensitivity." This is defined as the lowest concentration that can be measured in an assay with an interassay coefficient of variation of 20% (see also Chapter 2). This is used to establish a more realistic and robust detection limit for an assay used in patient care. Functional sensitivity is associated with the concept of assay generations, each successive generation representing a one-log concentration improvement in sensitivity (e.g., for a thyroid-stimulating hormone [TSH] immunoassay first generation 1 mIU/L, second generation 0.1 mIU/L, etc).

Examples of Labeled Immunoassays

Specific examples of different types of labeled immunoassays are discussed in the following section.

Radioimmunoassay. Radioimmunoassays (RIAs) use radioactive isotopes of iodine, ^{125}I and ^{131}I, and tritium (^{3}H) as labels. In practice, competition between radiolabeled and unlabeled antigen or antibody in an antigen-antibody reaction is used to analytically determine the concentration of the unlabeled antigen or antibody. It takes advantage of the specificity of the antigen-antibody interaction and the ability to measure very low quantities of radioactive elements. RIAs have been used to determine the concentration of antibodies or any antigen against which a specific antibody is produced. When used to measure the concentration of an antigen, RIA requires that the antigen be available in a pure form and be labeled with a radioactive isotope. An alternative assay design uses labeled antibody (e.g., immunoradiometric assay [IRMA]) and does not require purified antigen because the antigen need not be labeled. This obviates potential problems that may be caused by iodination of labile antigens. Antibodies are more stable proteins and are easier to label without damage to the function of the protein.

Nonseparation RIAs have also been developed that are based on the modulation of a tritium or ^{125}I label by microparticles loaded with a scintillant. These scintillation proximity assays (SPAs) have found routine application in high-throughput screening assays for drug discovery.

Although they were once popular, the use of RIAs in clinical laboratories has declined primarily because of concerns over safe handling and disposal of radioactive reagents and waste.

Enzyme immunoassay. EIAs use the catalytic properties of enzymes to detect labeled antigens or antibodies, discriminate between bound and free label, and quantify immunological reactions. Alkaline phosphatase (ALP), horseradish peroxidase (HRP), glucose-6-phosphate dehydrogenase (G6PDH), and β-galactosidase labels predominate in EIA.

Various detection systems have been used to monitor and quantify EIAs. Assays that produce compounds that can be monitored photometrically are very popular because compact, high-performance photometers are available. However, EIAs that use fluorescent- or chemiluminescent-labeled substrates or products are often preferred to photometry-based assays, owing to the inherent sensitivity of fluorescent and chemiluminescent measurements. Immunoassays that incorporate HRP as a label can be assayed by chemiluminescence using a mixture of luminol, peroxide, and an enhancer such as *p*-iodophenol (Fig. 15.7A), or by using an acridan derivative. A very sensitive assay for ALP labels uses a chemiluminescent adamantyl 1,2-dioxetane aryl phosphate substrate (see Fig. 15.7B). The enzyme dephosphorylates the substrate, which decomposes with a concomitant long-lived glow of light (detection limit for ALP using this assay is 10 zeptomoles [10^{-20} moles]).

Types of EIA include (1) enzyme-linked immunosorbent assay (ELISA), (2) EMIT, and (3) CEDIA.

Enzyme-linked immunosorbent assay. ELISA is a heterogeneous EIA technique. With this type of assay, one of the reaction components is attached to the surface of a solid phase, such as that of a microtiter well. This attachment may consist of nonspecific adsorption or chemical or immunochemical bonding and facilitates separation of bound and free labeled reactants. Typically with ELISA, an aliquot of sample or calibrator containing the antigen to be measured is added to and allowed to bind with a solid phase antibody. After the solid phase has been washed, an enzyme-labeled antibody different from the bound antibody is added and forms a "sandwich complex" of solid phase–Ab:Ag:Ab enzyme. Excess (unbound) antibody then is washed away, and enzyme substrate is added. The enzyme label then catalyzes the conversion of substrate to product(s), the amount of which is proportional to the quantity of antigen in the sample. Antibodies in a sample also are quantified through the use of an ELISA procedure in which antigen instead of antibody is bound to a solid phase, and the second reagent is an enzyme-labeled antibody specific for the analyte antibody. For example, in a microtiter plate format, ELISA assays have been used extensively for detection of antibodies to viruses and parasites in serum or whole blood. In addition, enzyme conjugates coupled with substrates that produce visible products have been used to develop ELISA-type assays with results that are interpreted visually.

Enzyme-multiplied immunoassay technique. EMIT is a homogeneous EIA that is very widely used in clinical analyses, an illustration of which is shown in Fig. 15.8. Because EMIT does not require a separation step, it is simple to perform and has been used to develop a wide variety of small molecule (drug, hormone, or metabolite) assays. Because of their operational simplicity, EMIT-type assays are easily automated and are included in the repertoire of many automated clinical and immunoassay analyzers. In this technique the patient's sample is incubated with (1) an enzyme conjugated to the antigen (drug, hormone, or metabolite), (2) antibody against the antigen, and (3) the enzyme substrate. Antigen in the patient's sample competes with the enzyme-conjugated antigen for the antibody. Binding of the antibody to the enzyme-antigen conjugate affects enzyme activity by physically blocking access of the substrate to the active site of the enzyme or by changing the conformation of the enzyme molecule and thus altering its activity. The relative change in enzyme activity resulting

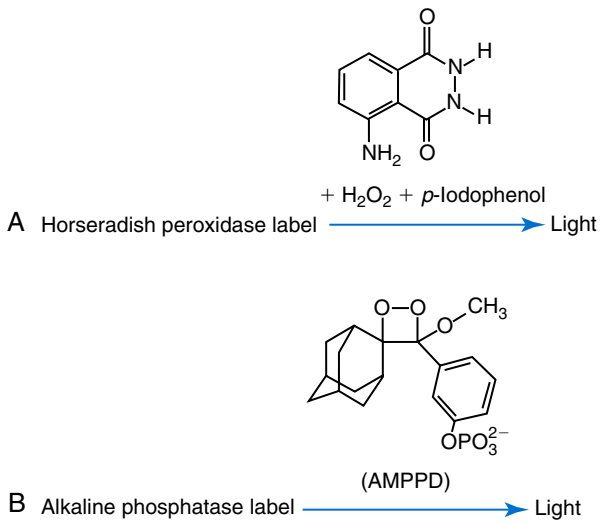


Fig. 15.7 Ultrasensitive assays using horseradish peroxidase and alkaline phosphatase labels. (A) Chemiluminescent assay for horseradish peroxidase label using luminol. (B) Chemiluminescent assay for an alkaline phosphatase label using AMPPD (disodium 3-(4-methoxyspiro[1,2-dioxetane-3,2′-tricyclo[3.3.1.1]-decan]4-yl)phenyl phosphate).

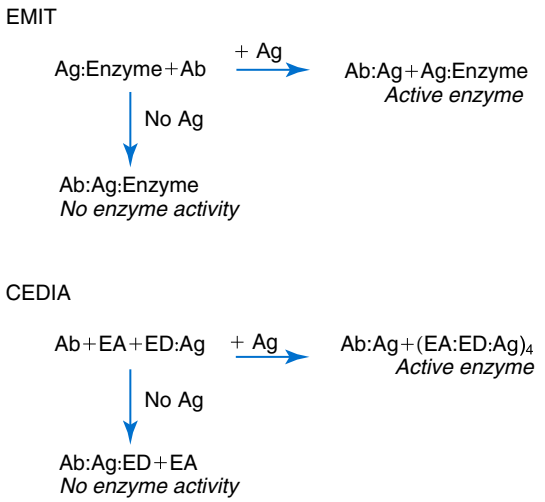


Fig. 15.8 Enzyme-multiplied immunoassay technique *(EMIT)* and cloned enzyme donor homogeneous immunoassay *(CEDIA). Ab,* Antibody; *Ag,* antigen; *EA,* enzyme acceptor; *EA:Ag,* antigen labeled enzyme acceptor; *(EA:ED:Ag)$_4$,* enzymatically active tetrameric β-galactosidase; *ED,* enzyme donor.

from the formation of the antigen-antibody complex is proportional to the antigen concentration in the patient's sample. Concentration of the antigen is calculated from a calibration curve prepared by analyzing calibrators that contain known quantities of the antigen in question.

Cloned enzyme donor immunoassay. As shown in Fig. 15.8, CEDIA is a homogeneous EIA; it was the first EIA designed and developed through the use of genetic engineering techniques. Two inactive fragments (the enzyme "donor" and "acceptor") of β-galactosidase are prepared by manipulation of the Z gene of the *lac* operon of *Escherichia coli.* These two fragments spontaneously reassemble to form active enzyme even if the enzyme donor is attached to an antigen. However, binding of antibody to the enzyme donor–antigen conjugate inhibits reassembly, and no active enzyme is formed. Thus competition between the antigen and the enzyme donor–antigen conjugate for a fixed amount of antibody in the presence of the enzyme acceptor modulates the measured enzyme activity. High concentrations of the antigen result in the greatest enzyme activity; low concentrations result in the least enzyme activity.

Fluoroimmunoassay. Fluoroimmunoassay (FIA) uses a fluorescent molecule as an indicator label to detect and quantify immunologic reactions. Examples of fluorophores used as labels in FIA and their properties are listed in Table 15.2. An early problem with FIA was that background fluorescence from within the sample limited its utility. This problem has been overcome by the use of time-resolved immunoassay techniques that use chelates of rare earth (lanthanide) elements as labels (see Chapter 9). These techniques are based on the fact that fluorescent emissions from lanthanide chelates such as (1) europium, (2) terbium, and (3) samarium have long lives (>1 μs) compared with the typical background fluorescence encountered in biologic specimens. In a time-resolved FIA, a europium chelate label is excited by a pulse of excitation light (0.5 μs), and the long-lived fluorescence emission from the label is measured after a delay (400 to 800 μs). The measurement after the delay ensures that any short-lived background signal has decayed.

Fluorescence polarization immunoassay (FPIA) is a type of homogeneous FIA that is widely used to measure drugs and other small molecules (Fig. 15.9). The polarization of the fluorescence from a fluorescein-antigen conjugate is determined by its rate of rotation during the lifetime of the excited state in solution. When the fluorescein-antigen conjugate is bound to the large antibody molecule, the fluorophore is constrained from rotating in the time between absorption of the incident radiation and emission of fluorescence; hence the fluorescence emission is still highly polarized. In contrast, when the fluorescein-antigen conjugate is free in solution, it can rotate more rapidly and the emitted light is depolarized. Thus binding to antibody modulates polarization, and a homogeneous assay is possible.

Förster or FRET is the distance-dependent transfer of energy from a fluorescent donor dye to a fluorescent acceptor dye. Distances at which there is 50% transfer efficiency are typically 1 to 20 nm, and up to approximately 20 nm is possible for some donor-acceptor pairs (e.g., lanthanide chelates and quantum dots). In, for example, a sandwich FRET immunoassay format used in the time-resolved amplified cryptate emission (TRACE) assay, an antibody labeled with a europium chelate (donor) is matched with an antibody labeled with an allophycocyanin dye (XL665; acceptor). The two antibodies form a sandwich immunocomplex with the antigen, and following irradiation with a pulsed excitation light (337 nm), energy transfer occurs from the donor europium chelate to the XL665 acceptor. In the absence of immunocomplex formation, the XL665 fluorescence is short-lived (nanoseconds). However, with immunocomplex formation, the europium chelate on the first antibody provides sustained energy to excite and prolong the normally short-lived fluorescence of the XL665 dye on the second antibody (microseconds).

TABLE 15.2 Properties of Fluorescent Labels

Fluorophore	Excitation, nm	Emission, nm	Fluorescence Quantum Yield[a]	Lifetime, ns
Cy3 (cyanine dye)	550	570	0.15	—
Cy5 (cyanine dye)	650	670	0.28	—
Europium (β-naphthoyl trifluoroacetone)	340	590,613	—	500,000
Fluorescein isothiocyanate	492	520	0.0–0.85	4.5
NN382	778	806	0.59	—
Phycobiliprotein	550–620	580–660	0.5–0.98	—
Rhodamine B isothiocyanate	550	585	0.0–0.7	3.0
Umbelliferone	380	450	—	—

[a]Fluorescence quantum yield: fraction of molecules that emit a photon.

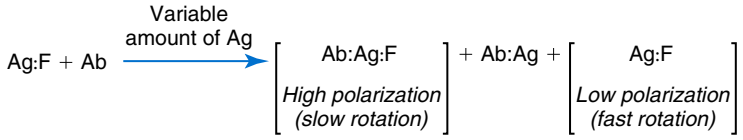


Fig. 15.9 Homogeneous polarization fluoroimmunoassay. *Ab,* Antibody; *Ag,* antigen; *F,* fluorescein.

Intensity of the 665-nm emission is proportional to antigen concentration. Another strategy used in a system for cardiac markers uses an antibody-coated latex particle containing a donor dye (a silicon phthalocyanine derivative) and a second antibody–coated latex particle containing an acceptor dye (a different silicon phthalocyanine dye). Analyte in the sample binds to create donor particle:analyte:acceptor particle complexes. Excitation at 670 nm excites the donor dye, and energy transfer to the acceptor dye occurs due to the proximity of the two differently dyed particles. The acceptor particle then emits light at 760 nm (long wavelength in the red portion of the spectrum). An advantage of this red emission is that scattering and fluorescence from blood plasma is minimized.

Phosphor immunoassay. A phosphor is a material that emits light (phosphorescence) over a relatively long time scale following exposure to excitation energy. A particular type of phosphor, an upconverting phosphor nanoparticle, can be used as a label in immunoassays (e.g., to label an antibody in a sandwich immunoassay). In one application, the upconverting phosphor nanoparticle (200- to 400-nm diameter) can be a crystalline lanthanide oxysulfide. This nanoparticle absorbs two or more photons of infrared light (980 nm) and produces light emission at a shorter wavelength (anti-Stokes shift). The phosphorescence is not influenced by reaction conditions (e.g., temperature, buffer), and no upconverted signal is obtained from biologic components in the sample (low background). Multiplexing is possible because different types of particles produce different wavelengths of phosphorescence (e.g., yttrium/erbium oxysulfides are green [550 nm], yttrium/thulium oxysulfide particles are blue [475 nm]).

Chemiluminescent immunoassay. Chemiluminescence is the name given to light emission produced during a chemical reaction. Isoluminol and acridinium esters are important examples of chemiluminescent labels used in chemiluminescent immunoassays. Oxidation of isoluminol by hydrogen peroxide in the presence of a catalyst, such as microperoxidase, produces a relatively long-lived light emission at 425 nm, and oxidation of an acridinium ester by alkaline hydrogen peroxide in the presence of a detergent (e.g., Triton X-100) produces a rapid flash of light at 429 nm. Acridinium and sulfonyl acridinium esters are high specific activity labels (detection limit for the label is 800 zeptomoles) that can be used to label both antibodies and haptens (Fig. 15.10A).

LOCI is a particularly important type of chemiluminescent immunoassay. It is one of the few homogeneous immunoassays that operate in a noncompetitive ("sandwich") format and can be used to assay large molecules. LOCI uses two reagent particles (sensitizer and chemiluminescer particles) that form a complex with the analyte of interest. The presence of analyte links the two reagent latex particles in close proximity. The first particle contains a photosensitizer (e.g., phthalocyanine) and, in the presence of light, converts oxygen to singlet oxygen. The second chemiluminescer particle contains a chemiluminescent agent (e.g., olefin) that reacts with singlet oxygen to form a dioxetane that decomposes and emits light. This reaction occurs only if the two particles are in close proximity and the singlet oxygen can diffuse efficiently from the sensitizer particle to the chemiluminescer particle. Singlet oxygen does not react with unbound chemiluminescer particles because of the short lifetime of this transient species in an aqueous environment. An example of a LOCI assay for TSH is illustrated in Fig. 15.11. TSH binds to a biotinylated anti-TSH antibody, and this complex links

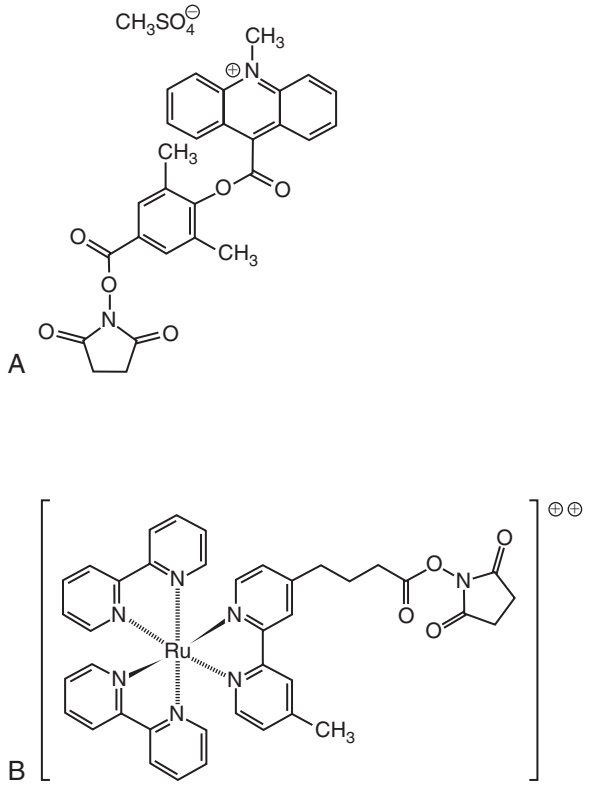


Fig. 15.10 Luminescent labels. (A) Chemiluminescent acridinium ester label. (B) Electrochemiluminescent ruthenium (II) tris(bipyridyl) NHS (*N*-hydroxysuccinimide) ester label. ([A], From Law, S.-J., Miller, T., Piran, U., Klukas, C., Chang, S., & Unger, J. [1989]. Novel poly-substituted aryl acridinium esters and their use in immunoassay. *Journal of Bioluminescence and Chemiluminescence, 4*, 88–98.)

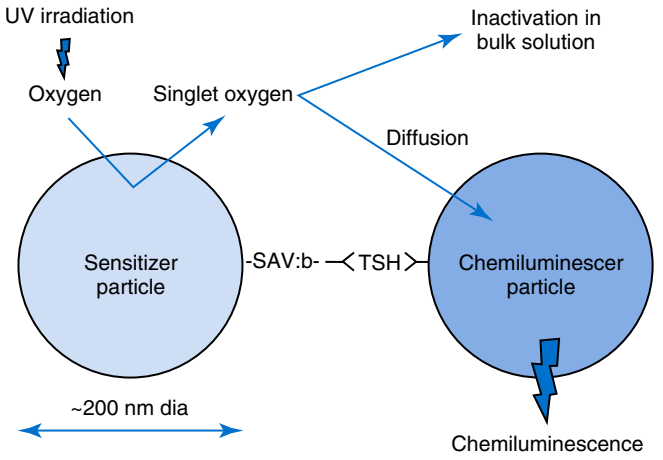


Fig. 15.11 Example of luminescent oxygen channeling immunoassay for thyroid-stimulating hormone *(TSH)*. In this assay, TSH links two latex microbeads—one containing a photosensitive reagent and the other, a precursor of a chemiluminescent reagent. *b, Biotin; SAV, streptavidin; UV, ultraviolet.*

the streptavidin-coated sensitizer particle to the anti-TSH antibody–coated chemiluminescer particle. Exposure of light results in emission of singlet oxygen from the photosensitive particle. The singlet oxygen activates the chemiluminescent particle to emit light that is then measured.

Electrochemiluminescence immunoassay. Ruthenium (II) tris(bipyridyl) (see Fig. 15.10B) undergoes an electrochemiluminescent reaction (620 nm) with tripropylamine at an electrode surface, and this chelate is now used as a label in competitive and sandwich electrochemiluminescence immunoassays. Using this label, various assays have been developed with magnetic beads as the solid phase. Beads are captured by a magnet at an electrode surface in a flow cell, and unbound label is washed out of the cell by a wash buffer. Label bound to the bead undergoes an electrochemiluminescent reaction, and the light emission is measured by an adjacent photomultiplier tube.

Particle immunoassay. Various micro- and nano-sized particles can be used as labels. For example, dyed-latex particles and gold sols are visible to the naked eye and have found wide application in simplified immunoassays based on the lateral flow principle (Chapter 17). A magnetic particle, detected via its reflective and light scattering properties, is used in an optomagnetic sandwich immunoassay. Magnetic nanoparticles coated with detection antibody bind to captured antigens and are detected using frustrated total internal reflection. An advantage of this assay format is speed (external fields are used to manipulate the magnetic particles), and it has been applied to a fast turnaround intraoperative parathyroid hormone assay.

Nuclear magnetic resonance detection of magnetic particle labels facilitates a homogeneous immunoassay design. The assay uses magnetic nanoparticles (~20-nm diameter) coated with antibody that binds to and clusters around analyte in solution. A consequence of the binding and clustering is that the microscopic environment of water in the reaction mixture is altered, and this can be detected by nuclear magnetic resonance as a change in the T2 relaxation signal (Fig. 15.12). An advantage of this type of immunoassay is that the signal is unaffected by the sample matrix (e.g., opacity), and this allows direct detection in various types of specimens (e.g., whole blood and sputum).

Immuno-polymerase chain reaction and bio-barcode immunoassay. DNA has been used in a number of different immunoassay designs. Immuno-PCR is a heterogeneous immunoassay in which a piece of single- or double-stranded DNA is used as a label for an antibody in a sandwich assay. Bound DNA label is amplified using the polymerase chain reaction (PCR). The amplified DNA product is separated by gel electrophoresis and quantitated by densitometric scanning of an ethidium-stained gel.

In a bio-barcode assay, capture antibody immobilized on a magnetic particle binds to the analyte (e.g., a protein) and the bound analyte is reacted with a gold particle (30-nm diameter) decorated with detection antibody and barcode dsDNA (Fig. 15.13; refer to the figure legend for details). Following complex formation and washing, one strand of the barcode dsDNA is released from the gold particle. Next, the released barcode dsDNA is hybridized to an immobilized capture DNA probe and to a gold particle-labeled detection probe. The bound gold particle–labeled detection probe is first decorated with silver and then detected scanometrically. This type of assay has achieved very low detection limits (e.g., pg/mL range for serum prostate-specific antigen).

Multiplexed immunoassays and protein microarrays. Multiplexed immunoassays, in which two or more analytes are detected/measured in a single assay, improve the efficiency of detecting/measuring multiple analytes. Two important strategies are based on discrete reaction zones on a planar surface (microarray) and on optically coded microspheres ("liquid array").

Arrays of hundreds or thousands of micrometer-sized dots of antigens or antibodies immobilized on the surface of a glass or plastic chip are emerging as an important tool in proteomic studies and in assessing protein-protein interactions. This format facilitates simultaneous multianalyte immunoassays using (e.g., fluorophore-labeled conjugates). The arrays

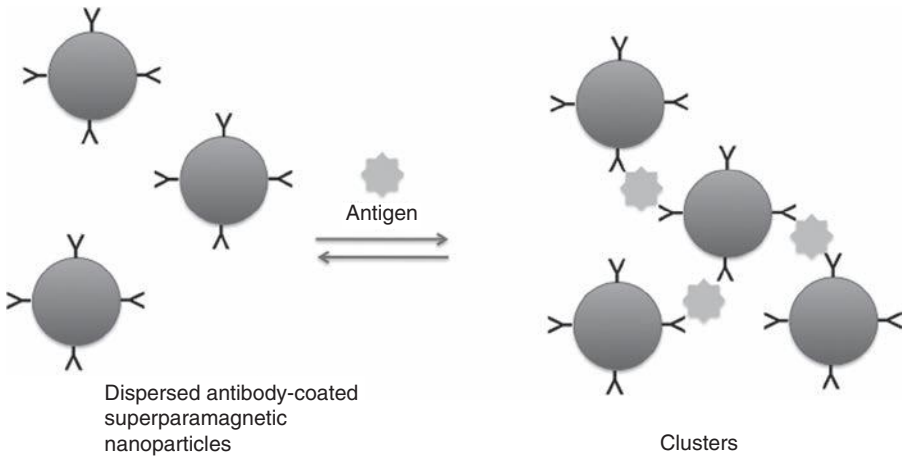


Fig. 15.12 Magnetic particle immunoassay. Antigen binds to magnetic nanoparticles (~20-nm diameter) coated with antibody. The binding and resulting clustering of the particles alters the microscopic environment of water in the reaction mixture and this can be detected by nuclear magnetic resonance (change in the T2 relaxation signal).

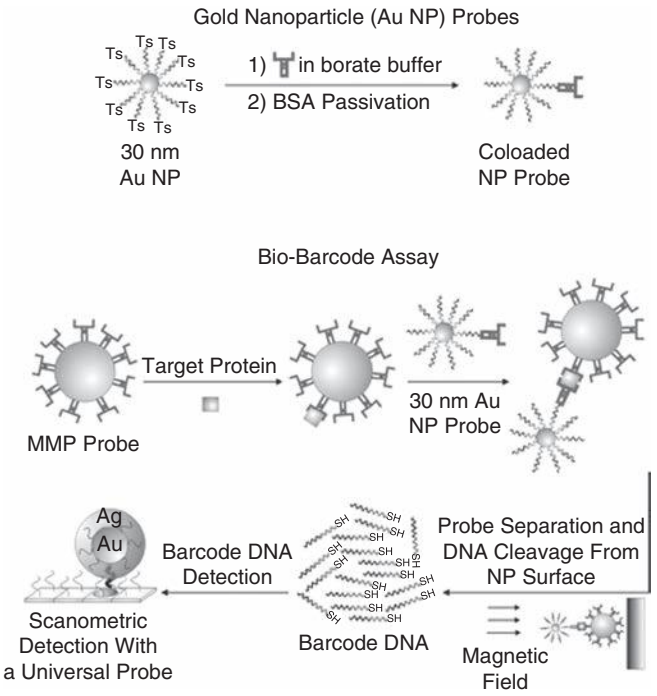


Fig. 15.13 Bio-barcode assay for a protein target. Barcode DNA–functionalized gold nanoparticles *(Au-NP)* (30-nm diameter) are conjugated to target protein–specific antibodies (via tosyl *[Ts]* groups) to generate the coloaded target protein Au-NP probes that are then passivated with bovine serum albumin *(BSA)*. The next steps in the sandwich immunoassay are reaction of the target protein with 1-μm magnetic microparticle probes *(MMPs)* coated with monoclonal antibodies to target protein, washing to remove excess serum components, and then reaction with the Au-NP. After magnetic separation and wash steps, the target protein–specific DNA barcodes are released into solution and detected using the scanometric assay (includes a Au-NP–catalyzed silver enhancement step). Approximately one-half of the barcode DNA sequence *(green)* is complementary to the "universal" scanometric Au-NP probe DNA, and the other half *(purple)* is complementary to a surface immobilized DNA sequence that is responsible for sorting and binding barcodes complementary to the target protein barcode sequence. (From Thaxton, C. S., Elghanian, R., Thomas, A. D., Stoeva, S. I., Lee, J.-S., Smith, N. D., et al. [2009]. Nanoparticle-based bio-barcode assay redefines "undetectable" PSA and biochemical recurrence after radical prostatectomy. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 18437–18442, with permission.)

are made by printing or spotting 1-nL drops of protein solutions onto a flat surface, such as a glass microscope slide. In a typical sandwich assay, the array on the surface of the slide is incubated with sample and then with conjugate. The fluorescence of the bound conjugate is detected using a scanning device. The pattern of the signal provides information on the presence and amount of individual analytes in the sample or the reactivity of a single analyte with the range of proteins arrayed on the surface of the slide.

Another type of microarray, the so-called liquid array, is based on collections of microbeads optically coded with fluorescent dyes. Each type of fluorescent bead is coated with a different antibody and following the immunologic reaction with sample and labeled antibody, the formation of an immunocomplex on each bead is assessed using flow cytometric or fluorescence imaging principles. Up to 500 different signatures can be created allowing multiplexing of up to 500 different assays in a single assay vessel.

Digital immunoassay. A long-cherished goal in immunoassay has been single molecule sensitivity. A new digital immunoassay design has led to dramatic improvements in sensitivity (Fig. 15.14). This ELISA type of assay is performed on small antibody-coated paramagnetic microbeads. At the end of the incubation of the capture antibody–coated beads with sample and β-galactosidase labeled antibody, the beads, in the presence of a fluorogenic substrate, are distributed into an array of 216,000 wells (one bead per well) each with a volume of 40 fL. A single target molecule captured onto the bead in an fL-sized well generates sufficient fluorescence signal for detection. The signal from the array of wells is imaged using a charge coupled device (CCD) camera and the number of fluorescent wells is counted (a digital read-out), thus counting the number of molecules in the sample. At the 10:1 bead-to-molecule ratio used in the assay, the number of beads that carry a labeled immunocomplex follows a Poisson distribution, such that at low protein concentrations each bead will capture one labeled immunocomplex or none. The combination of the sensitive label detection method and the low background achieved by the signal counting leads to protein assays with fg/mL sensitivity.

FIA can also be operated in a digital mode by combining a magnetic microparticle-based sandwich FIA and single-molecule counting technology. It integrates capillary flow, laser-induced fluorescence, and a highly sensitive detection optics module for sample analysis. Upon completion of sandwich formation, the fluorescent dye–labeled detection antibody is released from the captured antigen and is eluted into a very small volume (20 μL) for counting. Fluorescence detection occurs in a flow system in a very small interrogation space (5 μm) that is illuminated by a laser. Signals above the background are counted as digital events, and the sum of the digital events is related to the original concentration of the analyte in the sample.

Interferences. Immunologic assays are prone to interferences, in spite of the use of highly specific antibodies for molecular recognition of the analyte. Falsely low results can occur because of the hook effect at high antigen concentrations (see earlier discussion). False-negative or false-positive results are encountered if the sample contains anti–animal immunoglobulin antibodies. For example, in a two-site sandwich assay for human chorionic gonadotropin (hCG) based on mouse antibodies, any human anti-mouse antibodies (HAMAs) present in the specimen will recognize the immobilized mouse capture and mouse conjugate antibodies and will form a complex that is indistinguishable from an immobilized capture antibody:hCG:conjugate complex. This leads to a false-positive result. A false-negative result will be obtained if the HAMAs react with the capture antibody or the conjugate to such an extent that specific antibody binding to hCG is prevented. Many different types of circulating anti–animal immunoglobulin antibodies have been detected (e.g., human anti-goat, human anti-bovine antibodies) and shown to interfere in immunoassays. In practice, this type of interference is minimized by including additives in the immunoassay reagents such as nonimmune serum or IgG from the

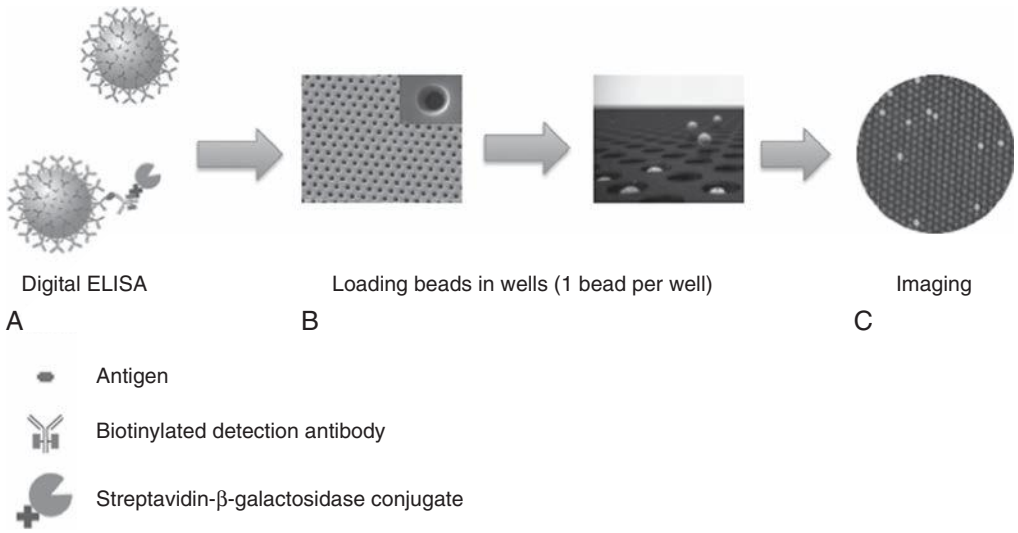


Fig 15.14 Design of a digital immunoassay (Simoa). (A) Reaction of the antigen with capture antibody–coated magnetic beads and biotinylated detection antibody. This is then followed by reaction of the bound complex with streptavidin–β-galactosidase conjugate to form single immunocomplexes of the magnetic beads. (B) The beads are loaded into wells, one bead per well, and then treated with a fluorometric substrate for β-galactosidase. (C) Fluorescence from the beads in the array detected by imaging. *ELISA*, Enzyme-linked immunosorbent assay. (Images used with permission of Quanterix Corporation.)

species used to raise the antibodies used in the assay. In practice, this type of interference can sometimes be uncovered by performing dilution studies (nonlinear response) or from changes in values following incubation of the sample in a heterophile blocking tube.

CELL-BASED AND TISSUE-BASED IMMUNOCHEMICAL TECHNIQUES

Other analytical methods of clinical interest that use antibodies include immunocytochemistry and agglutination assays. In immunocytochemical assays, labeled (e.g., horseradish peroxidase) antibody reagents are used as specific probes for protein and peptide antigens to evaluate single cells or pieces of tissue for their synthetic capability and/or phenotypic identity. In an agglutination method, the visible clumping of particulates, such as cells and latex particles, is used as an indicator of the primary reaction of antigen and antibody.

Hemagglutination, agglutination reactions in which the antigen is located on an erythrocyte, is used in the direct testing of erythrocytes for blood group, Rh, and other antigenic types.

REVIEW QUESTIONS

1. What does a strong signal in a microtiter well–based competitive enzyme immunoassay (EIA) for thyroxine signify?
   a. A low concentration of thyroxine in the sample
   b. A high concentration of thyroxine in the sample
   c. Enzyme substrate not added to the well
   d. Enzyme inhibitor present in the sample
   e. Bound antibody has detached from the surface of the well
2. What does a very weak signal in a microtiter well–based sandwich EIA for thyroid-stimulating hormone (TSH) signify?
   a. A low concentration of TSH in the sample
   b. A high concentration of TSH in the sample
   c. A hook effect
   d. Effective competition between the conjugate and the TSH for capture antibody
   e. Capture antibody saturated with TSH:conjugate complexes
3. The energy of interaction of a single antibody-combining site and its corresponding epitope on the antigen is referred to as:
   a. sensitivity.
   b. specificity.
   c. immunogenicity.
   d. affinity.
4. Which one of the following immunochemical assays requires separation of the free label from the bound labeled substance?
   a. Enzyme-multiplied immunoassay technique (EMIT)
   b. Enzyme-linked immunosorbent assay (ELISA)
   c. Cloned enzyme donor immunoassay (CEDIA)
   d. Luminescent oxygen channeling immunoassay (LOCI)
   e. Fluorescence polarization immunoassay (FPIA)

5. The addition of a linear polymer such as polyethylene glycol to a mixture of antigen and antibody causes_____ in the rate of immune complex formation and precipitation.
   a. a decrease
   b. an increase
   c. no change
6. Transferring electrophoretically separated proteins into a strip of nylon membrane by electroblotting is the second step in:
   a. radial immunodiffusion.
   b. enzyme-linked immunosorbent assay.
   c. Western blotting.
   d. immunofixation.
   e. crossed immunoelectrophoresis.
7. In a two-site antibody-based sandwich immunoassay, the presence of human anti-mouse antibodies produces a false-negative result by:
   a. binding a mouse immunoglobulin capture antibody and the mouse immunoglobulin conjugate to form a bridge.
   b. mimicking the specific analyte being assessed.
   c. binding to the mouse monoclonal capture antibody reagents.
   d. reacting with one of the assay reagents to prevent formation of the sandwich.
8. Which one of the following immunoassay procedures does not involve the coating of a solid phase with antibody or antigen?
   a. Enzyme-multiplied immunoassay technique (EMIT)
   b. Luminescent oxygen channeling immunoassay (LOCI)
   c. Enzyme-linked immunosorbent assay (ELISA)
   d. Particle-enhanced turbidimetric inhibition immunoassay (PETINIA)
   e. Bio-barcode immunoassay
9. The immunoglobulin most often used in immunochemical assays is:
   a. IgG.
   b. IgA.
   c. IgM.
   d. IgE.
10. Which one of the following statements is false?
   a. Dyed-latex microparticles are effective labels in simplified immunoassays because they are visible to the naked eye
   b. Binding of an antigen to an antibody involves van der Waals-London dipole-dipole interaction
   c. Addition of polymers to an antigen antibody reaction increases the rate of immune complex growth
   d. The functional sensitivity of an immunoassay is the lowest concentration that can be measured with a within-assay coefficient of variation of 20%
   e. Fluorescence resonance energy transfer from donor to acceptor dye is distance dependent

SUGGESTED READINGS

Blackburn, G. F., Shah, H. P., Kenten, J. H., Leland, J., Kamin, R. A., Link, J., et al. (1991). Electrochemiluminescence detection for development of immunoassays and DNA probe assays for clinical diagnostics. *Clinical Chemistry*, *37*, 1534–1539.

Kricka, L. J. (1991). Chemiluminescent and bioluminescent techniques. *Clinical Chemistry*, *37*, 1472–1481.

Luo, Z. X., Fox, L., Cummings, M., et al. (2016). New Frontiers in *in vitro* Medical Diagnostics by Low Field T2 Magnetic Resonance Relaxometry. *Trends in Analytical Chemistry*. https://doi.org/10.1016/j.trac.2016.02.025.

Omi, K., Ando, T., Sakyu, S., Uchida, Y., Oka, A., Ise, N., et al. (2015). Noncompetitive immunoassay detection system for haptens on the basis of antimetatype antibodies. *Clinical Chemistry*, *61*, 627–635.

Park, J. Y., & Kricka, L. J. (2013). Interferences in immunoassay. In D. Wild (Ed.), *The immunoassay handbook* (4th ed.). (pp. 403–416). Oxford: Elsevier.

Rissin, D. M., Fournier, D. R., Piech, T., Kan, C. W., Campbell, T. G., Song, L., et al. (2011). Simultaneous detection of single molecules and singulated ensembles of molecules enables immunoassays with broad dynamic range. *Analytical Chemistry*, *83*, 2279–2285.

Templin, M. F., Stoll, D., Schrenk, M., Traub, P. C., Vohringer, C. F., & Joos, T. O. (2002). Protein microarray technology. *Trends in Biotechnology*, *20*, 160–166.

Todd, J., Freese, B., Lu, A., Held, D., Morey, J., Livingston, R., et al. (2007). Ultrasensitive flow-based immunoassays using single-molecule counting. *Clinical Chemistry*, *53*, 1990–1995.

Turgeon, M. L. (2014). *Immunology and serology in laboratory medicine* (5th ed.) (pp. 584).Elsevier.

Ullman, E. F. (2013). Homogeneous immunoassays. In Wild (Ed.), *The immunoassay handbook* (4th ed.) (pp. 67–87). Amsterdam: Elsevier.

Wild, D. (2013). *The immunoassay handbook* (4th ed.) (pp.1013). Oxford: Elsevier.

16

Automation

*Charles D. Hawker, Jonathan R. Genzen, Carl T. Wittwer**

OBJECTIVES

1. List four benefits of using automation in the clinical laboratory.
2. State four advantages of using bar-coded labels to identify specimens.
3. List and describe the considerations that must be evaluated when an automated laboratory system is selected.
4. List the problems encountered when automated laboratory processes are used and integrated, including specimen identification, specimen delivery, reagent identification, and the chemical reaction phases.
5. Compare the following analyzer configurations with regard to specimen and reagent handling and processing:
   - Random-access
   - Continuous-flow
   - Sequential
   - Multiple-channel
   - Discrete
   - Single-channel
6. Compare an "open" versus a "closed" automated analyzer with regard to reagent usage.
7. Describe how automation has been applied to polymerase chain reaction (PCR) and deoxyribonucleic acid (DNA) sequencing.
8. Describe the trade-offs between automating complex processes and developing simpler processes that do not require automation.

KEY WORDS AND DEFINITIONS

Automation The process whereby an analytical instrument performs many tests with only minimal involvement of an analyst; also defined as the controlled operation of an apparatus, process, or system by mechanical or electronic devices without human intervention.

Batch analysis Type of analysis in which many specimens are grouped in the same analytical session.

Carryover The transport of a quantity of analyte or reagent from one specimen reaction into and contaminating a subsequent one.

Continuous-flow analysis Type of analysis in which each specimen in a batch passes through the same continuous stream at the same rate and is subjected to the same analytical reactions.

Molecular diagnostics Using the presence, quantity, or sequence of DNA or RNA to identify the nature of an illness.

Multiple-channel analysis Type of analysis in which each specimen is subjected to multiple analytical processes so that a set of test results is obtained on a single specimen; similar to random-access analysis.

Polymerase chain reaction (PCR) A molecular biology method to amplify a segment of DNA by repeated cycles of primer annealing, polymerase extension, and product denaturation.

Random-access analysis The most common configuration of an automated analyzer, in which analyses are performed on a collection of specimens sequentially and each specimen is analyzed for a different selection of tests.

Real-time PCR A variant of PCR where fluorescence is measured during amplification, allowing quantification of the initial template.

Sequential analysis Type of analysis in which each specimen in a batch enters the analytical process one after another, and each result or set of results emerges in the same order as the specimens are entered.

Single-channel analysis Type of analysis in which each specimen is subjected to a single process so that only

*The authors would like to acknowledge the original contributions of Ernest Maclin, PE; Donald S. Young, MB, PhD; and James C. Boyd, MD, upon which portions of this chapter are based.

results for a single analyte are produced; similar to batch analysis.

Southern blotting A method to identify and measure a specific DNA sequence or genes in a mixed extract, where DNA strands are cut with restriction enzymes, separated by gel electrophoresis, transferred to special filter paper, and hybridized.

Throughput The number of specimens processed by an analyzer during a given period of time, or the rate at which an analytical system processes specimens.

Workstation An area in the clinical laboratory dedicated to a defined task and containing appropriate laboratory instrumentation to carry out that task.

The term **automation** has been applied in clinical chemistry to describe the process whereby an analytical instrument performs many tests with only minimal involvement of an analyst. However, the term also applies to the automation of nonanalytical processes (preanalytical and postanalytical) that have become very important to the performance of laboratories over the past two decades. This chapter covers both nonanalytical and analytical automation in clinical laboratories.

The availability of automation systems and automated instruments enables laboratories to process much larger workloads without comparable increases in staff. The evolution of automation in clinical laboratories has paralleled that in the manufacturing industry, progressing from fixed automation, whereby an instrument performs a repetitive task by itself, to programmable automation, which allows the instrument to perform a variety of different tasks. Intelligent automation has also been introduced into some individual instruments or systems to allow them to self-monitor and respond appropriately to changing conditions.

One benefit of automation is a reduction in the variability of results and errors of analysis through the elimination of tasks that are repetitive and monotonous for most individuals. The improved reproducibility gained by automation has led to a significant improvement in the quality of laboratory tests.

Many small laboratories now have consolidated into larger, more efficient entities in response to market trends involving cost reduction. The drive to automate these mega-laboratories has led to new avenues in laboratory automation. No longer is automation simply being used to assist the laboratory technologist in test performance; it now includes (1) processing and transport of specimens, (2) loading of specimens into automated analyzers, (3) assessment of the results of the tests performed, and (4) storage of specimens.

Clinical laboratory automation as we know it today owes much to the pioneering efforts of Masahide Sasaki in Kochi, Japan. Beginning in 1981, Dr. Sasaki trained his medical technologists to assemble conveyor belts, construct electronic boards, and program robots, thus leading to the creation of the world's first automated laboratory. When he first revealed his automated laboratory to Western audiences in 1989, Sasaki demonstrated new possibilities for error reduction, cost containment, and unprecedented turn-around times. This led to many visitors to his laboratory as well as collaborations with diagnostic vendors that brought commercial automation products to clinical laboratories throughout Japan. Much of the automation in our laboratories today traces its lineage to these early automation systems in Japan. Thousands of laboratories worldwide now have either total laboratory automation (TLA) systems or automated systems that perform specific functions. Automation continues to grow because it is widely recognized as a principal means of eliminating errors; improving quality; and reducing labor, costs, and turn-around time.

NONANALYTICAL AUTOMATION FOR THE CLINICAL LABORATORY

Significant progress has been made in automating pre- and postanalytical activities and integrating these operations with analytical systems. Preanalytical automated systems range from robotics that perform a single function, such as aliquoting or sorting (also called *task-targeted automation systems*), to more complex **workstations** that perform several preanalytical functions. Several options are also available as modules for TLA systems, which are discussed later in this section on nonanalytical automation.

Preanalytical automation begins with automation of operations in the specimen processing area where specimens are (1) identified, (2) labeled, (3) scheduled for analysis, (4) centrifuged, and (5) sorted. After specimens are processed, they are transported to appropriate workstations in the laboratory, either manually or using conveyors, where they can then be analyzed with minimal or no human intervention. Rules-based expert system software (1) assists with the review of laboratory results by automatically releasing results that have no associated problems and (2) identifies any problematic results to bring to the attention of trained medical technologists. All specimens are catalogued after analysis and stored in a central storage facility, which may include automated storage and retrieval functions. The following sections cover various aspects of nonanalytical processes with descriptions of automation for these activities.

Specimen Identification

Typically, the identifying link (identifier) between the patient and specimen is noted at the patient's bedside. Maintenance of this connection is essential throughout (1) transport of the specimen to the laboratory, (2) subsequent specimen analysis, and (3) preparation of a report. Several technologies are available for automatic identification and data collection purposes (Box 16.1). In practice, automatic identification includes only technologies that electronically detect a unique characteristic or unique data string associated with a physical object. For example, identifiers such as (1) serial number, (2) part number, (3) color, (4) manufacturer, (5) patient name, (6) medical record number, and (7) accession number have been used to identify an object or patient through the use of electronic data

BOX 16.1 Technologies Used for Automatic Identification and Data Collection

Bar coding
Optical character recognition
Magnetic stripe and magnetic ink character recognition
Voice identification
Radiofrequency identification
Touch screens
Light pens
Hand print tablets
Optical mark readers
Smart cards

processing. In a clinical laboratory, labeling with a bar code has become the technology of choice for purposes of automatic identification.

Labeling

In many laboratory information systems (LISs), electronic entry of a test order in the laboratory or at a nursing station for a uniquely identified patient generates a specimen label bearing a unique laboratory accession number. A record is established that remains incomplete until a result (or a set of results) is entered into the computer against the accession number. The unique label is affixed to the specimen collection container when the specimen is obtained and the patient is properly identified. This may be at the bedside using preprinted labels carried from the laboratory or using a portable label printer connected wirelessly through the network to the laboratory computer. Many companies now provide these portable bar code labeling systems, and many hospitals are now using them. Proper alignment of the label on a specimen tube is critical for subsequent specimen processing when barcoded labels are used. The Clinical and Laboratory Standards Institute (CLSI) published a standard in 2011 *(Specimen Labels: Content and Location, Fonts, and Label Orientation)* that was intended to standardize specimen labels in clinical laboratories to reduce errors such as patient misidentifications. The focus of this standard was not the bar code but rather the human-readable content of the label, particularly the location and appearance of the patient name.

After the specimen is labeled at the patient's bedside and patient identification and collection information is provided, the labeled specimen is submitted to the laboratory. It may be accompanied by a requisition form, although wireless systems obviate that need and additionally automate the recording of phlebotomist identification (ID) and the date and time of collection. Arrival of the specimen in the laboratory is recorded by a manual or computerized log-in or accessioning procedure in which it is assigned an accession number.

After accessioning, specimens undergo various technical handling processes. Some automated analyzers sample directly from the original collection tube, simultaneously reading the accession number from the bar code label on the tube; other analyzers require an aliquot of the serum or plasma from the original tube. For processes requiring physical removal of serum or plasma from the original tube into a secondary tube, secondary labels bearing essential information from the original label must be affixed to any secondary tubes created. Secondary bar code labels, if necessary, may be generated at the time of accessioning or at the time the aliquots are prepared. This aliquoting step has been automated with different systems.

Bar Coding

A major advance in the automation of specimen identification in clinical laboratories is the incorporation of bar coding technology in analytical systems. A bar coding system consists of a bar code printer and a bar code reader, or scanner. In clinical laboratories, one-dimensional or linear bar code systems have historically been used. CLSI published a standard (AUTO02-A) in 2000 and updated it (AUTO02-A2) in 2003, specifying that clinical laboratories should only use the Code 128B symbology on specimen labels. However, as this chapter was in preparation, CLSI was developing a new standard (AUTO14) for the use of two-dimensional bar codes in clinical and anatomic pathology laboratories. Two-dimensional bar codes offer the prospect of encoding more data in smaller formats and having greater reliability than one-dimensional bar codes.

In practice, a bar code label (often generated by the LIS and bearing the sample accession number) is placed onto the specimen container and is subsequently "read" by one or more bar code readers placed at key positions in the analytical sequence. Examples of these key positions are a handheld bar code reader in a laboratory section or built-in bar code readers in a task-targeted automation system, the loading unit for a TLA system, and an automated analyzer. The resultant identifying and ancillary information is then transferred to and processed by the system software. Initiating bar code identification at a patient's bedside ensures greater integrity of the specimen's identity throughout these processes.

Unequivocal positive identification of each specimen is achieved in analyzers with bar code readers. Advantages of the use of bar code labels include the following:

1. Elimination of work lists for the system
2. Avoidance of mistakes made in the placement of tubes in the analyzer or during sampling
3. Avoidance of the need for analysis of specimens in a defined sequence
4. Decrease in identification errors

A best practice to reduce errors caused by bar code misreads is to set the bar code readers on analyzers and automation systems to only read the symbology expected on the specimen labels (e.g., Code 128B). This prevents reading errors in the event that a second bar code of a different symbology might be visible on an underlying label.

Identification Errors

Many opportunities arise for the mismatch of specimens and results. The risks begin at the bedside and are compounded with each processing step a specimen undergoes between

collection from the patient and analysis by the instrument. The risks are particularly great when hand transcription is invoked for accessioning, labeling and relabeling, and creation of load lists. An incorrect accession number, one in which the digits are transposed, or a load list with transposed accession numbers may cause test results to be attributed to the wrong patient. An additional hazard exists when specimens must be inserted into certain positions in the loading zone defined by a load list. Human misreading of the specimen label or the loading list may cause misplacement of specimens, calibrators, or controls. Automatic reading of bar code labels reduces the error rate from 1 in 300 characters (for human entry) to about 1 in 1 million characters barring (1) minor imperfections in printed bar codes, (2) improper bar code scanner resolution, or (3) skewed orientation of bar code labels on containers, all of which can result in read errors.

Specimen Preparation

The clotting of blood in specimen collection tubes, their subsequent centrifugation, and the transfer of serum to secondary tubes require a finite time to complete. If performed manually, the process may result in a delay in the preparation of specimens for analysis. To eliminate the problems associated with specimen preparation, systems are being developed to automate this process.

Use of Whole Blood for Analysis

When whole blood is used in an assay system, specimen preparation time is essentially eliminated. Automated or semiautomated ion-selective electrodes, which measure ion activity in whole blood rather than ion concentration, have been incorporated into automated systems to provide certain test results within minutes of collecting a specimen. This approach now is used commonly for assays of electrolytes and some other common analytes. Another approach involves manual or automated application of whole blood to dry reagent films and visual or instrumental observation of a quantitative change.

Specimen Delivery

Automated methods are often used to deliver specimens to the laboratory instead of more manual methods (e.g., phlebotomist transport, courier service, nurse delivery). Pneumatic tube systems have been used in many hospitals, but mobile robots can also transport specimens and other laboratory supplies within a laboratory or to the laboratory from other locations in the facility.

Pneumatic Tube Systems

Pneumatic tube systems provide rapid specimen transport and are reliable when installed as point-to-point services. However, when switching mechanisms are introduced to allow carriers (the bullet-shaped containers used to hold specimens) to be sent to various locations, mechanical problems may occur and may cause misrouting. In addition, close attention to the design of the pneumatic tube system is necessary to prevent hemolysis of the specimen. Avoidance of sudden accelerations and decelerations and the use of proper packing material inside the carriers can help to minimize hemolysis.

Mobile Robots

Automated guided vehicles (AGVs), also called mobile robots, have been used successfully to transport laboratory specimens both within a laboratory and outside a central laboratory. They are easily adapted to carry various sizes and shapes of specimen containers and are reprogrammable with changes in laboratory geometry. In addition, in a busy laboratory setting, delivery of specimens to laboratory benches by a mobile robot can be more frequent than human pickup and is cost-effective. Inexpensive models follow a line on the floor, but others have more sophisticated guidance systems. Their limitations include the need to batch specimens for greater efficiency and, in most cases, the requirement for laboratory personnel to place specimens onto or remove specimens from the mobile robot at each stopping place.

Single-Function Workstations

Some laboratories have gained efficiency, improved quality, and lowered turn-around time by implementing an automated system that performs a single task. This automation is also referred to as *task-targeted automation*. Examples of such automation include automated centrifuges, decappers, recappers, aliquotters, and sorters. Such systems began to be developed in the 1990s and early 2000s, when TLA systems were relatively expensive and thought to be affordable only by reference laboratories and the largest hospital laboratories. Today these systems continue to be developed and implemented because they offer affordable solutions for specific customer needs. Box 16.2 lists a number of vendors offering one or more single-function workstations. Fig. 16.1 is an example of a sorter, one of the different types of single-function workstations that are available. However, a laboratory could use a sorter for both preanalytical sorting for different testing areas and postanalytical sorting for archival storage.

Multifunction Workstations (Automated Specimen Processing)

Although the manual operations carried out in a specimen processing area may look simple, considerable complexity underlies them. Consequently, specimen processing has been one of the most difficult areas of the clinical laboratory

BOX 16.2 Vendors of Single-Function Automation Systems

Hettich: http://www.hettweb.com
IDS Co., Ltd.: http://www.idsma.com
m-u-t automation; see also *Sarstedt:* https://www.mut-gmbh.de/en/laboratory-automation/
Roche Diagnostics: http://www.usdiagnostics.roche.com/en/core_laboratory.html#Automation
Sarstedt: http://www.sarstedt.com
Yaskawa Motoman: http://www.motoman.com

to automate. Each specimen passing through a processing area has to undergo a series of operations, including (1) receiving the specimen; (2) inspecting it for appropriateness (labeling, container type, temperature, and quantity of specimen); (3) logging the specimens into the LIS; (4) recording the date and time of collection and phlebotomist ID, if not already entered; (5) labeling the specimen with an accession number; and (6) separating urgent and stat specimens from routine specimens. Also, specimens may have to be sorted for centrifugation and aliquoting and may have to be sorted or otherwise prepared for the appropriate laboratory station.

An example of a standalone specimen processing system is shown in Fig. 16.2. Such systems place processed specimens into racks that must be transported manually to the testing areas, with some exceptions. Some of these systems are approximately the size of a large automated analyzer; others may be slightly larger. They may be a good choice for laboratories (1) with daily workloads of 500 to 1500 specimens, (2) with space limitations, or (3) that desire ease of use with different analyzers from different vendors. Some laboratories may choose to use multiples of such a system based on higher daily volumes or to automate postanalytical archiving in addition to preanalytical specimen processing.

Fig. 16.1 Example of a single-function nonanalytical automated workstation. Pictured is a Motoman AutoSorter 1200, which has sorting speeds of up to 2000 specimens per hour, several different options for deck configurations, and the possibility of different analyzer-specific output racks. Photo shows two units side by side. (Courtesy Yaskawa America, Motoman Robotics Division, http://www.motoman.com/labauto.)

Fig. 16.2 Example of a multifunction nonanalytical automated workstation. Pictured is a Roche Cobas 612 system with automated centrifugation, decapping and recapping, aliquoting, and bar code labeling of secondary tubes, including sample quality inspection. (Courtesy Roche Diagnostics, http://www.cobas.com.)

Box 16.3 lists vendors offering multifunction or specimen processing workstations. Although there may be some variation in the functions that are included, these systems will typically (1) receive incoming specimens, (2) sort, (3) decap, (4) aliquot, and (5) label aliquot specimen containers with bar codes. All may be interfaced to the laboratory's LIS. Some systems may also include automated centrifugation. Some of the systems sort into instrument-specific racks for analyzers from a number of different vendors. In addition, some users apply these systems to aliquot and sort reference or "send-out" testing, saving considerable time in locating the original specimens after testing in their own laboratory.

Total Laboratory Automation Systems

Several manufacturers offer integrated or modular automation systems that use conveyor belts to connect preanalytical specimen processing and other functions directly to analyzers. Such systems may also include functions such as postanalytical storage or sorting of specimens or aliquots for referral testing or to be transported to low-volume testing areas. A list of automation vendors offering TLA systems can be found in Box 16.4; Fig. 16.3 illustrates a large TLA system with a variety of connected nonanalytical and analytical modules.

In addition to the nonanalytical functions listed in the preceding section, TLA systems typically add (1) conveyor transport; (2) interfacing to automated analyzers; (3) more sophisticated process control; and, in some cases, (4) specimen storage and retrieval. All systems are of modular design, allowing the customer to choose the modules or features that should be included. Some systems use an open design, which permits

BOX 16.3 Vendors of Multifunction Automation Systems (Pre- and Postanalytical)

Abbott Diagnostics: http://www.abbottdiagnostics.com
Aim Lab Automation Technologies: http://www.aimlab.com
Beckman Coulter: http://www.beckmancoulter.com
IDS Co., Ltd.: http://www.idsma.com
Roche Diagnostics: http://www.usdiagnostics.roche.com/en/core_laboratory.html#Automation
Sarstedt: http://www.sarstedt.com
Tecan: http://diagnostics.tecan.com/products/liquid_handling_and_robotics/fe500pro
Yaskawa Motoman: http://www.motoman.com

BOX 16.4 Vendors of Total Laboratory Automation Systems

Abbott Diagnostics: http://www.abbottdiagnostics.com
Beckman Coulter: http://www.beckmancoulter.com
Cerner Corporation: http://www.cerner.com/solutions/hospitals_and_health_systems/laboratory/cerner_automation
IDS Co., Ltd.: http://www.idsma.com
Inpeco: http://www.inpeco.com
Ortho-Clinical Diagnostics: http://www.orthoclinical.com
Roche Diagnostics: http://www.roche-diagnostics.us
Siemens Healthcare Diagnostics: http://www.usa.siemens.com/diagnostics
Sysmex America: http://www.sysmex.com/us
Yaskawa America, Motoman Robotics Division: http://www.motoman.com/labauto

Fig. 16.3 Example of a total laboratory automation system. Pictured is a Siemens Aptio Automation System, which can include options such as a hematology analyzer with autoslide maker or stainer, hemostasis analyzer, clinical chemistry and immunoassay systems, integrated chemistry systems, and data management system laboratory information systems (workstation and the following pre- and postanalytical modules: tube inspection, centrifuge, refrigerated storage, aliquotter, bulk input, rack input, input and output, sealer, desealer, decapper, and aliquotter recapper. (Courtesy of Siemens Healthcare. Copyright Siemens Healthcare Diagnostics, Inc., 2016. http://siemens.com/healthcare.)

interfaces to analyzers from a variety of vendors, but other systems are of a closed design and are interfaced only to the vendor's own analyzers. To achieve maximum effectiveness of an automation system, process control software should be able to read the specimen's ID bar code and obtain information from the LIS about specimen type and ordered tests. The software should then determine the processes the specimen requires and the exact route or course of action for each specimen. It should be able to (1) calculate the number of aliquots and the proper volume for each specimen depending on the tests requested, (2) route the specimens to preanalytical processing equipment and then to analyzers, (3) recap the specimens after testing, and (4) retain the specimens for automatic recall. The software should be able to monitor analyzers for in-control production status and automatically make decisions if a test is not available. Specimen integrity checking should be automatic; rules-based decisions should monitor specimen quality and make these decisions as appropriate. Finally, most process control software should include "autoverification," which is validation of analyzer results by making rules-based decisions that flag exceptions for technologist review, and "autoretrieval" of specimens for repeat, reflex, and dilution testing. Although most of these systems are restricted to handling specific types of specimen containers, they are capable of processing much of the daily workload of a large clinical laboratory. A few laboratories with workloads of less than 1000 tubes per day have justified these systems because of a shortage of technical help, but typically these systems are designed for laboratories handling 1000 to 25,000 specimens per day. In addition to process control software and an LIS interface, each of these systems incorporates some or all of the following components:

1. *Specimen input area:* This is a holding area where bar code–labeled specimens are introduced into the system.
2. *Bar code reading stations:* Multiple bar code readers are placed at critical locations in the processing system to track specimens and provide information for their proper routing to various stations. Alternately, bar code readers may only be required at limited locations in the system if each transport carrier (puck) has a unique ID such as a radiofrequency identification (RFID) chip, and the specimen's bar code ID is "married" to the puck's ID when first placed in the puck.
3. *Transport system:* Segments of a conveyor belt line that move specimens to the appropriate location.
4. *A high-level device to sort or route specimens:* A device that separates specimens based on processing requirements or order code and passes them to the transport system or to a system using racks. A high-level sorter is often used to separate specimens that require centrifugation or other processing steps from specimens that do not, or to route specimens into completely different pathways within the total automation system.
5. *Automated centrifuge:* An area of the specimen processor in which specimens requiring centrifugation are removed from the conveyor belt, introduced into a centrifuge with automatic balancing, centrifuged (refrigerated or at room temperature), and then placed back on the transport system.
6. *Level detection and evaluation of specimen adequacy (specimen integrity):* An area where sensors are used to evaluate the volume of specimen in each specimen container and to look for the presence of hemolysis, lipemia, or icterus. Very few TLA systems have such capability outside of automated chemistry analyzers.
7. *Decapper station:* An area or device in the automated system where specimen caps or stoppers are automatically removed and discarded into a waste container.
8. *Recapper station:* An area or device in the automated system where specimen tubes are automatically recapped with new screw caps, stoppers, or other air-tight closures.
9. *Aliquotter:* A machine that aspirates appropriately sized aliquots from each original specimen container and places them into bar-coded secondary specimen containers for sorting and transport to multiple analytical workstations. The process controller generally instructs the aliquotter as to how many aliquots of what volumes are required based on the bar code ID of the specimen.
10. *Interface to automated analyzer:* A direct physical connection to an automated analyzer that permits the analyzer's sampling probe to aspirate directly from an open specimen container while the container is still on the conveyor or that may robotically lift the container from the conveyor and place it onto the analyzer.
11. *Sorter:* An automated sorter to sort specimens not going to a conveyor-interfaced analyzer or workstation. Such a sorter typically sorts into 30 to 100 different sort groups in racks or carriers. In some systems, the racks may be specific to certain analyzers for convenience. The sorter may also serve as a temporary holding area for specimens that may require a repeat test (see take-out stations) or may sort specimens for referral testing.
12. *Take-out stations:* Temporary storage areas for specimens before or after analysis. The take-out station may be the same as the sorter described earlier, where specimens are sorted for manual delivery. However, it may also serve as a holding area (stockyard) for specimens awaiting autoverification of results if a repeat test is required.
13. *Storage and retrieval system:* This unit may serve the same function as the take-out station or stockyard—that of holding specimens after analysis in case a specimen is necessary for a repeat test—but it has one major difference. These units are typically refrigerated and hold many more specimens (3000 to 30,000) than the typical take-out station or stockyard. Depending on daily workloads, the laboratory may be able to retain up to 1 week's worth of specimens for possible repeat or additional tests. Specimen containers are loaded and retrieved with a robot. Some storage systems include automated specimen discard directly into biohazard waste bins after a predetermined storage duration has been reached.

Conveyor Belts

One main feature of integrated TLA systems is the use of conveyor belts. Ordinary industrial conveyor belts have been used successfully when only transportation is required.

However, in the past, when conveyors were integrated with other robotic systems to automate pre- or postanalytical functions, this technology had difficulty handling the large variety of specimen containers found in the clinical laboratory. To increase the variety of specimen containers that are carried on a conveyor belt system, specimens are placed into specially designed carriers that fit on the conveyor belt line. Sometimes known as "pucks" or "racks" (depending on whether they carry individual specimens or groups of specimens), the carriers have receptacles for variously sized tubes, generally ranging from 13 × 75 mm to 16 × 100 mm. The industrial, motor-driven conveyor belts used in typical laboratory automation systems have transport rates of a few hundred pucks per hour up to 2000 per hour.

Two types of conveyor belt systems are typically used as illustrated in Fig. 16.4. Fig. 16.4A depicts a loop conveyor that has a single module for both input of new specimens and removal of completed specimens. The flow around the loop takes specimens to the processing and analytical modules. Specimens may be sampled directly by the analytical instrument while on the conveyor, or a robot attached to the workstation may remove selected specimens from the conveyor for analysis. Fig. 16.4B depicts a unidirectional conveyor in which specimens are loaded at one end of the belt, flow past various processing and analytical modules, and arrive at the opposite end where they are removed. Depending on the vendor, both of these conveyor types can have bypass lanes that enable specimens not designated to stop at certain modules to bypass them entirely and proceed to other modules. The advantage of this approach is that it does not require specimens to be aliquoted because specimens pass by all workstations where tests are performed. If specimens are robotically removed from their carriers on the line for testing, systems are required that queue empty carriers to return the tubes to the conveyor and to identify which specimen is in each carrier.

Transfer of specimens from the conveyor belt to the laboratory workstation has been implemented in various ways. For example, some manufacturers have equipped their laboratory instruments with devices to move specimen containers from conveyor belt systems onto the analyzers, and others have installed the ability for the analyzer to sample directly from specimen containers remaining on the conveyor. In practice, the automation system requires a device that stops the tube in the exact location required by the analyzer for sample transfer and verifies and transfers the tube's bar code identification to the analyzer. Depending on the particular automation system, specimens for immediate short turn around testing (STAT) may or may not be processed and analyzed at speeds that are faster than if manually handled. Laboratories that are

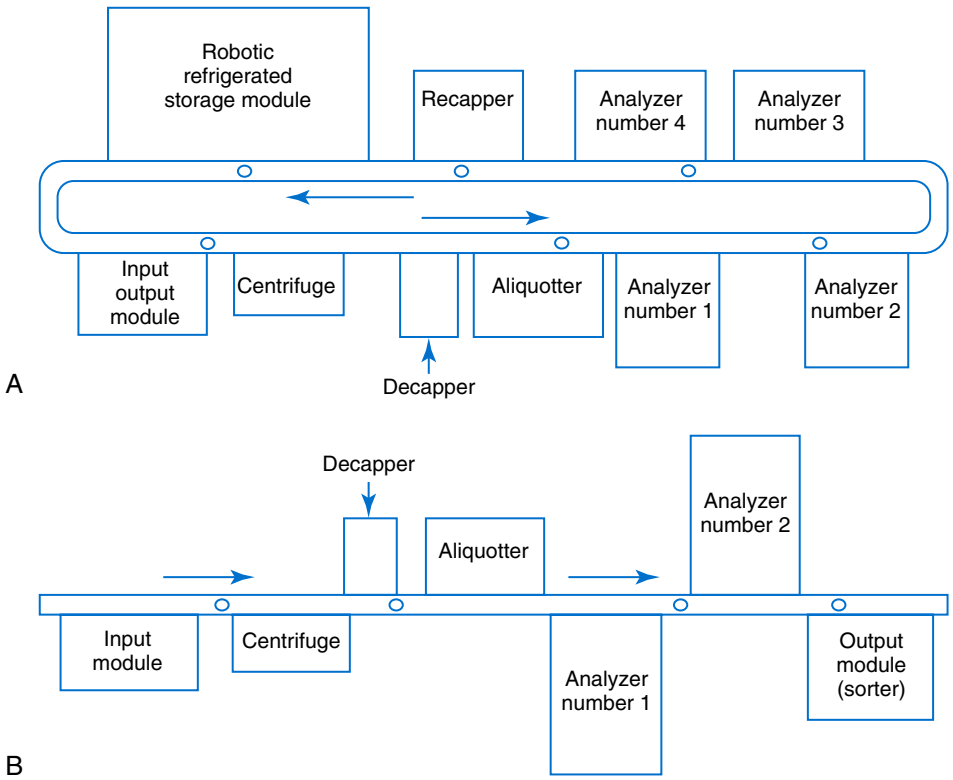


Fig. 16.4 (A) A depiction of a laboratory automation system using a continuous loop–type conveyor. In this concept, a single module is used for loading of new specimens and unloading of completed specimens. Depending on the needs of the laboratory, all other modules illustrated are options that can be included or not. (B) A depiction of a laboratory automation system using a unidirectional conveyor. In this concept, separate modules are required for loading of new specimens and unloading of completed specimens. The output module may also perform sorting of specimens for other testing. Other modules can be included, depending on the laboratory's needs.

purchasing laboratory automation systems are advised to learn how each system being considered handles STAT specimens.

Automated Specimen Storage and Retrieval

The automated capability to store and retrieve specimens on demand and with readily known exact storage locations is an important aspect of automated specimen delivery systems. In addition to automated storage and retrieval options in some of the integrated systems described earlier, several automated or semiautomated options for storage and retrieval, as well as LIS modules and PC-based software systems, permit laboratories to track trays or racks of specimens in their own freezers or refrigerators. Some large reference laboratories have adapted large storage systems commonly used in other industries to their laboratory settings.

PRACTICAL CONSIDERATIONS

In this section, the practical considerations that influence a laboratory's decision to automate part or all of its operations are discussed. Some of the suggested readings provide more insights and details about these practical considerations, including workflow mapping.

Evaluation of Requirements

Any consideration of total or modular laboratory automation should start with an evaluation of requirements. Such an evaluation begins with mapping of the current laboratory workflow from the arrival of patient specimens through completion of testing and reporting of results. Box 16.5 lists potential workflow steps that should be mapped. Mapping of material (specimen) flows and data flows is directly related to process flow and assists the laboratory in determining process steps that (1) are bottlenecks, (2) waste labor, and (3) are prone to error. Workflow mapping thus enables the laboratory to better identify what steps should be considered for automation.

Some laboratorians use an "80% rule" in guiding decisions about automation. Clinical laboratories have many exceptional tests, specimen containers, and handling situations. Nevertheless, if 80% of the daily workload of specimen containers, covering most routine handling situations, can be standardized and automated, the laboratory should achieve a significant reduction in its labor usage and costs, sufficient to justify the investment in automation and the planning and evaluation time involved.

After the laboratory's existing workflow is mapped and its requirements are identified, alternative solutions are considered. Use of process improvement techniques such as LEAN and Six Sigma can help to ensure that automated solutions ultimately improve operations. Vendors are invited to make presentations and to host visits of the laboratory management team to other laboratories where vendors have successful installations. It is important at this stage to focus on the requirements identified by workflow mapping and not to allow the vendor to try to sell equipment that may not be necessary. Key points to remember when a laboratory is considering either new automation or an addition to existing automation are listed in the following Points to Remember box.

BOX 16.5 Clinical Laboratory Steps for Workflow Mapping

Unpacking from transport containers
Presorting
Temperature preservation
Order entry
Document management (e.g., requisitions)
Labeling
Sorting
Centrifugation
Labeling of aliquot tubes
Decapping
Pouring of aliquots
More sorting
Delivery to laboratory sections
More sorting
Preparing work lists
Labeling analyzer-specific tubes for specimens
Pouring or pipetting analyzer-specific specimens
Loading tubes on analyzers
Performing tests (steps such as extraction, centrifugation, precipitation, and dilution are not specifically listed)
Unloading analyzers
Recapping
Data manipulations (calculations)
Results review and verification
Reporting of results
Delivery of specimens to archival storage system
Archival storage of specimens
Reflexive testing
Repeat testing, diluting, if necessary
Additional physician-ordered testing
Specimen retrieval for additional or repeat testing
Disposal of expired specimens

POINTS TO REMEMBER

When considering new automation or expansion of existing automation:

- Understand the laboratory's needs (moving from the current state to the desired state).
- Review the laboratory's specimen volumes by hour and by day (determine peaks and troughs).
- Review the percentages that are centrifuged, aliquoted, refrigerated or frozen, and shared among lab sections.
- Consider logistics, handling, facilities, and space issues.
- Map the laboratory's workflow. (What are the paths for all specimens?)
- Time the workflow to find bottlenecks and time wasters.
- Rank the laboratory's needs in order of importance and differentiate needs that are "must haves" from needs that are "desired."
- Identify possible solutions that will meet the identified needs.
- Evaluate alternatives that will meet those needs.
- Use performance measures for productivity, turn-around time, and quality to establish baselines against which postautomation performance can be measured.
- Cost justification will likely be required by management, but the performance measures, if credibly predicted, will likely provide the basis for justification.

Problems of Integration

Building a highly integrated laboratory generates many potential problems. Most vendors of clinical laboratory automation systems prefer customer settings in which the integrated analyzers are their own brand. However, many laboratories may prefer to use analyzers from multiple vendors, including the automation system vendor, making integration of instruments and robotic devices from different manufacturers necessary. Decisions must be made concerning which device will be the master controller and which vendor will develop the software that provides overall control of the automation scheme. In addition, individuals or firms that will be responsible for configuration of the automation to the geometry and production schedule of the laboratory must be recruited and trained. Over the past decade, automation vendors have gained considerable experience implementing integrated systems. In addition to expertise provided by the automation vendor, readers should find the CLSI automation standards helpful when integrating devices in their automation model.

Device Integration

One objective in developing an integrated laboratory is to link laboratory instruments and devices into an automated system to maximize the number of automated functions. Automatic specimen introduction requires the development of mechanical interfaces between each laboratory analyzer and devices such as conveyor belts, mobile robots, or robot arms. Some systems have added enhancements, such as electronic interfaces for laboratory instruments, to allow remote computer control of front-panel functions, notification of instrument status information, and coordination of the distribution of specimens between instruments. In the ideal integrated clinical laboratory automation system, the process controller would integrate, automate, and monitor many of the decision-making tasks that occur in the daily activity of a laboratory. For example, the process controller would notify an operator that a particular analyzer (even if it is from a different vendor) may be running low on a reagent or may have a fault condition. The process controller would control and schedule nonanalytic modules such as centrifuges, aliquotters, decappers, and so forth. Most existing LISs have no process control capability, and their interfaces with laboratory analyzers provide only the ability to download accession numbers and the tests requested on each specimen and to upload results generated by the analyzer. The distribution of tasks must be carefully specified in developing such a communications network.

AUTOMATION OF THE ANALYTICAL PROCESSES

The following sections describe basic concepts inherent to automating the analytical process, with a focus on how these concepts have been integrated into clinical chemistry and immunochemistry instrumentation and automated solutions. Advances in analytical automation in clinical chemistry, immunochemistry and immunology, urinalysis, hematology, coagulation, transfusion medicine, microbiology, mass spectrometry (MS), and molecular diagnostics are then discussed separately.

Basic Concepts

The individual steps required to complete an analysis are referred to as *unit operations* (Box 16.6). Many of these operations (e.g., specimen identification, delivery, and processing) are shared with capabilities of nonanalytical automation discussed earlier in this chapter, although they may also be integrated into the analytical instrumentation itself.

Automated analyzers are generally designed to perform mechanized versions of previously manual laboratory techniques and procedures. Analytical automation provides opportunities for efficiency, reproducibility, and throughput that may not be possible with manual processes. As such, modern instrumentation is engineered in a wide variety of configurations that are influenced by the manufacturer's hardware design, the testing to be performed, the environment in which the instrumentation is to be used, and the instrument and LIS connectivity that are ultimately required.

The most common automated instrument configuration is the random-access analyzer. In **random-access analysis**, testing is performed sequentially (or prioritized, i.e., STAT) on a set of specimens, with each specimen analyzed for a different selection of tests based on its respective clinical orders. Tests in random-access analysis are performed through the use of different vials, packs, or kits of reagents stored onboard the analyzer. This approach permits measurement of both a variable number and variety of analytes in each specimen loaded onto the instrument.

Other analyzer configurations include continuous-flow and centrifugal analyzers. *Continuous-flow analyzers* (e.g., the Technicon AutoAnalyzer) were the first automated analyzers used in clinical laboratories. Initially, these analyzers were used in a **single-channel analysis** configuration and carried out a **sequential analysis** of each specimen. Subsequently, **multiple-channel analysis** versions were developed, in which analysis of each specimen was performed on every channel in parallel. Results of nonrequested tests in the test profile were discarded as unnecessary after the analysis was complete. Inflexibility in the menu of tests that could be performed on these analyzers eventually led

BOX 16.6 Unit Operations in an Analytical Process

Specimen identification
Specimen delivery
Specimen processing
Sample introduction and internal transport
Sample loading and aspiration
Reagent handling and storage
Reagent delivery
Chemical reaction phase
Measurement phase
Signal processing, data handling, and process control

to their replacement in the marketplace by more versatile discrete processing configurations.

Centrifugal analyzers used discrete pipetting to load aliquots of specimens and reagents sequentially into discrete chambers in a rotor. The specimens were subsequently analyzed in parallel *(parallel analysis)* by spinning the rotor to exert centrifugal force to mix the specimens and reagent and to drive the mixtures into cuvettes located on the periphery of the rotor. Such analyzers could be operated in a multiple-specimen/single-chemistry or a single-specimen/multiple-chemistry mode. Centrifugal analysis was not sufficiently versatile to compete with other random-access analyzers and has largely been abandoned for use in high-volume routine laboratory testing. Centrifugal microfluidics, however, is an important component in near-patient testing in which whole-blood specimen processing on small instrumentation enables plasma separation and multiple analyses in separate micro wells.

Specimen Loading

The loading zone of an analyzer is the area where specimens are held in the instrument before they are analyzed. This holding area is often a circular tray, a rack or series of racks built into a cassette, or an array of individual containers ("pucks") into which tubes are inserted. Although most automated instrumentation is now able to identify tubes through techniques such as bar code scanning, some instruments may still require loading of specimens in a defined sequence as specified by a loading list. In most large-scale automated analyzers, specimens may be added continuously by the operator as they become available. This configuration is also most compatible with modular automation and TLA solutions, in which specimens are automatically distributed to component instruments.

A desirable feature of any automated analyzer is the ability to insert new specimens ahead of specimens already in place in the loading zone. This feature allows the timely analysis of a specimen with a high medical priority (i.e., STAT specimens). When specimen identification is machine read, it is usually possible for the operator to reposition specimens in the loading zone. When specimen identification is tied to a loading list, however, insertion or repositioning of specimens must be accompanied by revision of the loading list. The ability to perform add-on testing when a specimen is present on the instrument is another characteristic that can improve efficiency and decrease manual interventions.

In clinical chemistry testing, the specimen for automatic analysis is often serum or plasma. Most automated chemistry analyzers are able to directly aspirate specimen from primary collection tubes of common container sizes. Instrument manufacturers typically specify acceptable specimen container types to ensure physical compatibility of the tube and to prevent inadvertent damage to the instrument. Although the outer tube diameter is most frequently evaluated for instrument compatibility, the inner tube diameter (e.g., double-walled collection tubes) may also impact compatibility with instrument probes or pipettes. The inner tube diameter may also impact accurate estimations of sample volume in systems with automated volume detection. Many primary collection tubes contain gel separator materials (i.e., separator tubes) that form a barrier after centrifugation between serum or plasma and the cellular elements. Sample aspiration in automated analytical systems must be designed to prevent or withstand probe crashes with the bottom of the tube, as well as crashes or plugging due to gel separators in the specimen container.

Many analyzers also sample from secondary containers (cups or aliquot tubes) filled with sample transferred from an original specimen tube. Often the design of the sampling cup is unique for a particular analyzer. Each cup should be designed to minimize *dead volume*, that is, the excess serum that must be present in a cup to permit aspiration of the full volume required for testing. Many instruments use *short-sample cups* or *false-bottom tubes* to help facilitate testing by automated analyzers with minimal dead volume. When the creation and placement of aliquots or pour-offs into secondary containers is not automated, careful attention must be used in preventing specimen mix-ups, particularly when the secondary container is too small to be directly labeled with appropriate patient identifiers. Secondary containers should be disposable to minimize cost and the potential for cross-contamination, and their shape should, even without a cap, limit evaporation by minimizing the surface area of sample exposed to air.

Specimens may undergo other forms of degradation in addition to evaporation. For example, specimens that contain thermolabile constituents may undergo degradation if held at ambient temperatures. Thermolability is minimized when both specimens and calibrators are held in a refrigerated loading zone. Other constituents such as bilirubin are photolabile. Photodegradation is reduced by the use of semiopaque cups and placement of smoke- or orange-colored plastic covers over the specimen cups. The potential effects of both thermolability and photolability, however, can be minimized by automated instrumentation that tests specimens promptly and efficiently.

Specimen Sampling

The instrument sampling mechanism aspirates specimen from a loaded container for subsequent testing. Automated analyzers may sample fixed or variable volumes depending on the instrument and assay configuration. Sampling takes place using fixed reusable probes or disposable pipettes. Fixed probes limit disposable costs, reduce the need for supply management, and eliminate ongoing operator loading of disposable pipettes. Onboard wash stations for fixed probes are required, however, to reduce the risk of sample or reagent carryover.

Many automated analyzers incorporate *liquid-level sensing* to detect potential short samples, ensure appropriate specimen aspiration, and minimize the risk of probe crashes. Liquid-level detection is often conducted using either pressure or capacitive (radiofrequency) methods. Disposable capacitive pipette tips are available for processes that are vulnerable to carryover. Automated analyzers may

also integrate methods for *clot detection* because clots may result in inadequate specimen volumes being aspirated, as well as plugging of the probe, tubing, or fluidic channels. Instrument "clot errors" that occur in specimens without evidence of detectable clots should raise the suspicion for potential hyperviscosity.

Closed-container sampling systems have also been developed for use in automated chemistry and hematology analyzers. In these systems, the specimen probe passes through a hollow needle that initially penetrates the primary container's rubber stopper. This configuration prevents damage or plugging of the specimen probe while allowing the level sensor to remain active. After the specimen probe is withdrawn, the outer hollow needle is also withdrawn, so the stopper reseals, and no specimen escapes. Closed-container sampling is used widely in hematology analyzers because specimen mixing through gentle tube inversion is commonly required before analysis. Such mixing may even be incorporated into instrument functionality. Closed-container sampling also reduces the risk of contamination or splatter associated with removing the specimen container stopper.

Transmission of infectious diseases by automated equipment is a concern in clinical laboratories. The potential method of transmission by automated equipment is primarily through splatter of serum or blood during the acquisition of samples from rapidly moving specimen probes. The use of level sensors, which restrict the penetration of sample probes into specimens, and provision of software for smoother motion control greatly reduce the possibility of splatter. Additional engineering controls include closed-tube sampling, covers over the specimen probe sampling areas, automated discard to waste bins for disposable sample probes, specimen cups, reaction cuvettes, direct-line plumbing of instrument liquid waste, and appropriate decontamination protocols before routine maintenance or service.

Specimen Introduction and Internal Transport

The method used to introduce a sample into the analyzer and its subsequent transport within the analyzer is the major difference between *continuous-flow* and *discrete processing* systems. In continuous-flow systems, the sample is aspirated through the sample probe into a stream of flowing liquid, whereby it is transported to analytical stations in the instrument. In discrete systems, the sample is aspirated into the sample probe and is then delivered, often with reagent, through the same orifice into reaction cups or other containers where the chemical reaction and analysis take place.

Continuous-Flow Systems

Peristaltic pumps and plastic tubing are used to advance the sample and reagents in **continuous-flow analysis**. Peristaltic pumps trap a "slug" of fluid between two rollers that occludes the tubing. As the rollers travel over the tubing, the trapped fluid is pushed forward and, as the leading roller lifts from the tubing, is added to the fluid beyond it. Peristaltic pumps are still used in some hematology analyzers, instruments with ion-selective electrodes, and wash systems.

Discrete Processing Systems

Air- or positive-liquid displacement pipettes are used for sampling in most discrete automated systems in which specimens, calibrators, and controls are delivered by a single pipette. A pipette may be designed for one of two operational modes: (1) to dispense only aspirated sample into the reaction receptacle or (2) to flush out sample together with diluent. Both systems use a plastic or glass syringe with a plunger, the tip of which is usually made of Teflon. Instrument pipettes may be designed as (1) fixed, (2) variable, or (3) selectable volume. Selectable-volume pipettes allow the selection of a limited number of predetermined volumes. Although pipettes with unlimited or selectable volumes are typically used in systems that allow many different applications, fixed-volume pipettes are used for samples and reagents in instruments dedicated to the performance of only a small variety of tests.

Carryover

Carryover is defined as the transfer of a quantity of analyte or reagent from one specimen or reaction into a subsequent one. Because it erroneously affects analytical results from the subsequent reaction, carryover should be minimized or eliminated. In discrete systems with disposable reaction vessels and measuring cuvettes, carryover is generally caused by the pipetting system. In instruments with reusable cuvettes or flow cells, carryover may also arise from incomplete cleaning of the cuvettes or flow cells between assays.

Most manufacturers of discrete systems reduce sample-to-sample carryover by using disposable pipette tips or by incorporating wash stations for the sample probe that flush the internal and external surfaces of the probe with copious amounts of wash solution. An adequate ratio of flush and rinse to specimen volume controls carryover in many cases to acceptable limits. Appropriate choice of sample probe material, geometry, and surface conditions also influences carryover. Some systems clean the outside of the sample probe to prevent transfer of a portion of the previous specimen into the next specimen cup. Use of disposable sample probe tips eliminates the risk of contamination of one sample by another inside the probe, as well as the risk of carryover of one specimen into a specimen in the next cup, because a new pipette tip is used for each pipetting.

Although many high-volume automated chemistry analyzers have fixed reusable probes, instrumentation associated with processes that are more vulnerable to carryover and contamination (e.g., serology and molecular diagnostics) generally incorporate disposable pipettes for sample aspiration. For example, the reduction of sample-to-sample carryover is a more stringent requirement for automated analyzers that perform immunoassays in which some analytes (e.g., human chorionic gonadotropin) have a wide range of concentrations. Some systems use extra steps, such as additional washes, or an additional washing device to reduce carryover for selected tests to acceptable limits. Because extra steps reduce overall throughput, additional rinsing functions are initiated (by computer operator selection) only for assays with large analytical measurement ranges (AMRs).

Sample-to-sample carryover can be detected by the preparation of two sample pools—one having a very high analyte concentration (H), the other having a low concentration (L). By running sequences of tests (e.g., HHHLLLHHLLHHLL) if sample-to-sample carryover is present, higher results will be noted in the low-concentration sample analyses that immediately follow a high-concentration sample.

Reagent-to-reagent carryover can also occur on discrete systems that use a common probe for pipetting reagents. Its minimization or elimination requires use of the same approaches described for sample-to-sample carryover. Detection of reagent-to-reagent carryover can be difficult for end users and usually requires involvement of the instrument vendor.

Sample Pretreatment

Analytical procedures sometimes require the capability to remove proteins or other interferents from specimens to ensure the specificity of an analytical method. Dialysis, column chromatography, extraction, and filtration have been used for this purpose. Additional pretreatment processes include specimen concentration and/or dilution to bring an analyte of interest into an assay's AMR. Automation of pretreatment processes through offline or integrated specimen processing systems can improve the efficiency of these procedures. Sample pretreatment steps may also be integrated into assays designed for automated platforms. Examples include the addition of reagents that displace analytes from their endogenous binding proteins and instrument addition of red blood cell (RBC) lysing reagents in automated glycosylated hemoglobin or HbA1c measurements.

Reagents

Reagent Management and Configuration

Reagent management is an important consideration in choosing automated instrumentation. Analytical systems may require that different reagents be stored at ambient, refrigerated (4°C), or frozen (−20°C or −80°C) temperatures. Availability of adequate reagent storage close to instrumentation is an important consideration in implementing efficient analytical systems. Reagents, calibrators, and quality control (QC) materials that are supplied as lyophilized (dry) powder requiring reconstitution can require significant ongoing manual processing on the part of instrument operators. Some analytical systems have the capacity for onboard reconstitution of certain reagents, QC, or calibrators.

Many automated systems use liquid reagents stored in plastic or glass containers. For analyzers with a working inventory maintained in the system, the volumes of reagents stored onboard are often optimized based on the desired number of tests to be performed without operator intervention. Such analyzers often track the amount of reagent that is currently onboard and notify the operators when a replacement is required. In larger systems, reagent storage compartments may be maintained at refrigerated temperatures to increase onboard stability. Some analyzers have the ability to cap reagents while not in use to decrease evaporation and reagent degradation. Larger automated analyzers may have the ability to load new reagent packs (and discard empty packs) without interrupting the testing process or placing the instrument in a "standby" mode. Additionally, many automated instruments allow the loading of more than one similar type of reagent pack so that testing is not delayed when the first pack runs empty. Having an excess number of available reagent positions on automated instrumentation allows for optimization of reagent management to match the laboratory's daily workflow with minimal operator involvement. For analyzers in which specimens are not processed continuously, in-use reagents are often stored in laboratory refrigerators and are only introduced onto instruments when required for testing.

Reagent Identification

Labels on reagent containers include information such as (1) reagent identification, (2) volume of the contents or the number of tests for which the contents of the containers are to be used, (3) expiration date, and (4) lot number. Many reagent containers carry bar codes or RFID tags that contain some or all of this information. Advantages of using reagent bar codes include (1) facilitation of inventory management, (2) the ability to insert reagent containers in random sequence, and (3) the ability of the system to track in-use reagent lot numbers and document that appropriate calibration and QC procedures were performed on a specific lot or pack before testing. Furthermore, when a bar code reader is coupled with a level-sensing system on the reagent probe, it alerts the operator as to whether a sufficient quantity of reagent exists to complete a workload. Bar codes on reagent containers may also contain information about calibrators, such as the definition of a calibration curve algorithm and values of curve constants defined at the time of reagent manufacture. Accompanying calibrator materials provided in their own bar code tubes at the time of manufacture ensure that calibration functions are integrated properly into the analysis. Inventory management systems (tied to bar codes or RFID tags) can be used to track reagent usage and reorder reagent to prevent depletion of inventory and corresponding test delays. Such systems (combined with vendor reagent lot sequestering programs) can decrease the need for new-lot workups and therefore maximize instrument and operator productivity.

Vendor-Supplied Versus User-Defined Reagents

Automated analyzers can be classified as "open" or "closed" with respect to reagent compatibility. A "closed" system analyzer is one in which reagents or calibrators (or both) are provided only by the instrument manufacturer and other reagents or methods cannot be used. Closed systems can be restricted based on proprietary licensing or technical restrictions of reagent components, reagent pack size and shape, or software restrictions associated with compatibility of vendor bar codes or the ability to configure custom test parameters. In an "open" analyzer, the operator can change the parameters related to an analysis and load user-defined reagents (UDRs) by preparing "in-house" reagents or using reagents from alternative suppliers. Such analyzers usually have considerable flexibility and

adapt readily to new methods and analytes. Many vendors of automated analyzers allow a limited number of UDRs on an otherwise closed system, with restrictions based on technical (instrument-related) or contractual limitations. Advantages of UDRs include decreased cost, increased flexibility, and the ability to optimize assay performance. Disadvantages of UDRs may include additional regulatory requirements for laboratory-developed tests, manual processes associated with reagent preparation, lack of manufacturer-associated bar code information on reagent packs, and smaller peer group sizes in proficiency testing challenges.

Reagent Delivery and Mixing

Liquid reagents are often acquired and delivered to mixing and reaction chambers by pumps (through tubes) or by positive-displacement syringe devices. In analyzers in which more than one reagent is acquired and dispensed by the same syringe, washing or flushing of the probe is essential to prevent reagent carryover. Reagent carryover may interfere with subsequent reactions depending on which combinations of tests were sequentially performed on the instrument. Reagent carryover experiments are particularly important when considering UDR kits because UDRs would not have been evaluated by the instrument vendor for potential interference because of constituent carryover.

Chemical Reaction Phase

Sample and reagents react in the chemical reaction phase. Factors that are important in this phase include (1) the vessel where the reaction occurs; (2) the cuvette where the reaction is monitored; (3) the timing of the reaction(s); (4) the mixing and transport of reactants; (5) the thermal conditioning of fluids; and (6) for some immunoassay systems, the separation of bound and unbound fractions.

Type of Reaction Vessels and Cuvettes

In continuous-flow systems, each specimen passes through the same continuous stream and is subjected to the same analytical reactions as every other specimen and at the same rate. In such systems, the reaction occurs in the flow-through component. In discrete systems, each specimen has its own physical and chemical reaction space, separate from every other specimen. Discrete analyzers use either individual (disposable or reusable) reaction vessels transported through the system after sample and reagent have been dispensed, or they use stationary reaction chambers.

Reaction vessels (typically cuvettes) are reused in many instruments. The time before reusable reaction vessels must be replaced depends on their composition and the washing mechanisms used. Some manufacturers have computer algorithms that automatically control when individual reaction vessels should be replaced, depending on how many assays and what types of assays have been performed in a given cuvette.

Mixing of Reactants

Various techniques are used to mix reactants. In discrete systems, these include (1) forceful dispensing, (2) magnetic stirring, (3) vigorous lateral displacement, and (4) physical stirring. Dry reagent systems obviate the need for mixing because the serum completely interacts with the dry chemicals as it flows through the matrix of the reaction unit. However, regardless of the technique used, mixing can be a difficult process to automate and one that may contribute to reaction-to-reaction carryover among reused components. Inadequate mixing can lead to erroneous test results that may only be detectable by QC failures.

Thermal Regulation

Thermal regulation requires the establishment of a controlled-temperature environment in close contact with the reaction container and efficient heat transfer from the environment to the reaction mixture. Various technologies have been used for temperature regulation, including air baths, water baths, and Peltier devices. Onboard water baths often require specific maintenance and cleaning, particularly if the reaction monitoring system involves passage of light through the water bath to a reaction cuvette. Automated analyzers may also be influenced by external operating temperatures in the laboratory. Most vendors define acceptable temperature (and humidity) limits for their instruments.

Separation in Immunoassay Systems

Automation of immunoassay procedures often requires the separation of free and bound fractions of heterogeneous immunoassays; see Chapter 15 for additional information. Several approaches have been used to accomplish this separation. For example, automated immunoassay analyzers may use bound antibodies or proteins in a solid-phase format. With this approach, the binding of antigens and antibodies occurs on a solid surface to which the antibodies or other reactive proteins have been adsorbed or chemically bonded. Different types of solid phases are used, including (1) beads, (2) coated tubes, (3) microtiter plates, (4) magnetic and nonmagnetic microparticles, and (5) fiber matrices.

Measurement Approaches

Automated chemistry analyzers have traditionally relied on photometers and spectrophotometers to measure the absorbance of the reaction produced in the chemical reaction phase; see Chapter 9 for additional information. Alternative approaches now incorporated into analyzers include reflectance photometers, fluorometers, and luminometers. Immunoassay reactions that produce fluorescence, chemiluminescence, and electrochemiluminescence enhance sensitivity. Ion-selective electrodes and other electrochemical techniques also are quite common. Many of these methods are discussed in detail elsewhere in this textbook (see Chapters 9 and 15).

Postassay Processing

After a test is performed on a specimen, additional automated or manual specimen processing may still be required to obtain a final, clinically reportable result. Two such processes that are commonly performed are dilutions and repeat testing before reporting.

Dilutions

If initial results are above the assay's AMR, dilutions may be required to bring the analyte concentration into the linear range. Dilutions may also be used to exclude the possibility of antigen excess ("hook effect"; see Chapter 15). Some instruments can perform automated dilutions by (1) pipetting less specimen volume into the reaction mixture, (2) pipetting more reagent into the reaction mixture, or (3) using onboard dilution vessels. Instrument dilutions may be performed automatically (based on analyzer settings) or triggered by operator request. Analyzers that are unable to perform autodilutions may still notify operators when an offline dilution is required. Many analyzers will then allow the operator to input the dilution factors so that final results are calculated automatically to limit the chance of error.

Repeat Testing

Many laboratories have protocols that require certain assays to be automatically repeated before reporting. Some examples include critical results or tests for life-altering diagnoses. Improved analytical methods and instrumentation have challenged the need for routine repeat testing. Regardless, many automated analyzers allow configurations for repeat testing without operator intervention, thus decreasing the time for result availability and communication to the ordering provider.

Signal Processing, Data Handling, and Process Control

The interfacing and integration of computers into automated analyzers and analytical systems has had a major impact on the acquisition and processing of analytical data. Analog signals from detectors are converted to digital signals. Computer software then processes the digital data into useful and meaningful output. Data processing has allowed automation of complex calibration relationships. Computer workstations provide a central point of communication with the user regarding (1) instrument status, (2) order entry, (3) result display, (4) QC functionality, (5) instrument troubleshooting, (6) instrument maintenance, and (7) result release to the LIS. For small laboratories, these processes may be tracked primarily within an individual instrument computer and the LIS. For laboratories that have multiple analyzers (often from different vendors), tracking of these processes within additional computer systems between the instruments and LIS (termed *middleware*) can enhance overall efficiency.

Test Autoverification

Autoverification is the process whereby patient results generated from interfaced instruments are compared by computer software against laboratory-defined acceptance parameters. If results fall within these defined parameters, the results are automatically released for reporting with no additional laboratory staff intervention. Any data that fall outside the defined parameters are reviewed by laboratory staff before reporting. Autoverification has been implemented in a variety of ways, such as by using LIS to compare results against acceptance parameters, or alternatively, by performing this function in middleware. Most autoverification systems use rule-based decision logic. The rules applied in such systems require careful development and validation by laboratories before and after they are implemented. Regulations regarding autoverification are incorporated into most laboratory accreditation programs.

Fig. 16.5 Example of an integrated chemistry and immunoassay system. The ARCHITECT *ci*8200 (Abbott Diagnostics) combines a c8000 clinical chemistry module *(left)* with an *i*2000SR immunoassay module *(right)* in one connected platform. (Courtesy Abbott Diagnostics, Lake Forest, IL.)

Instrument Clusters

To reduce labor costs, instrument manufacturers have developed approaches that allow a single technologist to simultaneously control and monitor the functions of several instruments or modules. Initially, such workstations were configured with clusters of identical instruments, such as chemistry, immunochemistry, or hematology analyzers. Advanced instrument clusters may incorporate different chemistry, immunochemistry, or hematology analyzers into the same cluster or combine different types of detection systems into a platform (see Figs. 16.3, 16.5, and 16.6).

A cluster of analyzers has its own central computer control module with software designed to assist the operator in monitoring the functions of each analyzer and to aid in the review of laboratory results generated by the cluster. Access to the many front-panel functions of each analyzer is provided by the interface between the analyzer and the central control module. Thus, the operator loads specimens onto each instrument (or a shared input module) and then monitors subsequent instrument operation and result review from a central workstation. Incorporating the activities of multiple instruments into a single integrated workstation can improve overall efficiency.

Microtiter Plate Systems

Microtiter plate systems are commonly used in immunoassays and nucleic acid analyses. Microtiter plates (as used for enzyme-linked immunosorbent assays [ELISAs]) are usually made of polystyrene and have 48 or 96 wells that are coated with an antibody that is specific for the antigen of interest. After incubation of serum in the microtiter plate well, the well is washed to remove unbound antigen, and a second antibody

Fig. 16.6 Example of an integrated chemistry and immunoassay system. The Cobas 8000 (Roche Diagnostics) modular analyzer series configuration shows (from *left* to *right*) a core unit, an ion-selective electrode module, as well as c702 and c502 chemistry and e602 immunoassay instruments. Module sample buffers allow additional capacity and random access distribution of racks. (Courtesy Roche Diagnostics, Switzerland.)

with conjugated indicator enzyme is added. After a second incubation period, the well is washed to remove the unbound conjugate. A color-producing product is developed by the addition of enzyme substrate, and the reaction is terminated at a specific time. With the development of automated pipetting stations (see next section), the liquid handling steps have been fully automated to make microtiter plate assays a viable technology for carrying out large numbers of immunoassays. Measurement of absorbance and data processing are also automated in many instruments.

Automated Pipetting Stations

Automated pipetting stations have a Cartesian (also referred to as linear) robot with a pipette fixed to the end of a probe that moves within a rectangular space. The probe is capable of moving in the x-, y-, and z-axes. Liquids may be aspirated and dispensed in any location within the rectangular space. Pipetting stations may be used to automate an analytical procedure for which an automated analyzer does not exist or cannot be cost justified. Pipetting robots should (1) be relatively easy to program, (2) rarely malfunction, and (3) be capable of delivering aliquots of liquids with extreme precision and accuracy. Multiple-channel pipetting robots allow parallel processing of specimens with 8- or 12-channel probes to handle microtiter plates.

AREAS OF ANALYTICAL AUTOMATION

Although the automation principles discussed earlier focus primarily on clinical chemistry applications, the following sections describe aspects of analytical automation that are unique to individual disciplines, including (1) clinical chemistry, (2) immunology and immunochemistry, (3) urinalysis, (4) hematology, (5) coagulation, (6) transfusion, (7) microbiology, (8) mass spectrometry, and (9) molecular diagnostics.

Clinical Chemistry

Automated systems in clinical chemistry can be differentiated based on (1) the volume of testing that is performed on the instruments(s) and (2) their ability to connect to instrument clusters or a larger laboratory automation system.

Fig. 16.7 Example of an automated bench-top special protein analyzer (Optilite, The Binding Site Group). (Courtesy The Binding Site Group Ltd., Birmingham, United Kingdom.)

Laboratories that perform a lower volume of testing or that have limited available floor space may choose smaller analyzers to minimize cost, complexity of maintenance, and instrument footprint. Many low-volume laboratory automated chemistry instruments include similar functionalities to those seen on high-volume instruments, including a diverse assay menu, multiple methods of analytical detection, compatibility with UDRs, onboard QC management, and LIS connectivity. Smaller instruments may also be used in large-volume laboratories for specialized methods, where test order volume or assay compatibility is not supported or not cost-effective on higher volume instruments (Fig. 16.7). For example, use of smaller instruments for esoteric testing may be amenable to **batch analysis** with less reagent waste.

Automated instruments for large-volume laboratories typically improve efficiency by handling more tests per hour

Fig. 16.8 Example of a high-throughput automated clinical chemistry system. Shown is a Beckman Coulter AU5840 Clinical Chemistry System, which includes loading and rack buffer units, dual ion-selective electrodes, and four analytical modules. (AU 5800 Copyright 2016 Beckman Coulter, Inc. All rights reserved. Used with permission.)

with a larger capacity for specimens and reagents (Fig. 16.8). Larger instruments may have more physical plant (infrastructure) requirements than their smaller counterparts, including floor space, ventilation, water supply, electrical, and drainage. Although these requirements may place greater initial costs on a larger automated system, when they are in place, they support greater efficiency of instrument operation. Larger automated instruments also typically offer greater options for connectivity in instrument clusters and TLA solutions.

Larger systems may also have more complex maintenance procedures. Furthermore, instrument downtime for service may have a significant impact on clinical operations with high-volume consolidated platforms. For this reason, many high-volume laboratories are designed to incorporate instrument and assay redundancy so that if one analyzer is down for service, clinical testing can still be conducted on separate instruments. This is particularly important for laboratories that support acute-care, inpatient, and emergency department operations.

Along with routine chemistry analyzers that perform traditional photometric and potentiometric measurements, automation in clinical chemistry also includes specialized instrumentation for specific clinical needs. Blood gas analysis (including instrumentation in clinical laboratories, STAT laboratories, and point of care or POC) is conducted on automated instrumentation. These platforms may include built-in QC assessment; single- or multiple-use reagent systems; and quantification of additional analytes that may be useful to near-patient testing, including electrolytes, glucose, ionized calcium, lactate, pH, hemoglobin, or hematocrit. Automated platforms also exist for other specialized techniques such as HbA1c measurement, protein electrophoresis, osmometry, viscosity, and erythrocyte sedimentation rate.

Immunochemistry and Immunology

Antibody-based detection methods are incorporated into many automated instruments and include nephelometry, turbidimetry, and a wide range of immunoassays incorporating enzymes, labels, fluorophores, and chemiluminescent detection methods (see Chapters 9 and 15). One advantage of using antibodies in automated instrumentation is that by varying reagents (antibodies), the same instrumentation may be used for the detection and quantification of many different analytes.

Fig. 16.9 Example of an automated immunofluorescence microscopy system (EUROPattern, EUROIMMUN). (Courtesy EUROIMMUN.)

Along with discrete processing systems, automated immunoassay platforms may also be designed for microplates and microwells, including ELISAs and cell-based immunofluorescence assays. Automated indirect immunofluorescence microscopy systems are also available (Fig. 16.9). Automated systems are offered for processing of Western blots and immunoblots, techniques that have otherwise been very manually intensive. Finally, bead-based multiplex systems allow for the detection of multiple analytes in the same reaction mixture. Such systems can decrease the specimen volume required for parallel analyses.

Urinalysis

Although many of the same analytical principles are used for the quantification of serum and urine chemical constituents, traditionally it has been more challenging to fully automate testing of urine because of the broad range of urine analyte concentrations. This requires a low limit of detection to measure low concentrations and expanded linearity to permit

Fig. 16.10 Example of an automated urinalysis system. The Iris iRICELL3000 (Beckman Coulter) combines urine chemistry analysis (iChemVELOCITY, *right*) with a digital flow morphology module (iQ200SPRINT, *left*). (iRICELL Copyright 2016 Beckman Coulter, Inc. All rights reserved. Used with permission.)

measurements of high concentrations without dilution. This requirement, together with the relatively low demand for urine tests compared with that for serum, slowed the development of analyzers designed specifically for urine constituents. Several automated urinalysis instruments are now commercially available for chemical analyses (automated and semiautomated test strip reading by reflectometry) and analysis of sediment, particles, or cellular elements using flow cytometric or microscopic analysis. Many systems combine both chemical and particle analysis modules into a workcell (Fig. 16.10) so that particle analysis can be conducted automatically on specimens with abnormal test strip results. Instruments for microscopic urine analysis have also been applied to automating body fluid cell counting. Finally, although manual reading of urine test strips is still commonly performed as a POC test, small semiautomated instruments are available for Clinical Laboratory Improvement Amendments (CLIA)-waived automated reagent strip interpretation.

Hematology

Analyzers that perform a complete blood count have been automated through the use of (1) the "Coulter principle," which is based on cell conductivity; (2) light scatter; (3) flow cytometry; and more recently, (4) digital microscopic analysis. Individual blood cells are analyzed by application of one or more of these techniques. The Coulter principle is based on changes in electrical impedance produced by nonconductive particles suspended in an electrolyte solution as they pass through a small aperture between electrodes. RBCs, white blood cells (WBCs), and platelets are identified by their sizes.

Flow cytometry typically counts cells (often stained with supravital or fluorescent dyes) that travel in suspension, one by one, past a laser light source. Scattered light and emitted light are collected in front of the light source and at right angles, respectively. Information derived through measurement of light scatter when a cell is struck by the laser beam is then used to estimate (1) cell shape, (2) cell size, (3) cellular granularity, (4) nuclear lobularity, and (5) cell surface structure. Although some cell counters classify WBCs using the Coulter principle, cell conductivity, and light scattering of unstained cells to differentiate cell types, other cell counters use multiple flow cytometry channels or a combination of approaches.

When abnormal cells or cell counts outside of predetermined limits are observed in automated counting methods, it may be necessary to generate a blood smear for additional microscopic review or generation of a manual differential. Integrated hematology systems may include automated slide makers and slide stainers to limit the manual intervention involved in creating and labeling slides (Fig. 16.11). In addition, advances in slide imaging have enabled the development of automated systems to classify and count cell types based on a digital scan of the slide or a representative number of regions of interest. Automated digital imaging systems may also offer electronic review of categorizations by an operator, as well as electronic storage and sharing of images with clinicians outside the laboratory.

Coagulation

Most coagulation testing is now performed on automated analyzers. A variety of methodologies are used, including optical and mechanical clot detection, as well as turbidimetric, chromogenic, and antibody-based analyses. Many instruments permit use of UDRs for more specialized coagulation testing. Connectivity to TLA systems is an option with some instrumentation, and this allows for automated specimen handling and centrifugation before coagulation testing. Additional instrumentation and methods used in coagulation testing include aggregometers, platelet function analyzers, and thromboelastography. Automated POC and near-patient devices are also available for many coagulation tests including prothrombin time, international normalized ratio, partial thromboplastin time, and activated clotting time.

Transfusion Medicine

Advances in laboratory automation have also impacted most areas of blood banking and transfusion medicine. Robust information technology systems to track donor history, blood component traceability, bar coding, transfusion-transmissible diseases, immunohematological testing, inventory

Fig. 16.11 Example of an automated hematology system. A scalable Sysmex XN-9000 Automated Hematology System, including three XN-10 hematology test modules and one SP-10 module for slide making and staining *(left)*. (Courtesy Sysmex America, Inc., Lincolnshire, IL.)

Fig. 16.12 Example of an automated transfusion medicine analyzer. The ORTHO VISION Analyzer (Ortho Clinical Diagnostics) offers automated immunohematology testing. (Courtesy Ortho Clinical Diagnostics, Raritan, NJ.)

management, and component administration all work to improve the safety and traceability of transfused blood products. Automation at blood collection sites through apheresis donations helps to maximize the yield of specific blood component donations. Automated component processing systems at blood centers improve the efficiency of postdonation handling of blood products, thus enabling fresh components of blood products (with less chance of processing errors) to ultimately be distributed for transfusion.

Traditionally, immunohematological testing in the blood bank was conducted primarily through test tube RBC agglutination techniques. Both small- and large-scale systems are now available to automate many of these techniques, including RBC typing (ABO, Rh), antigen typing, screening and identification of alloantibodies, direct antiglobulin testing (DAT), dilutions, titrations, reflex testing, and crossmatching of potential products for transfusion (Fig. 16.12). In general, automated blood bank systems are as reliable and sensitive as routine manual testing. Advantages include less opportunity for processing and transcription errors. The decision to automate immunohematology testing includes a variety of factors, such as staffing and expertise, test volumes, variation in orders between day and night shifts, workflow considerations, and overall expense. Finally, RBC molecular typing (genotyping) may further refine how donor and recipient screening is automated within the blood bank.

Microbiology

Clinical microbiology has recently undergone a transition toward far greater automation. Although standalone instrumentation for aseptic plate preparation, blood culture incubation and microbial detection, biochemical identification, and antimicrobial susceptibility have all been commercially available for years, much of the processing and interpretation steps within the microbiology workflow remained largely manual. Both standalone and TLA solutions are now available, however, to automate many of these processes.

TLA microbiology solutions with tracking systems have been designed to handle preanalytical, analytical, and postanalytical steps (Fig. 16.13) and may include specimen decapping and recapping; Petri dish bar code labeling; plate inoculation and streaking; Gram stain preparation; automated incubation (O_2 and CO_2); high-resolution digital imaging of plates; and colony sampling for biochemical identification, antibiotic susceptibility testing, and seeding of plates for subsequent matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometric analysis. MALDI-TOF is a rapidly growing area of microbiology and is used for organism identification.

POINTS TO REMEMBER

Analytical Automation

- Analytical automation has impacted all areas of laboratory medicine.
- Along with automating the testing phase, analyzers may incorporate specimen handling, reagent management, QC, and result review.
- Automated processes, including liquid-level sensing, clot detection, closed-container sampling, and instrument autodilution, help to improve the efficiency of laboratory testing.
- Analytical systems can be organized into clusters or TLA solutions, allowing operators to control multiple instruments.

Fig. 16.13 Example of an automated microbiology system. The WASPLab system (Copan) includes automated specimen processing, management, incubation, and digital imaging. (Courtesy Copan Diagnostics, Murrieta, CA.)

Molecular Diagnostics

Automation has been crucial for the expansion of molecular diagnostics into the clinical laboratory. Once a guaranteed money sink that only academic laboratories could afford, the realm of DNA and RNA analysis is now a cherished part of most laboratories, large or small. **Molecular diagnostics** has come to mean analysis of DNA or RNA, unfairly excluding proteins, electrolytes, small molecules, and other polymers.

Molecular diagnostics is a relatively new field, with the structure of DNA only described in 1953. The first molecular method with the necessary sensitivity and specificity to detect sequence changes in DNA was **Southern blotting**, introduced by Edwin Southern in 1975. Southern blotting was the antithesis of automation requiring multiple manual steps, including DNA isolation, treatment with restriction enzymes, electrophoresis, blotting, preparation of radioactive probes, hybridization of the probes, and finally development of signal by exposure to film. Early sequencing was also very labor intensive, requiring cloning of the fragment to be sequenced and multiple steps to eventually read the sequence base by base from different sized fragments separated by electrophoresis. Although many of the steps in Southern blotting and sequencing can be automated, it was the simplicity of the **polymerase chain reaction (PCR)** in the 1980s and **real-time PCR** in the 1990s that popularized molecular diagnostics in clinical laboratories.

Automation has had major impacts in (1) nucleic acid preparation, (2) liquid handling, (3) amplification, and (4) analysis of molecular diagnostic assays. Nucleic acid preparation typically involves (1) cell lysis, (2) separation from proteins, and (3) purification from interfering substances. Sample preparation instruments today often rely on reversible binding to silica-coated paramagnetic beads that can easily be washed to remove impurities. Most molecular methods require several mixing or transfer steps that can be automated by liquid handling. For example, PCR requires mixing reaction components together before thermal cycling. Although many components are common to all reactions and can be combined as a "master mix," template DNA and primers are unique and must be pipetted individually for each reaction. One common use for the Cartesian pipetting stations mentioned previously is to automate the setup of such reactions when large numbers of samples or assays need to be processed. Instead of onsite automation just before the assay, another alternative is to perform the automation well before the assay is performed. For example, primer plates can be made, dried, and stored by a Cartesian automation station that patterns different primer pairs on a 96-well plate when multiple assays are performed together, greatly reducing the complexity for the performing technologist. Commercial vendors can provide one- or two-dimensional arrays of targets, including millions of hybridization targets on a chip for expression profiling, genotyping, or copy number assessment. The commercial use of automation can provide products to the laboratory that greatly reduce the complexity at the site where the assay is performed, albeit often with a loss in flexibility.

Although single-molecule techniques are now available, most molecular methods require target or signal amplification for visualization. For example, PCR requires repetitive temperature cycling of multiple reactions between two or three temperatures, a task begging for automation that has resulted in many different instruments. Isothermal methods of amplification were developed many years later and nicely match rapid, near-patient testing. Automated analysis of molecular assays can be as simple as size interpolation of samples from a ladder of standards on electrophoretic gels or as complicated as aligning billions of massively parallel sequence reads against a reference human genome. Such massive data analysis is only possible because of a parallel increase in computer storage and computational speed enabled by modern computers.

Sequencing Automation

Traditional Sanger sequencing is a complicated multistep process, originally performed with radioactivity on 2D slab gels. By using fluorescence instead of radioactivity and by acquiring, storing, and analyzing sequence information from a one-dimensional tube gel directly on a computer, the analysis part of sequencing was automated. Although this was a great advance and led to the widely used capillary sequencers of the 1990s and 2000s, only the final task of data acquisition and analysis was automated. The process still required (1) initial amplification by PCR, (2) cleanup of the PCR product, (3) addition of cycle sequencing reagents, (4) cycle sequencing with dideoxy termination, (5) cleanup of the termination products, and (6) transfer to a capillary electrophoresis instrument. Although many of these steps were eventually automated in the largest sequencing centers, completely integrated liquid handling for sequencing was seldom obtained in clinical laboratories. Massively parallel sequencing in the mid-2000s introduced entirely new workflows. Sequencing by synthesis is now most common and uses the stepwise addition of fluorescently labeled reversible terminators. The amplification, synthesis, and data acquisition steps are entirely automated. Even though sample preparation is complex and mostly manual, the effort is justified by the massive amounts of data generated. Data analysis is also complex and not standardized, requiring custom software "pipelines" to align each sequenced read to a reference sequence, determine or call each base, and determine the clinical significance of each variant from the reference sequence.

Polymerase Chain Reaction Automation

PCR was first automated before thermostable polymerases were available. Because the enzyme was inactivated in each cycle at the high temperatures needed for DNA denaturation, enzyme needed to be replenished in each cycle. Automation required not only temperature cycling but also the addition of reagents in each cycle. The combination of thermal control and repetitive reagent addition was complex and expensive. Although the liquid handling could be automated, it was made obsolete by the use of thermostable polymerases that could withstand the required DNA denaturation temperatures. A simple change in the process greatly simplified the automation requirements.

The thermal cycling provided by PCR instruments was at first slow, requiring 4 hours for amplification. Most common platforms used metal blocks that held samples in tubes or plates that were convenient to handle, although circulating systems allow better thermal control and faster cycling. Rapid-cycle PCR was developed in the early 1990s and reduced PCR amplification times down to 10 to 15 minutes. More recently, extreme PCR amplifies small products in as little as 15 seconds, indicating that the speed of PCR is limited by the instrumentation, not the biochemistry. Often, physical automation cannot keep up with the speed of molecular reactions.

Real-time PCR combines fluorimetry with thermal cycling, allowing fluorescence monitoring of double-stranded DNA production during PCR. When the fluorescent signal is acquired once each PCR cycle, template quantification can be automated. When the fluorescent signal is acquired during temperature transitions, melting curves are generated that monitor DNA hybridization. Both applications of quantification and melting curve analysis are enabled by the simple automation of fluorescence signal acquisition. Nucleic acid quantification, previously very difficult to obtain by PCR, is greatly simplified and used extensively in transcript quantification. Melting analysis and its more precise sibling, high-resolution DNA melting, have replaced manual gels in many applications and provide closed-tube genotyping, scanning, and relative quantification without probes. When both control and acquisition are automated, powerful systems, such as real-time PCR can result, similar to many automated laboratory instruments.

Throughput and Molecular Applications

Throughput is defined as the number of samples processed in a given period of time. Automation can improve throughput either by increasing the batch size of a run (the number of samples analyzed at once) or by decreasing the turn-around time (the time it takes to process each sample or run). Automation has successfully improved throughput by increasing batch size, from single samples, to strips of 8 or 12, to 96-, 384-, or even 1536-well plates. For example, large automated molecular microbiology systems are used for infectious disease testing for HIV, hepatitis C virus, and several other major infections. Less emphasis has focused on decreasing turn-around time, although rapid (<1-hour) sample-to-answer systems for molecular targets are now approved by the Food and Drug Administration. The symbiosis of automation and molecular testing will continue to lead to faster, better, and less expensive solutions for clinical laboratories.

REVIEW QUESTIONS

1. Which type of specimen label has resulted in a reduction in preanalytical specimen identification errors and is used extensively in automated specimen identification?
   - **a.** Social Security numbers stamped on all individual specimen containers
   - **b.** Hospital identification numbers written on tubes of blood
   - **c.** Labels containing bar codes that are unique identifiers
   - **d.** Medical accounting numbers written on the lids of all containers
2. One serious issue with the use of accelerating/decelerating pneumatic tube systems to deliver specimens to the laboratory is the occurrence of:
   - **a.** clot formation.
   - **b.** hemolysis.
   - **c.** specimen breakage.
   - **d.** specimen volume loss.

3. Which of the following activities are important in the evaluation of the need for automation and in identifying the solution?
   a. Workflow mapping
   b. Plotting the volumes of arriving specimen by the hour of the day
   c. Vendor analyses and presentations of findings
   d. Visiting laboratories with installed systems and asking frank questions
   e. All of the above
4. The transport of a quantity of analyte or reagent from one specimen reaction into a subsequent one is referred to as
   a. batching.
   b. carryover.
   c. misdelivery.
   d. indiscrete handling.
5. What makes automation of the polymerase chain reaction much easier?
   a. Capillary electrophoresis
   b. Fluorescence feedback to automatically control the denaturation temperature
   c. Primer specificity
   d. Thermostable polymerase
6. Future opportunities for automating massively parallel sequencing include:
   a. Sample preparation
   b. Capillary electrophoresis
   c. Thermostable polymerase
   d. Quantitative reverse transcriptase PCR

SUGGESTED READINGS

Armbruster, D. A., Overcash, D. R., & Reyes, J. (2014). Clinical chemistry laboratory automation in the 21st century–Amat Victoria curam (victory loves careful preparation). *Clinical Biochemistry Reviews*, *35*, 143–153.

Boyd, J. C., Felder, R. A., (2002). Preanalytical automation in the clinical laboratory. In: Ward-Cook, K. M., Lehmann, C. A., Schoeff, L. E., et al., eds. *Clinical diagnostic technology: the total testing process, Vol. 1. The preanalytical phase*. Washington, DC: AACC Press, 107–129.

Dorak, M. T. (Ed.). (2006). *Real-time PCR*. New York: Taylor and Francis.

Hawker, C. D. (2007). Laboratory automation: total and subtotal. In R. C. Friedberg, & R. L. Weiss (Eds.), *Clinics in laboratory medicine* (pp. 749–770). Philadelphia: Elsevier.

Hawker, C. D., Garr, S. B., Hamilton, L. T., et al. (2002). Automated transport and sorting system in a large reference laboratory. Part 1: evaluation of needs and alternatives and development of a plan. *Clinical Chemistry*, *2002*(48), 1751–1760.

Hawker, C. D., Genzen, J. R., & Wittwer, C. T. (2018). Automation in the clinical laboratory. In N. Rifai, A. R. Horvath, & C. T. Wittwer (Eds.), *Tietz textbook of clinical chemistry and molecular diagnostics* (pp. 370.e1–370.e24). Philadelphia: Elsevier.

Landegren, U., Kaiser, R., Caskey, T., et al. (1988). DNA Diagnostics—molecular techniques and automation. *Science*, *242*, 229–237.

Lyon, E., & Wittwer, C. T. (2009). LightCycler technology in molecular diagnostics. *The Journal of Molecular Diagnostics*, *11*, 93–101.

Melanson, S. E. F., Lindeman, N. I., & Jarolim, P. (2007). Selecting automation for the clinical chemistry laboratory. *Archives of Pathology & Laboratory Medicine*, *131*, 1063–1069.

Smith, L. M., Sanders, J. Z., Kaiser, R. J., et al. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature*, *321*, 674–679.

Point-of-Care Testing

Andrew St. John, Maurice O'Kane, Christopher P. Price

OBJECTIVES

1. Define point-of-care testing (POCT) and other terms used to describe this process.
2. Describe the analytical requirements and technological considerations for POCT.
3. Describe examples of POCT devices.
4. Describe the interrelationship between informatics and POCT.
5. Describe key factors in the implementation and management of POCT programs.
6. Discuss types of evidence that can be used to demonstrate the effectiveness of POCT.

KEY WORDS AND DEFINITIONS

Accreditation The formal recognition by an independent body that the process or service operates according to defined standards.

Analyte The substance that is to be measured (also known as a *measurand*).

Audit The examination of the quality of performance of a process measured against a predetermined standard.

Connectivity The property of a device that enables it to communicate with an information system (e.g., a laboratory information system) for the primary purposes of transmitting patient data from the device to the patient's record, and for monitoring the performance of the device.

Effectiveness Demonstrating through evidence that POCT brings benefits to patients and the overall clinical care process.

Fluidics Process by which liquid moves within a confined space, as in the case of a narrow tube or a porous matrix. Such processes include surface tension, diffusion, and the use of pumps.

Informatics The structure, creation, management, storage, retrieval, dissemination, and transfer of information.

Operator Interface The part of a device that the operator is required to use in order for the device to work (e.g., switch on a reader, enter a patient or sample identification, or calibrate the device).

Point-of-Care Testing (POCT) A mode of testing in which the analysis is performed close to the patient rather than in a laboratory.

Quality Management The act of overseeing all activities and tasks needed to maintain a desired level of excellence.

Recognition Element A molecule that is used for the recognition (and sometimes capture) of the analyte of interest (e.g., enzyme, antibody, affinity ligand).

Sensor A device that receives and responds to a signal or stimulus. There are many examples in life, including the receptors of the tongue, the ear, and so on.

Transducer A substance or device that converts input energy in one form into output energy of another form.

Point-of-care testing (POCT) can be simply defined as "medical testing at the site of patient care" that therefore allows more rapid clinical decision-making with the possibility of improved health outcomes. Many other terms have been used to describe POCT, including "bedside," "near patient," "physician's office," "extra-laboratory," "decentralized," "off-site," "satellite," "kiosk," "ancillary," and "alternative site" testing. POCT also includes "self-testing," where the patients perform the tests themselves. Self-monitoring for blood glucose and prothrombin time or international normalized ratio (INR) together form the largest segment of the POCT market.

Advances in a range of technologies and manufacturing processes such as thin film sensors, semiconductor engineering, plastic moulding, microfluidics, nanotechnology, and consumer electronics, make it possible to adapt most of the methods used in the laboratory to the point-of-care setting. In turn, these advances have now enabled health care providers and patients to make a choice about where testing is conducted (Box 17.1).

BOX 17.1 Environments Where Point-of-Care Testing Might Be Employed

Primary Care
Home
Community pharmacy
Health centers (general practice, primary care)
Workplace clinic
Physician's office and community clinic
Diagnostic and treatment center
Paramedical support vehicle (ambulance, helicopter, aircraft)

Secondary and Tertiary Care
Emergency room
Admissions unit
Ambulatory diagnostic and treatment center
Operating room
Intensive care unit
Ward
Outpatient clinic

BOX 17.2 Potential Advantages and Disadvantages of Point-of-Care Testing

Advantages
Reduced turnaround time for test results
Improved patient management; improved interaction between patient and caregiver
Improved patient morbidity and mortality
Reduction in the administrative work associated with test requesting and reporting
Reduced risk of errors
Reduction in clinic visits
Reduction in hospital admissions
Improved cost of care

Disadvantages
Increase in administrative work associated with training and certification of operators
Care giver required to perform test
Increased risk of errors if tests performed by untrained operators

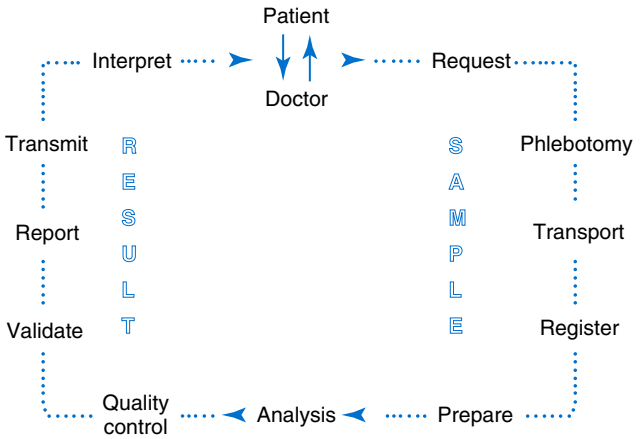


Fig. 17.1 Schematic representation of the key steps in requesting, delivering, and using a diagnostic test result.

Clinical needs met by POCT can include more rapid clinical decision-making (especially in situations of clinical urgency), with the possibility of improved clinical outcomes. POCT can also reduce the complexity of the testing process—so-called process needs—which include improving the immediate patient/clinician/caregiver interaction, as well as expediting the patient journey, and improving the patient's experience and patient empowerment. The patient testing process is shown in Fig. 17.1, while the advantages and disadvantages of POCT are summarized in Box 17.2.

TECHNOLOGY

The technology used to measure **analytes** outside of the conventional laboratory setting has come about partly through the long-term trend of miniaturization—that is, reducing the size of electronic devices, as in the case of mobile phones. Advances in information and communications technologies have also played an important role. Starting with instruments to measure electrolytes, blood gases, and glucose, it is now possible to measure many other analytes using increasingly smaller devices. Accompanying the process of miniaturization has been the development of dry, stable reagents that can be incorporated in disposable unit-dose devices. While the throughput of tests for these devices is low (e.g., typically <10 samples per hour depending on test type), the time required to produce the results is usually short (e.g., may vary from seconds for capillary blood glucose testing to several minutes for immunoassay based tests).

Required Features of Point-of-Care Testing Devices

As with any device, it is important that designers first consider the needs of users, and these depend to some extent on the clinical setting or application. Some simply stated but critical requirements are as follows:

1. Devices should be simple to use.
2. Reagents and consumables must be robust over extended periods of time.
3. Results from POCT devices should be concordant with established laboratory methods, and the analytical quality of results should be appropriate to the intended clinical purpose.
4. The device together with its associated reagents and consumables must be safe to use.

Design criteria become more specific for particular clinical settings. For example, criteria exist for devices designed for use in the developing world for sexually transmitted diseases (STDs), a major unmet medical need that is just beginning to be addressed effectively. Provided by the World Health Organization (WHO), the so-called ASSURED criteria are as follows:

- Affordable—for those at risk of infection
- Sensitive—minimal false negatives
- Specific—minimal false positives
- User-friendly—minimal steps to carry out test

TABLE 17.1 Classification of Types of Point-of-Care Testing (POCT) Instruments or Devices

Classes of Technology	Analytical Principles Used in the Classification	Example Analytes
Small handheld, single-use, POCT devices	Reflectance	Urine and blood chemistry
	Lateral-flow or flow-through immunoassays	Infectious disease agents, cardiac markers, hCG
Small handheld, single-use, POCT devices with a monitoring device	Reflectance	Glucose
	Electrochemistry	Glucose
	Reflectance	Blood chemistry
	Light scattering/optical motion	Coagulation
	Lateral-flow, flow-through, or solid phase immunoassays	Cardiac markers, drugs, CRP, allergy, and fertility tests
	Immunoturbidimetry	HbA1C, urine albumin
	Spectrophotometry	Blood chemistry
	Electrochemistry	pH, blood gases, electrolytes, metabolites
	Fluorescence, electrochemistry with PCR	Infectious agents
Larger cartridge-type and bench-top POCT devices	Electrochemistry	pH, blood gases, electrolytes, metabolites
	Fluorescence	pH, blood gases, electrolytes, metabolites
	Multiwavelength spectrophotometry	Hemoglobin species, bilirubin
	Time-resolved fluorescence	Cardiac markers, drugs, CRP
	Electrical impedance	Complete blood count
	Polymerase chain reaction	Bacteria and viruses

- Rapid and robust—short turnaround time (TAT) and no need for refrigerated storage
- Equipment-free—no complex equipment
- Delivered—to end users

Connectivity to information systems, including the patient health record, is a growing requirement in modern health care systems.

A more recent design challenge is the need to simultaneously measure multiple analytes in the same cartridge—so-called multiplexing. The Piccolo chemistry analyzer (Abaxis, Union City, CA) and blood gas or critical care analyzers are some of the relatively few examples of such technologies. The main benefit is the reduced overall time to result.

Design

There is a great diversity of devices being used for POCT (Table 17.1), the majority relying on technologies devised more than two decades ago. Many of the devices use the same analytical principles as those found in conventional laboratory analyzers, but newer technologies are starting to appear, some of which will be described.

Irrespective of the type of technology, all POCT devices should have several key components, which include (1) the operator interface, (2) bar code identification systems, (3) sample and reagent delivery mechanisms, (4) reaction cell, (5) sensors, (6) control and communications systems, and (7) data management and storage. The main objectives of the design are to (1) enable the required reaction to take place that allows measurement of the analyte of interest, (2) ensure reliable and stable performance of the device over a period of time, and (3) minimize the risk of operator error.

Operator Interface

The **operator interface** for a POCT device should (1) require minimal operator interaction; (2) guide the user through the operation; (3) indicate whether any key steps have failed; and (4) allow identification of the (A) operator, (B) patient, and (C) test to be measured.

Advances in information technology (IT) and electronics have had a major impact on this area. Types of user interfaces include (1) keypads, (2) bar code readers, (3) printers, and (4) touch screens.

Bar Code Identification Systems

Many POCT devices incorporate bar code reading systems for a number of purposes. These include (1) identifying the reagent package to the system—whether it be a single-use disposable or a multiuse reagent pack, (2) incorporating factory calibration data, and in some cases, (3) programming the instrument to process a particular test or group of tests. Some POCT devices use magnetic strips as a way of storing similar information, such as lot-specific calibration data. Other functions of the bar code reader include the capability to identify both the operator and the patient sample to the system.

Sample Types and Sample Delivery

The vast majority of POCT devices use whole blood or urine specimens; in the case of blood, this commonly involves a capillary or finger-stick sample. In this way the sample can be introduced to the POCT device directly. Alternatively, a venous or arterial blood specimen can be collected and a small sample removed for POCT; in some cases the blood collection tube can be inserted directly into the POCT instrument.

Sample access and delivery to the actual sensing component of the strip, cassette, cartridge, or fluidic cell are also key processes that are prone to error. Ideally, following the addition of the sample, there should be no further need for operator intervention. Devices are now available that minimize the degree of interaction to the point where a closed tube simply

has to be placed in a holder and the instrument automatically performs all other functions.

Reaction Cell

Where the analytical reaction takes place varies from a simple porous pad to a microfluidic cartridge, or surface within a cartridge. Advances in **fluidics** and fabricating techniques have been critical to the development of POCT devices. As analytical reactions have increased in complexity, more reaction cells/zones have been added. Thus, in the case of molecular tests, different reaction zones may be required for DNA extraction, amplification, and probe detection. Another consideration in the design of reaction cells is to ensure that the requisite mixing of sample and reagents occur, which is critical when using very small volumes.

Sensors

Much of the development focus in POCT devices has been concerned with the advances in **sensor** design of which there are various types. A chemosensor is one where the analyte has an intrinsic property, such as electrochemical, that enables it to be detected without a **recognition element**. More commonly, chemosensors used in many POCT devices have a transducing element such as a chemical indicator or binding molecule that recognizes the analyte to be measured and produces a signal, usually electrical or optical. A biosensor is distinguished from a chemosensor by having a biological or biochemical component as the recognition element. Enzymes are the most common biological element used followed by antibodies; transduction typically is via an optical or electrical signal, although there is now an increasing use of antibody and nucleic acid recognition elements.

Control and Communications Systems

In even the smallest device, there is a control subsystem that coordinates all the other systems and ensures that the required processes for an analysis take place in the correct order. Operations that require control include (1) insertion or removal of the strip, cartridge, cassette, or reagent; (2) temperature control; (3) sample injection or aspiration; (4) sample detection; (5) reagent addition—especially in a fluidic-based system; (6) mixing; (7) timing of the detection process; (8) detection of measurement signal; and (9) waste removal. Fluid movement is often accomplished by mechanical means through pumps or centrifugation, and by fluidic properties, such as surface tension; the latter is often a critical element in the design of simple strip tests and in microfabricated systems.

Data Management and Storage

IT is crucial to POCT devices and may include rules-based assessment of internal quality control results, computing, and reporting of patient results. In some systems, data transfer and management take place when the meter or reader is linked to a small bench-top device called a docking station, but increasingly devices use wireless technology for communication. These and other devices include communication protocols that allow data to be transferred to other data management systems, electronic medical records, and mobile computing devices.

Types of Existing Point-of-Care Testing Technology

POCT devices can be classified into small, handheld devices that often use qualitative or quantitative reagent strips (or similar technology with a reading device), and larger, bench-top devices that are often variants of ones used in conventional laboratories. This classification is somewhat arbitrary since there is no sharp demarcation. There are gradations in both size and internal complexity of devices, although differentiation can sometimes be made on the need for external power. The clear trend is toward devices becoming smaller and with greater functionality.

Small Handheld, Single Use Point-of-Care Testing Devices

Many POCT devices fall into this category, including (1) single- or multipad tests that are read visually, by reflectance photometry or electrochemically; (2) more complex pads that use light reflectance for measurement; and (3) fabricated cassettes or cartridges that incorporate methodologies such as immunochromatography and are used as immunosensors. All these devices are truly portable and for that reason their operational mode (i.e., proximity to the patient) differs from larger bench-top devices. For example, devices that require application of a whole blood sample such as glucose strips obviate the need for sample containers, labeling or transport, as compared with larger POCT instruments that may be some distance from the patient.

Single and multipad strip tests. These are relatively simple in construction and are composed of a pad of porous material, such as cellulose, that is impregnated with reagent and then dried. Samples are added to the device either by dipping (e.g., in urine analysis) or by spotting (e.g., in blood glucose analysis). In the latter case, there has been a gradual transition from optical to electrochemical detection, avoiding the need to wipe the strip to remove surplus blood prior to measurement. A critical operator factor when spotting the sample onto the pad is the need to cover the whole pad with the sample. In addition, because the reactions often do not proceed to completion, it is necessary for visually read tests to be timed between placing the sample on the pad and comparing the resulting color to a color chart. More complex pads are composed of several layers, the uppermost of which is a semipermeable membrane that prevents red cells from entering the matrix.

Immunostrips. These are biological sensors in which the recognition agent is an antibody that binds to the analyte. Detection of the binding event or signal **transducer** is usually via an optical mechanism, either reflectance or fluorescence spectrophotometry, although visual inspection has also been used. Immunosensors typically use solid phase technologies in conjunction with (1) flow-through, (2) lateral-flow, or (3) immunochromatography processes. In the flow-through

format, a heterogeneous immunoassay takes place in a porous matrix that acts as the solid phase and requires an additional reagent to wash and separate bound and unbound label. In lateral flow, the separation stage takes place as the sample passes along the porous matrix, and no additional wash solution is required. In all these different formats, uniform and predictable flow of the sample through or along the solid phase matrix is a major determinant of the reproducibility of the technique.

In a typical example of an immunostrip for cardiac troponin T (cTnT), the blood sample is added and first flows through a glass fiber fleece, which separates the plasma from whole blood. Simultaneously, two monoclonal anti-human cTnT antibodies, one conjugated to biotin and one labeled with gold particles as the signal antibody, bind to troponin T in the sample to form a sandwich complex. The antibody-troponin sandwich complex then flows in a lateral direction along the cellulose nitrate test strip until it reaches the capture zone, which contains streptavidin bound to a solid phase. The biotin in the antibody troponin complex binds to the streptavidin and immobilizes the complex. The complex is then visualized as a purple band by the gold particles attached to the signal antibody. The unreacted gold-labeled antibody moves farther down the strip, where it is captured by a zone containing a synthetic peptide consisting of the epitope of human cTnT and is visualized as a separate but similar colored band. The presence of this second band serves as an internal quality indicator because it shows that the sample has flowed along the test strip and the device has performed correctly.

While the majority of quantitative lateral flow immunoassay devices only measure a single analyte at a time, some can measure a panel of analytes, such as cardiac markers, fertility tests, and drugs of abuse.

Small Handheld, Single Use Point-of-Care Testing Devices With a Monitoring Device

Glucose Measurement. In clinical practice, capillary glucose measurement is the most common POCT application. These devices are biosensors because they all use an enzyme as the recognition agent—glucose oxidase (GO), hexokinase (HK), or glucose dehydrogenase (GDH)—with photometric (reflectance) or, now increasingly, electrochemical detection.

Modern glucose strips use *thick-film* technology with several layers, each having a specific function. These are shown diagrammatically for a reflectance-based strip in Fig. 17.2. When blood is added to a strip, both water and glucose pass into the analytical layer. For some photometric systems, erythrocytes are excluded by a spreading or separating layer that contains various components, including glass fibers, fleeces, membranes, and special latex formulations. In photometric systems, a spreading layer is also important for the fast, homogeneous distribution of the sample. In contrast, electrochemical strips use systems that rely on capillary action to move the blood sample over the electrodes. The support layer is usually a thin plastic material that in the case of reflectance-based strips may also have reflective properties.

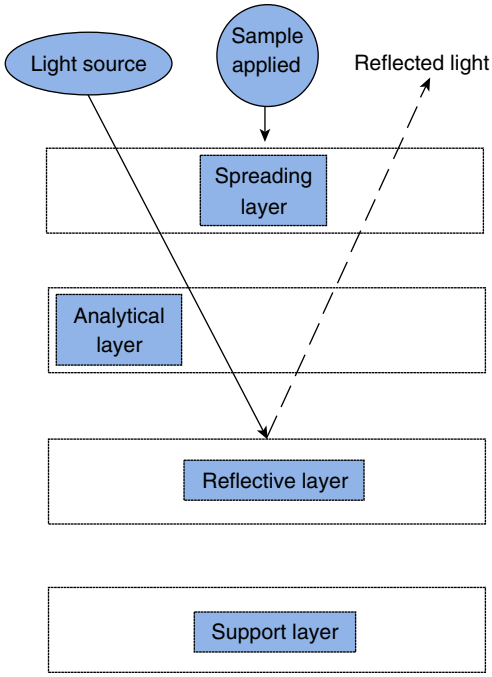


Fig. 17.2 Schematic diagram of the components that make up a typical reflectance-based glucose strip.

Additional reflectance properties have been achieved through the inclusion of substances such as titanium oxide, barium sulphate, and zinc oxide.

Glucose strips are produced in large batches, and after extensive quality assurance procedures by the manufacturer, each batch is given a code that is stored in a magnetic strip on the underside of each test strip. This code describes the performance of the batch, including the calibrating relationship between the photometric or electrochemical signal and the concentration of glucose. More recently, electrochemical glucose strips that do not require coding have become more commonplace. These are composed of an Ag-AgCl reference electrode and a carbon-based active electrode, both manufactured using screen printing technology with ferrocene or its derivatives contained in the printing ink. The conversion of glucose is accompanied by the reduction of ferrocene and the release of electrons. Continuing innovations have also facilitated the production of smaller meters, nonwipe strips, less need to clean instrument optics, more rapid results, and smaller sample volumes.

Although electrochemical systems have enabled the design of strips that are less subject to interferences, problems still persist. Strips that use GO are more substrate specific but are affected by oxygen tension, with high PO_2 values leading to falsely low results. Blood oxygen tension does not affect GDH based strips, but GDH strips that use the pyrroloquinolinequinone form of the enzyme are subject to interference by maltose. Excessive concentrations of the latter, such as in some parenteral solutions, can cause falsely high glucose results, masking hypoglycemia, and several deaths have been attributed to these errors. Hematocrit is another important interference, the effects of which have been reduced in newer strip designs.

Fig. 17.3 Epoc test card for blood gas analysis. (Courtesy Alere.)

INR Measurement. Another common meter-type device used by patients at home and by health care professionals in clinics are those employed to monitor warfarin (Coumadin) therapy and measure prothrombin time (reported as INR levels). First devised some two decades ago, a number of different technologies have been developed to measure INR at point of care, including optical and electrochemical detection. Innovation has been applied both to the strip, where a drop of blood is placed, and to the meter into which the strip is placed. In the HemoSense device (Alere, Waltham, MA), clotting is detected by change in impedance, while in a Roche device (Roche Diagnostics, Basel, Switzerland), a change in current is detected. Both devices incorporate QC systems that are activated when a patient sample is placed on the strip. In the case of HemoSense, the strip has three channels: one for the patient test and the other two for different concentrations of internal QC material.

Handheld integrated cartridge technology. The prime example of this type of technology is the i-STAT analyzer (Abbott Point-of-Care, Princeton, NJ), a handheld device originally designed to measure blood gases but now with a considerably extended menu that includes electrolytes, glucose, creatinine, coagulation parameters, and cardiac markers. In contrast to the thick-film technology described above, the single-use sensors are constructed using *thin-film* technology. The latter involves microlithographic processes to deposit thin layers of 1 to 10 μm (as opposed to 10 to 50 μm layers in thick film technology), similar to the construction of computer chips and other silicon-based devices. Each single use i-STAT cartridge contains an array of electrochemical sensors for a set of particular analytes, and it is operated in conjunction with a handheld reading device. Because the sensor layer is very thin, analytes permeate to this layer quickly, and the sensor cartridge, once it reaches room temperature (if it has been refrigerated), can be used immediately. This is an advantage over some thick-film sensors that require an equilibration or "wet-up" time before they are used.

A disadvantage of thin-film technology is that manufacturing costs are high due to the need for special clean air facilities. Thus more recent technology to measure blood gases and related parameters utilizes less costly, so-called smart card technology. An example is the Epoc system (Epocal, Ottawa, Canada), which is used in conjunction with a handheld analyzer to provide a range of critical care tests, including immunoassay measurement. Fig. 17.3 shows two views of the Epoc test card, which is manufactured on a 35 mm tape on reel format; on the right is the top view of the test card showing the sample entry port, the sealed calibrator reservoir, and the sensor module at the top; on the left is the bottom view of the card with the sensor module, where measurements take place, and below this are details of the test panel.

Larger Cartridge Type and Bench-Top Point-of-Care Testing Devices

This section includes a wide range of devices that are all quantitative, not handheld but varying considerably in size. The smaller ones typically incorporate small, compact detectors, such as a charge-coupled device (CCD) camera that is a multichannel light detector, similar to a photomultiplier tube in a spectrophotometer. This type of technology enables quantitative measurements of lateral-flow immunoassay strips and is incorporated in the many different POCT devices available on the market.

Cartridge-type devices. A number of single-use, quantitative POCT devices are available that employ a cassette or cartridge design rather than lateral-flow strips. Several

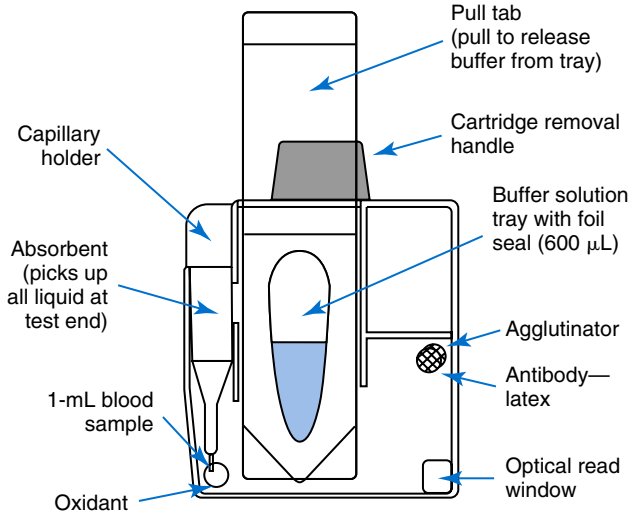


Fig. 17.4 Schematic diagram of the Bayer DCA 2000 HbA1C immunoassay cartridge. (Courtesy Siemens Medical Diagnostics.)

cassette-based systems have been developed for the measurement of hemoglobin. In one such system, red cells are lysed in a minicuvet, hemoglobin converted to methemoglobin, and the methemoglobin measured at 570 nm; turbidity is corrected for by an additional measurement at 880 nm.

Another type of cartridge design uses a light-scattering immunoassay to measure glycated hemoglobin, together with a photometric assay for total hemoglobin. The cartridge is a relatively complex structure that contains all the reagents required for the immunoassay, as well as lysing reagents and potassium ferricyanide reagent for measurement of total hemoglobin (Fig. 17.4). Agglutination takes place between the latex bound antibody and a synthetic polymer that contains the immunoreactive portion of HbA1C that leads to increased scattering of light. Addition of HbA1C in the patient sample reduces the agglutination by competing for the latex bound antibody, and the reduction in light scattering is proportional to the concentration of HbA1C.

Measurement takes place when the cartridge is placed into a temperature-controlled reader, and the analytical performance of the Siemens DCA device (Siemens Healthcare GmBH, Erlangen, Germany) is sufficient for quantitative monitoring of glycemic control, and can also be used for measurement of urinary albumin and creatinine. Following placement into the instrument, the analysis proceeds without further involvement of the operator. The most recent of these types of instruments is also considerably smaller, and that trend is likely to continue.

Multianalyte desktop analyzers. A typical example of these is the Piccolo (Abaxis), which can measure a wider range of general chemistry analytes on a very small sample volume. It uses a small disposable rotor that contains all the required reagents and diluents to perform a battery of related tests such as liver function tests.

The Radiometer AQT 90 (Radiometer Medical, DK-2700, Bronshoj, Denmark) bench-top, random access immunoassay analyzer utilizes a proprietary dry-chemistry concept and a detection method based on europium nonenhanced time-resolved fluorescence (TRF) technology. The europium-based chelates have certain natural advantages that allow sensitive and rapid detection methods. Analysis takes place in a small reagent cup on which are immobilized biotinylated capture antibodies, as well as streptavidin and europium-labeled tracer antibodies. Following addition of the patient sample and incubation, an antibody-antigen-antibody "sandwich" complex is formed, and after washing and drying, the europium response to excitation light is measured, the amount of which is directly proportional to the antigen present. The AQT 90 also has a unique, walk-away, closed-tube sampling device from which the instrument dispenses a sample into a cup containing dried reagents including antibodies to the analytes of interest and europium lanthanide chelate as the signal reagent. The menu includes troponin, brain natriuretic peptide (BNP), D-dimer, beta human chorionic gonadotropin (hCG), and C-reactive protein (CRP).

Critical care analyzers. These remain the most commonly used type of POCT bench-top analyzers; the expanded menu of analytes allows use in many clinical areas. These devices use the same type of thick-film sensors or electrodes in strips, as described above, but in this case the sensors are reusable. They are manufactured from thick films of paste and inks using screen-printing techniques to produce individual or multiple sensors for metabolites, blood gases, and electrolytes. Further details of the analytical principles behind these sensor technologies can be found in Chapter 10 of this textbook. The sensors have been incorporated with reagents and calibrators into a single cartridge or pack, which is placed in the body of a small- to medium-sized, portable critical care analyzer. Each pack contains reagents sufficient to test a certain number of samples during a specific time period, after which it is relatively simple to replace. An example of such a device is the GEM Premier 4000 (Instrumentation Laboratory, Bedford, MA; Fig. 17.5).

Other key developments in critical care instruments include liquid calibration systems that use a combination of aqueous base solutions and conductance measurements to calibrate the pH and PCO_2 electrodes, with oxygen being calibrated with an oxygen-free solution and room air. In addition, automated QC packages are integrated into these instruments to ensure that QC samples are analyzed at regular intervals. These comprise packs or bottles of QC material that are contained within the instrument and sampled at predetermined intervals with on-board software, interpreting the results and generating alerts, if necessary. Box 17.3 shows the various features that are incorporated into a critical care analyzer that contribute to its ease of use and reduce the potential for errors.

Hematologic and immunologic analyzers. Small bench-top devices are now available to perform hematologic and immunologic analysis at point of care. Again, they are often scaled-down versions of laboratory instruments but have been reduced in complexity, thus reducing the risk of operator error by non–technically trained staff. The range of hematology instruments include those that measure just hemoglobin

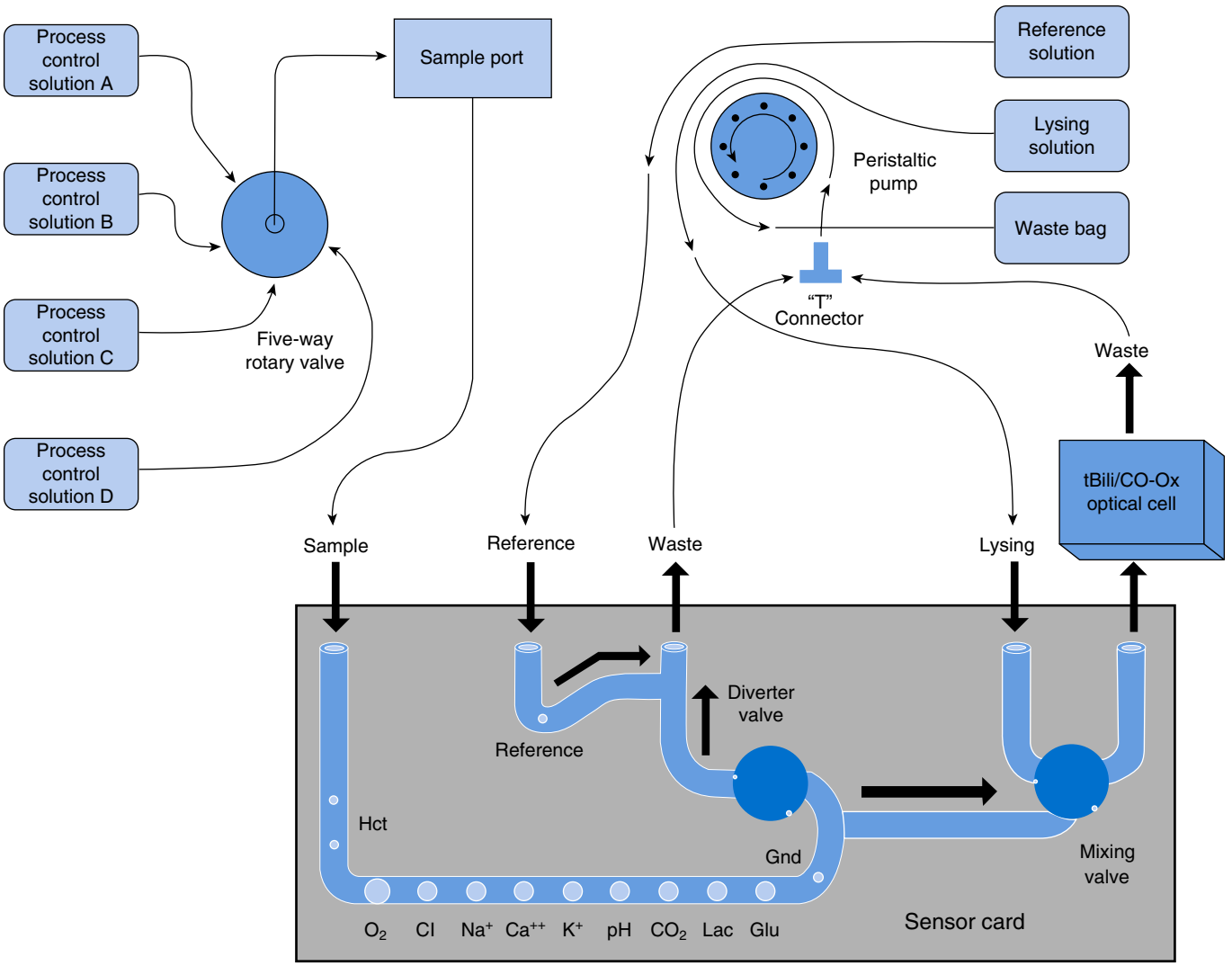


Fig. 17.5 Schematic diagram of the components and patient sample fluidic pathways of the GEM Premier 3000 critical care analyzer. (Courtesy Instrumentation Laboratory.)

BOX 17.3 Features of Critical Care Analyzers That Contribute to Ease-of-Use and Reduce the Risk of Errors

- Long-life, maintenance-free electrodes or disposable sensor packs
- Touch screens as the user interface
- Software that can demand user and patient identification
- Built-in bar code scanners
- Sample aspiration instead of injection
- Reduced sample sizes
- Clot detection within analysis chamber
- Sample detection to prevent short samples
- Liquid calibration systems instead of gas bottles
- Automated calibrations
- Automated quality control sampling
- Sophisticated QC programs including interpretation of data
- Connectivity to information systems allowing remote monitoring and control
- Built-in videos for training purposes

and white cell counts to the Sysmex pocH-100i (Sysmex, Lincolnshire, IL), which can measure up to 17 parameters.

Another relatively new product is the PIMA flow cytometer device (Alere) that can be used for measurement of T-helper cells, or CD4 counts. The PIMA uses the same imaging and cell counting technology employed in laboratory instruments but redesigned into a smaller, more compact instrument, which is simple to operate. Following mixing of the sample with the reagents and incubation, fluorescence is measured in the stained sample by a CCD camera and the results are displayed on the screen.

Newer and Emerging Point-of-Care Testing Technologies

Many laboratory technologies have been refined and packaged into smaller devices, and in many cases, there has been an improvement in analytical performance and a reduced risk of error. That trend in incremental improvements will continue, because no device is risk-free and in some cases there remains a need for better analytical performance. In addition,

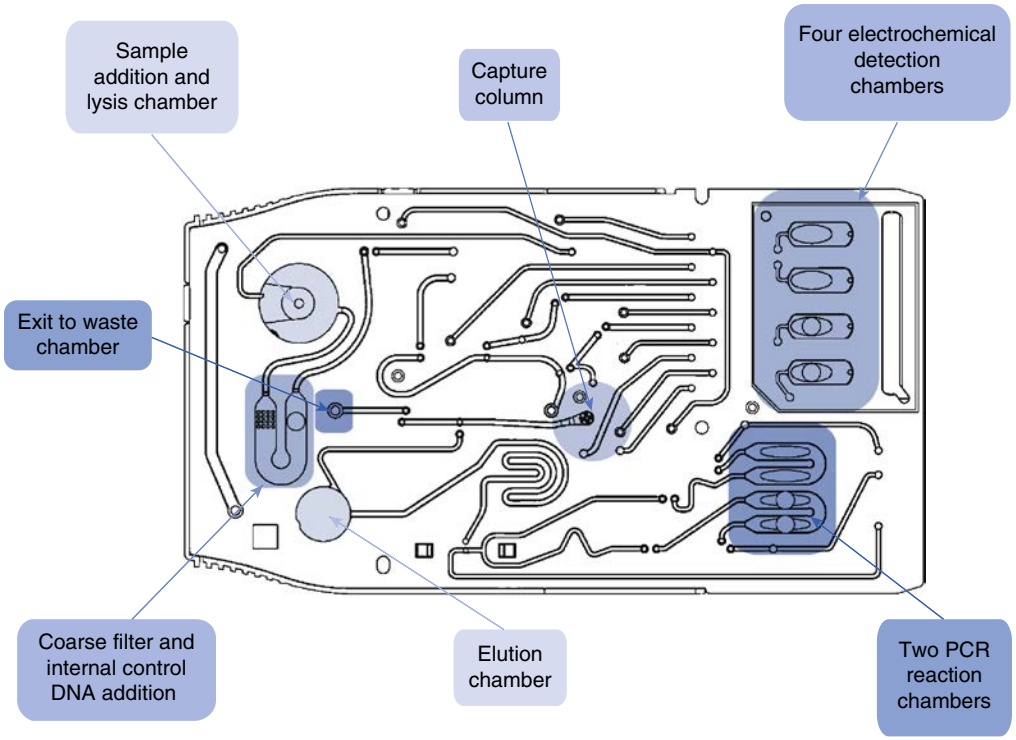


Fig. 17.6 Schematic diagram of a microfluidic card employing a molecular test for *Chlamydia trachomatis*. (Courtesy Atlas Genetics, Trowbridge, UK.)

the search for fundamentally different technologies continues because there are known limitations to existing technologies. For example, lateral flow strip technology continues to dominate the POCT market, but they have a number of limitations, including inadequate sensitivity and difficulty with multiplexing or measurement of more than two or three analytes on the same strip. An alternative measurement concept for the future is the so-called Lab on a Chip (LOC), also referred to as a microchip, adapted from the microelectronics industry through techniques of miniaturization and microfabrication. LOC devices have been defined as ones that perform analysis at microscopic scales (i.e., 1 to 500 μm) and incorporate microfilters, microchannels, microarrays, micropumps, microvalves, and bioelectronics chips.

Despite the continuing advances taking place in silicon chip manufacturing, these advances have not as yet translated into the development of many LOC devices for point-of-care use. One such LOC device is the Cepheid GeneXpert system (Cepheid, Sunnyvale, CA). This bench-top device can perform real-time quantitative polymerase chain reaction (PCR) in approximately 90 minutes with minimal operator interaction, thus enabling the potential to perform rapid molecular testing in situations where the need for results is urgent. Until recently the only technology available for point-of-care infectious disease testing such as influenza has been lateral flow strips with their previously described problems of limited sensitivity, which are overcome by PCR related techniques such as the Cepheid instrument. The system uses single-use cartridges that each contain multiple chambers to hold the sample, various purification and elution buffers, and all the PCR reagents, including enzymes; in addition, all waste is retained within the cartridge. Through a series of valves in the syringe, the sample is moved through the various stages of the PCR process using the reagents stored in the cartridge, and culminating in real-time detection of the amplified products.

Rapid molecular and other types of testing are also possible using technology developed by Atlas Genetics (Atlas Genetics, Trowbridge, Wilts, UK). The system comprises a test-specific disposable cartridge, containing all the necessary reagents to perform the test and shown in Fig. 17.6. Following addition of the specimen into the sample well of the cartridge and insertion of the cartridge into the Atlas io reader, the test is performed under complete instrument control with no further user input required. After sample addition, lysis of sample and internal control material, the mixture is pumped to the capture column where DNA is captured. Following PCR, the amplified mixture moves to the detection chambers for electrochemical measurement. The system is capable of multiplexing by virtue of a number of independent fluidic channels and the opportunity to use a number of electrochemical labels.

Informatics and Point-of-Care Testing

Most analytical devices used in clinical laboratories are directly linked or connected via an electronic interface to a laboratory information system (LIS). In this progression, many different **informatics** functions are used, including the electronic transfer of data from the analyzers to the LIS and ultimately into a patient's electronic medical record. This provides health care professionals with quick, accurate, and appropriate access to the patient's medical history and clinical information.

Considerable effort has been expended to incorporate these same processes into POCT devices because of the vital

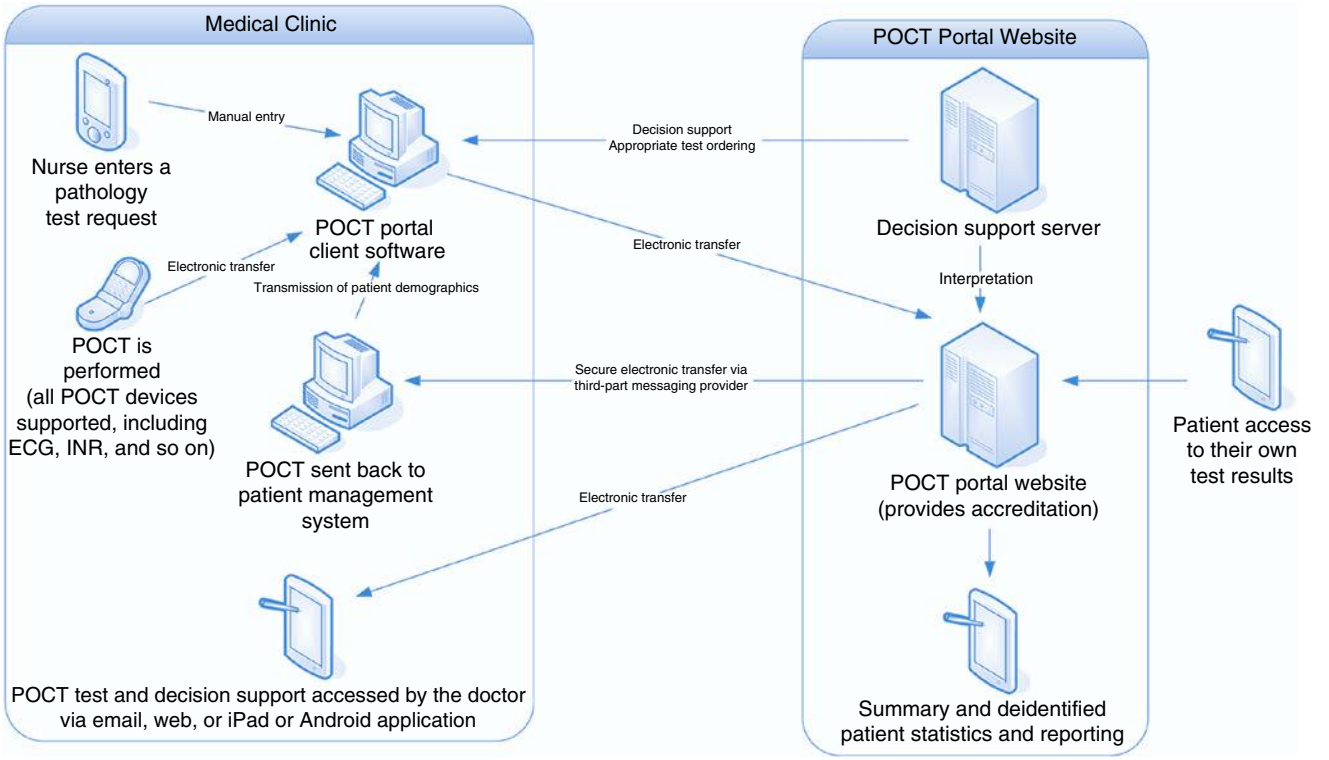


Fig. 17.7 Schematic diagram of a point-of-care testing (POCT) network showing the relationships between the user, POCT device, other health information systems, and the patient's caregiver. *INR,* International normalized ratio. (Courtesy R. Tirimacco, ICCNET.)

importance of capturing analytical data in a patient's medical record in order to avoid the errors associated with manual transcription of data. Thus newer POCT devices have addressed this problem by incorporating the prerequisite hardware and software into their design, and in the last decade so-called connectivity standards have also been introduced, which facilitate linking devices to information management systems. Adherence to these connectivity standards ensures that POCT devices meet critical user requirements, such as (1) bi-directionality, (2) device connection commonality, (3) commercial software interoperability, (4) security, and (5) QC and/or regulatory compliance.

Improved connectivity has enabled two major trends. First, and primarily due to demand from those responsible for the management of POCT, there has been the development of a range of commercial and specialized products for data management of POCT results with software devoted to ensuring compliance with procedures and managing data. These systems offer a range of features and benefits that include the ability of central laboratories to both monitor and remotely control their instruments in locations outside of the main laboratory.

The second trend is that connectivity standards, together with less expensive network technology and widespread access to the internet and smart mobile phones, have provided the capability to establish POCT networks across major secondary and tertiary care institutions and, perhaps more importantly, in the community or primary care where POCT has the potential to make a much bigger impact than in the hospital setting. The structure of such a typical community POCT network is shown in Fig. 17.7.

The most important benefit of connectivity remains that of facilitating the transfer and capture of patient POCT and quality-related data into permanent medical records. Additional benefits include that of providing alerts of abnormal results, particularly those that may be critical. However, the increasing ability to interface devices more easily to other information systems should also enable the development of applications, which add value to patient data such as the use of decision support software to assist with interpretation and clinical decision-making. One such example is the use of dosing software in combination with POCT INR measurements to manage warfarin treatment.

Points to remember—Point-of-Care Testing Technology

- POCT devices need to be simple to use and the reagents and consumables need to be robust over extended periods of time.
- Analytical results need to have an analytical quality appropriate for the clinical requirement.
- Glucose strips and lateral flow immunosensors remain the most common type of POCT technology.
- Newer technologies using molecular techniques are now starting to emerge.
- All devices should either have a means of data storage or be interfaced to an information system so that data can be stored, transmitted, and recalled.

BOX 17.4 Assessing the Need for a Point-of-Care Testing (POCT) Service

Which tests are required?
What is the turnaround time required?
What clinical question is being asked when requesting this test?
What clinical decision is likely to be made upon receipt of the result?
What action is likely to be taken upon receipt of the result?
What outcome should be expected from the action taken?
Why isn't the laboratory able to deliver the required service?
Will POCT provide the required accuracy and precision of result?
Are there staff available to perform the test?
Are there adequate facilities to perform the test and store the equipment and reagents?
Will you abide by the organization's POCT policy?
Are there operational benefits to this POCT strategy?
Are there economic benefits to this POCT strategy?
Will a change in practice be required to deliver these benefits?
Is it feasible to deliver the change in practices that might be required?

QUALITY MANAGEMENT

Management and maintenance of a POCT service requires planning, oversight, inventory control, and assurance of the reliability of test results through adequate training and QC. Several guidelines document these requirements, based on the ISO standard 22870 ("Point-of-Care Testing [POCT]—Requirements for Quality and Competence"), and are available from various professional laboratory organizations.

Establishment of Need

The decision to implement a POCT service requires (1) establishment of the unmet need, (clinical, operational, or economic); (2) critical appraisal of the evidence supporting the claimed clinical, operational, and economic benefits; and (3) examination of the costs and changes in the clinical process involved. Addressing the questions listed in Box 17.4 is useful for establishing the requirement for a POCT service.

When organizing and implementing a POCT service, it is important to consult with all the stakeholders, by establishing a POCT coordinating committee. Membership should include representatives of those who use the service and those who deliver the service, together with a representative of the organization's management team. The users may include (1) physicians, (2) physician assistants, (3) family nurse practitioners, (4) nurses, (5) other health care providers, and (6) patients. Typically a laboratory professional will chair such a committee, and it is also recommended that the committee report to the medical director of the hospital or other care provider organization. The work of the committee should be governed by the organization's policy on POCT.

BOX 17.5 Elements of a Point-of-Care Testing (POCT) Policy

Catalog Information—Review Time
- Approved by
- Original distribution
- Related policies
- Further information
- Policy replaces

Introduction—Background
- Definition
- Accreditation of services
- Audit of services

Laboratory Services in the Organization—Location
- Logistics
- Policy on diagnostic testing

Management of POCT—Committee and Accountability
- Officers
- Committee members
- Terms of reference
- Responsibilities
- Meetings

Equipment and Consumable Procurement—Criteria for Procurement
- Process of procurement

Standard Operating Procedures—Training
- Training and certification of staff
- Certification
- Recertification

Quality Control and Quality Assurance—Procedures
- Documentation and review

Health and Safety Procedures

Bibliography

Point-of-Care Testing Policy and Accountability

Implementation of a POCT service requires a POCT policy that establishes all of the procedures required to ensure the delivery of a high-quality service, together with the responsibility and accountability of all staff associated with the POCT. This may be (1) part of the organization's total **quality management** system, or (2) part of its clinical governance policy, and will be required for **accreditation** purposes. The elements of a POCT policy are listed in Box 17.5.

Equipment Procurement and Evaluation

After identifying a requirement for POCT, developing a POCT policy, and establishing a coordinating committee, the next stage in the process is equipment procurement. This involves identifying candidate POCT equipment having the prerequisite analytical and operational capabilities to meet the clinical requirements of the POCT service. A CLSI protocol (Clinical Laboratory Standards Institute, Document POCT09-P. Selection Criteria for Point of Care Testing Devices) describes

BOX 17.6 The Main Elements of a Point-of-Care Testing Training Program

Understanding the Context of the Test
- Clinical requirement for the test
- Action taken on basis of result
- Nature of test and method used

Patient Preparation Required
- Relevance of drug therapy

Sample Requirement and Specimen Collection

Preparation of Analytical Device—Machine and/or Consumables

Performance of Test

Performance of Quality Control

Documentation of Test Result and Quality Control Result

Reporting of Test Result to Appropriate Personnel

Interpretation of Result and Sources of Advice

Health and Safety Issues (e.g., disposal of sample and test device, cleaning of machine and test area)

the features of the evaluation process, and further guidance can be sought from other evaluations described in the literature. In addition, the educational and certification requirements of the device operator also have to be identified, and the potential for operator error determined. Independent validation of these analytical and operational characteristics is obtained from (1) the manufacturer, (2) published evaluations performed by government agencies, (3) reports in the peer-reviewed literature, and (4) discussions with colleagues.

Training, Competency, and Certification

Many of the agencies involved in the regulation of health care delivery now require that all personnel associated with the delivery of diagnostic test results demonstrate their competence, and this applies equally to operators of POCT equipment. The elements of a training program are listed in Box 17.6. Clearly the extent of this program will, to a certain extent, depend on how well the complexity of the analytical method has been engineered to minimize operator requirements. In practice, such a program is tailored to meet the needs of the individual and the organization.

Training may include formal presentations, self-directed learning, or computer-aided learning. It is important to document the satisfactory completion of training, and that the individual has been tested and found competent, with a combination of questions probing the understanding and practical demonstration of the skills gained. The latter is achieved by performing tests on a series of QC materials and repeat testing of patient samples that have recently been analyzed in the laboratory or on another device—so-called parallel testing.

Competence on a long-term basis is maintained through regular practice of skills and continuing education, and it is important to build these features into any POCT program. Regular review of performance in QC and quality assurance programs will provide a means of overseeing the competence of operators. The error log may also highlight when problems are arising. However, it is important to encourage an open approach to the assessment of competence so that operators themselves seek help if they believe that problems are occurring.

Quality Control, Quality Assurance, and Audit

Quality control, external quality assurance (EQA), or proficiency programs and **audits** provide a formal means of monitoring the quality of a service. The QC program gives a relatively short-term view and typically compares the current performance of a device with prior analyses. EQA is a more comprehensive process that includes comparing testing performance of different sites and/or different pieces of equipment or methods. Audit measures the performance of all aspects of the testing process against predetermined standards. However, the foundation to ensuring good quality remains a successful training and competency scheme.

Classically quantitative QC involves the analysis of a sample for which the analyte concentration is known, and the mean and range of acceptable results is quoted for the method used. There are several challenges to the classical approach with POCT. The first concern is the frequency of testing—should a QC sample be analyzed every time that (1) a sample is analyzed, (2) a new operator uses the system, (3) a new lot number of reagents is used, or (4) the system is recalibrated? There is no consistent agreement on the correct approach, and one probably has to be guided by the reproducibility and overall analytical performance of the system, but it should never be less than what is recommended by the manufacturer or what is necessary to meet regulatory requirements.

For single-use POCT disposable devices, the above strategy does not completely monitor the quality of the test system. For example, when conventional QC material is analyzed on a unit-use or single-test POCT system, only that testing unit is monitored. Under these circumstances, there is greater dependence placed on the manufacturing reproducibility of the devices to ensure a good quality service.

There are also other QC approaches, but many do not test the whole process. For example, the use of a plastic surrogate reflectance pad as a QC sample will only test the performance of the reflectance meter and so on. Similarly, some forms of electronic internal QC also do not test the sampling technique, but simply the functionality of the cassette and the docking station. As with conventional QC, any electronic process checks should not be less than those recommended by the manufacturer.

EQA, or proficiency testing, is a systematic approach to QC monitoring in which samples are analyzed by one or more laboratories to determine the capability of each participant. In this approach, the operator has no knowledge of the analyte concentration, and therefore it is considered closer to a "real testing situation." The results are transmitted to a

central authority, which then prepares a report and returns a copy to each participating laboratory. More details of quality management as applied to POCT can be found in Chapter 7 of this textbook.

Maintenance and Inventory Control

The implementation and maintenance of a POCT service require that a supply of devices and associated consumables be maintained at all times and that a formal program for doing so is employed. The key points in this process are to (1) adhere to the recommended storage conditions, (2) be aware of the stated shelf life of the consumables, and (3) ensure that stocks are released in time for any preanalytical preparation to be accommodated (e.g., thawing).

Clear guidelines should be available from the manufacturer and should be adhered to rigorously. Issues that usually require particular vigilance include expiration dates, biocontamination, electrical safety, maintenance of optics, and inadvertent use of inappropriate consumables.

Documentation

It is critically important to keep an accurate record of the (1) test request, (2) result, and (3) action taken, in the medical record. Some of the issues concerning documentation are now being resolved with the advent of the patient electronic record, electronic requesting, and better connectivity of POCT instrumentation to information systems and the patient record.

Accreditation and Regulation of Point-of-Care Testing

Regulatory and/or accrediting bodies of medical laboratories generally require compliance to minimum standards, and this applies equally to POCT services. Thus the Clinical Laboratory Improvement Amendments of 1988 (CLIA) legislation in the United States stipulates that all POCT must meet certain minimum standards. In the United States, the Centers for Medicare and Medicaid Services, the Joint Commission on Accreditation of Healthcare Organizations, and the College of American Pathologists are among those responsible for inspecting sites, and each is committed to ensuring compliance with testing regulations for POCT.

POINTS TO REMEMBER—QUALITY MANAGEMENT

- Any point-of-care testing (POCT) service should be based on satisfying a documented clinical or operational need.
- A POCT policy agreed by all stakeholders should determine how the service is run within an institution.
- All POCT equipment should be evaluated by those responsible for the service.
- Training and assessment of competency of POCT users is essential since many will not be trained in laboratory science.
- POCT needs to be accredited to the same levels as the laboratory.

EVIDENCE OF EFFECTIVENESS

The objective of employing POCT is to enable clinical decisions to be made at the time of the consultation with the patient, where the decisions may relate to screening, diagnosis, treatment, or monitoring. The outcomes of these decisions should lead to clinical, operational, and/or economic benefits if implementation is to be considered. Typically those involved in the management decisions to implement POCT will look for evidence of clinical and cost **effectiveness**; those involved in implementation will, in addition, look for evidence of operational effectiveness, as that evidence will inform the implementation plan. It follows therefore that the evidence of effectiveness is employed in business cases proposing the adoption of POCT, and that it informs the implementation plan. The needs of all stakeholders involved in delivering the care package must be addressed. This can be challenging, as stakeholder perspectives may differ across a care pathway, for example, the current desire to shift care closer to home may have an adverse impact on the revenue received by the hospital.

Clinical Effectiveness

The core outcome measures of clinical effectiveness are mortality and morbidity. Surrogate or intermediate outcome measures are generally used for evidence of effectiveness in both the generation of evidence, as well as in performance management and audit. Some examples of outcome measures are given in Table 17.2. The generation of evidence of clinical effectiveness in the field of diagnostics including POCT is challenging, primarily because (1) the timely delivery of a result in itself does not lead to improved outcomes (i.e., outcomes are dependent on the clinical management decision based on the test result), (2) variability can occur with both the decisions made and the action taken on receipt of the result, and (3) it is difficult to design studies that minimize the generally accepted risks of introducing bias (e.g., operator bias; see Chapter 2).

Examples of tests where there is good evidence for the clinical effectiveness of POCT include (1) rapid diagnosis of *Chlamydia* infection, (2) managing warfarin treatment with INR self-monitoring, (3) ruling out a diagnosis of deep vein thrombosis, and (4) ,managing patients with diabetes using HbA1C.

Cost-Effectiveness

Economic assessment is becoming more important because of the growing debate around "value-for-money" in health care, and consequently health care economics is seen as an integral part of evidence-based practice. An economic assessment of POCT should ask the question: What are the complete consequences and costs of using POCT compared with testing done in the laboratory? The consequences would include any operational changes involved with introducing POCT—particularly in the clinical setting, in addition to the impact on clinical outcomes. The reality is that relatively few economic assessments have been conducted in this way for laboratory tests and even fewer for POCT.

TABLE 17.2 Outcome Measures That Can Be Used to Assess Point-of-Care Testing Technologies

OUTCOME MEASURE			
Clinical	**Process**	**Application**	**Test**
Reduce antibiotic use	Reduce the need for laboratory tests	Rule out UTI	Urinalysis
Early detection of renal disease	Advise the patient at the clinic visit	Rule out albuminuria	Urine albumin creatinine ratio
Faster and earlier diagnosis	Reduced clinic visits Use of echocardiography	Rule out a diagnosis of heart failure	Natriuretic peptide
Reduced complication rate	Reduced nonattendance at the clinic Reduced clinic visits	Screening for infection	*Chlamydia*
Increased period within the therapeutic window Reduced complication rate	Reduced clinic visit Reduced hospitalization	Self-monitoring Self-dose adjustment	INR
Improved glycemic control indicated by HbA1C	Reduced clinic visits	Primary care management of diabetes	HbA1C in primary care
Faster triage to therapeutic intervention	Reduced length of stay in ED	Rule out acute coronary syndrome	Troponin I or T in ED
Faster diagnosis and treatment	Reduced clinic visits	Rule out DVT	D-dimer in primary care

DVT, Deep venous thrombosis; *ED*, emergency department; *INR*, international normalized ratio; *UTI*, urinary tract infection.

Instead of robust economic studies, there has been a focus in the past on simplistic comparisons of the cost of POCT technology to that used in the central laboratory without any consideration of the patient outcomes from using the tests. Thus many studies in the literature are of the cost minimization type and not surprisingly have shown that POCT is more expensive than the central laboratory, where economies of scale can deliver obvious efficiencies. However, this situation is changing to one that moves beyond a comparison of testing costs to one that includes assessment of the complete process of patient care and identification of the economic outcomes that can be achieved with POCT. Some of the economic benefits may result from a clinical impact upon the patient, but others may be operational benefits such as more rapid treatment, or reduction in length of stay.

POINTS TO REMEMBER—EFFECTIVENESS OF POINT-OF-CARE TESTING

- Point-of-care testing (POCT) may offer increased operational efficiency, for example, more rapid patient triage in emergency department.
- POCT may be associated with improved clinical outcomes, for example, reduced morbidity and mortality for home INR monitoring of coumadin treatment.
- POCT may offer increased patient accessibility and convenience, for example, testing and treatment in one visit for sexually transmitted diseases.

Consideration of the cost effectiveness must examine the impact of POCT on the entire care pathway, including patient outcomes.

REVIEW QUESTIONS

1. Which of the following statements best describes the value of point-of-care testing (POCT)?
   - **A.** Testing performed outside of the laboratory
   - **B.** Testing performed at the bedside
   - **C.** Testing that enables the clinician or caregiver to make a decision at the time the clinical problem arises
   - **D.** Testing near to the patient
2. The benefits of using POCT are best assessed by:
   - **A.** comparing the analytical results of the device with those obtained by the central laboratory.
   - **B.** performing a study of clinical and cost effectiveness.
   - **C.** assessing the performance of the device in an external quality assurance program.
   - **D.** asking the patients what they think.
3. The technical performance of a POCT device is best assessed by:
   - **A.** comparing the results obtained by the person(s) who will be operating the device in the intended setting.
   - **B.** reviewing the manufacturer's literature.
   - **C.** testing the device in the laboratory.
   - **D.** looking for information on the Internet.
4. Quality control of POCT for lateral flow strips read with a reflectance meter is best achieved by:
   - **A.** analysis of a designated QC specimen on at least one occasion in each shift.
   - **B.** use of the reflectance control strip supplied by the manufacturer.

C. comparison of analysis of a specimen using the POCT with the results from sending the same specimen to the laboratory.
D. analysis of an aqueous solution prepared for the purpose.
E. reliance on the manufacturer to provide a quality controlled product.

5. The first step to be considered in the implementation of a POCT solution is:
A. the analytical imprecision of the POCT device.
B. the analytical bias of the POCT device.
C. the unmet clinical need.
D. the cost of each test.

SUGGESTED READINGS

Adams, E. J., Ehrlich, A., Turner, K. M., Shah, K., Macleod, J., Goldenberg, S., et al. (2014). Mapping patient pathways and estimating resource use for point of care versus standard testing and treatment of Chlamydia and gonorrhoea in genitourinary medicine clinics in the UK. *BMJ Open*, *4*, e005322.

Attia, U. M., Marson, S., & Alcock, J. R. (2009). Micro-injection molding of polymer microfluid devices. *Microfluidics and Nanofluidics*, *7*, 1–29.

Clinical Laboratory Standards Institute. (2006). *Clinical Laboratory Standards Institute. Point-of-Care Connectivity: Approved Standard – Second Edition. CLSI Document POCT01–A2*. Wayne, Pa, USA: CLSI.

International Organization for Standardization. ISO 22870:2006. (2009). *Point-of-care testing (POCT) – Requirements for quality and competence. Clinical Laboratory Standards Institute. Selection Criteria for Point of Care Testing Devices, CLSI Document POCT09-P*. Wayne, Pa, USA: CLSI.

International Organization for Standardization, 2015 International Organization for Standardization. ISO 22870:2006. Point-of-care testing (POCT) – Requirements for quality and competence. http://www.iso.org/iso/catalogue_detail.htm?csnumber=35173 (Accessed April 2015).

Kost, G. J. (Ed.). (2002). *Principles and practice of point-of-care testing*. Philadelphia: Lippincott Williams & Wilkins. several chapters.

Oliver, N. S., Toumazou, C., Cass, A. E., & Johnston, D. G. (2009). Glucose sensors: a review of current and emerging technology. *Diabetic medicine: a journal of the British Diabetic Association*, *26*, 197–210.

Peeling, R. W., Holmes, K. K., Mabey, D., & Ronald, A. (2006). Rapid tests for sexually transmitted infections (STIs): the way forward. *Sexually transmitted infections*, *82*(suppl 5), v1–v6.

Posthuma-Trumpie, G. A., Korf, J., & van Amerongen, A. (2009). Lateral flow (immuno)assay: its strengths, weaknesses, opportunities and threats. A literature survey. *Analytical and Bioanalytical Chemistry*, *393*, 569–582.

Price, C. P., & St John, A. (2012). *Point-of-care testing. Making innovation work for patient-centred care*. Washington, DC: AACC Press.

Price, C. P., St John, A., & Kricka, L. J. (Eds.). (2010). *Point-of-Care Testing Needs, Opportunities and Innovation* (3rd ed) Washington, DC: AACC Press. several chapters.

Raja, S., Ching, J., Xi, L., Hughes, S. J., Chang, R., Wong, W., et al. (2005). Technology for automated, rapid, and quantitative PCR or reverse transcription-PCR clinical testing. *Clin Chem*, *51*, 882–890.

St John, A., & Price, C. P. (2013). Economic evidence and point-of-care testing. *The Clinical Biochemist Reviews / Australian Association of Clinical Biochemists*, *34*, 61–74.

Tonyushkina, K., & Nichols, J. H. (2009). Glucose meters: a review of technical challenges to obtaining accurate results. *J Diabetes Sci Technol*, *3*, 971–980.

Turner, A. P. F. (1987). In I. Karube, & G. S. Wilson (Eds.), *Biosensors: fundamentals and applications* (pp. 1–770). Oxford: Oxford University Press.

US Department of Health and Human Services. (1992). Medicare, Medicaid and CLIA programs: regulations implementing the clinical laboratory improvement amendments of 1988 (CLIA). *Final rule. Federal Register*, *57*, 7002–7018.

US Department of Health and Human Services. (1993). Medicare, Medicaid and CLIA programs: regulations implementing the clinical laboratory improvement amendments of 1988 (CLIA) and clinical laboratory act program fee collection. *Federal Register*, *58*, 5215–5237.

Valkirs, G. E., & Barton, R. (1985). ImmunoConcentration–a new format for solid-phase immunoassays. *Clin Chem*, *31*, 1427–1431.

Walter, B. (1983). Dry Reagent Chemistries. *Analytical Chemistry*, *55*, 498A–514A.

Wong, R., & Tse, H. (2009). *Lateral Flow Immunoassay*. New York: Humana Press, 223 pp.

18

Amino Acids, Peptides, and Proteins

Dennis J. Dietzen

OBJECTIVES

1. Define the following:
   - Acute-phase response
   - Bence-Jones proteins
   - Complement
   - Essential amino acid
   - Immunoglobulin
   - Isoelectric point
   - Paraprotein
   - Peptide bond
2. Diagram and describe the basic chemical structure of an amino acid.
3. Diagram the formation of a peptide bond from two adjacent amino acids.
4. Describe amino acid and protein analyses, including specimen requirements and quantitative analytical techniques.
5. Describe how amino acid carbon contributes to energy production.
6. Match the peptide(s) to the physiologic function:
   Peptide:
   - Hepcidin
   - Natriuretic peptides
   - Angiotensins
   - Vasopressin

   Function:
   - Promote Na retention and increased blood pressure
   - Inhibit iron mobilization
   - Promote water retention in collecting duct
   - Promote Na and water excretion
7. Explain the principle of serum protein electrophoresis, including a description of staining methods, and immunofixation.
8. Compare plasma and serum with regard to protein concentration.
9. List the acute-phase response proteins and the negative acute-phase response proteins.
10. List the various classes of protease molecules and representative members of each class.
11. List the various protein posttranslational modifications, and cite specific examples of proteins modified by each of the modifications.
12. Discuss the complement proteins, including their clinical significance, abundance, pathway involvement, and disease associations.
13. List the immunoglobulin classes. Describe their biochemistry, function, and clinical significance, and explain how they are affected by disease.
14. Discuss primary, secondary, tertiary, and quaternary protein structure.
15. Explain the presence of protein in cerebral spinal fluid (CSF) and the significance of increased CSF protein.

KEY WORDS AND DEFINITIONS

Acute-phase response Body's response to injury or inflammation.

Amino acid An organic compound containing both amino ($-NH_2$) and carboxyl ($-COOH$) functional groups.

Aminoaciduria An excess of amino acids in the urine.

Bence-Jones proteins Small light chains of immunoglobulin found in the urine.

Complement system Complex system of proteins found in blood that combines with antibodies to destroy pathogenic bacteria and other foreign cells.

Essential amino acids Amino acids that are not synthesized by humans and therefore are essential dietary constituents for maintaining health or growth.

Immunoglobulins A family of proteins also known as *antibodies* that contain highly specific antigen-binding

sites consisting of two identical heavy (H) chains encoded on chromosome 14 and two identical light (L) chains encoded on chromosome 2.

Isoelectric focusing An equilibrium technique that is used to separate charge variants of proteins and is applied to the analysis of certain genetic variants of proteins.

Kwashiorkor A form of protein-energy malnutrition produced by severe protein deficiency.

Marasmus A form of protein-energy malnutrition predominantly due to prolonged severe caloric deficit.

Monoclonal gammopathy of undetermined significance (MGUS) A condition in which a paraprotein is found in an individual's blood.

Multiple myeloma A cancer in which antibody-producing plasma cells grow in an uncontrolled and malignant manner.

Paraprotein A monoclonal immunoglobulin produced in excessive amounts in disorders such as multiple myeloma.

Peptide A compound consisting of two or more amino acids linked in a chain via peptide bonds.

Peptide bond The amide bond formed between the carboxyl group of one amino acid and the amino group of another.

Protein A polymer of amino acids linked by peptide bonds with a specific sequence that folds into a defined structure; any of a group of complex organic compounds that contain carbon, hydrogen, oxygen, nitrogen, and usually sulfur (the characteristic element being nitrogen).

Proteome The total complement of proteins expressed by the genetic material of an organism under a given set of environmental conditions.

Amino acids, peptides, and proteins are crucial for virtually all biologic processes. **Amino acids** serve as structural subunits of peptides and proteins but also play diverse roles in metabolism, neurotransmission, and intercellular signaling. **Peptides** serve as autocrine and endocrine signaling molecules that control appetite, vascular tone, and electrolyte homeostasis, as well as carbohydrate and mineral metabolism. **Proteins**, longer peptide chains with molecular mass typically greater than approximately 6000 Da, serve as (1) intracellular and extracellular structural components, (2) biologic catalysts, (3) mediators of contractility and motility, (4) agents of molecular assembly, (5) ion channels and pumps, (6) molecular transporters, (7) mediators of immunity, and (8) components of intracellular and intercellular signaling networks.

The human genome contains more than 20,000 open reading frames that encode proteins. The structural and functional complexity of these proteins is enhanced by alternative splicing of messenger RNA (mRNA), somatic recombination, mutation, proteolytic processing, and posttranslational modification. The **proteome** represents the complete set of proteins in an organism or compartment of an organism such as the plasma space. Diagnostic exploitation of the proteome promises to improve the detection and treatment of many pathologic conditions—most notably cancer.

This chapter begins with a discussion of the chemistry, metabolism, and analysis of amino acids. Discussion of the peptide bond and methods for measuring peptides follows. The protein narrative begins with an account of protein structure and the role of co- and posttranslational modifications in protein function. Important constituents of the proteome in various body fluids are then defined followed lastly by a description of methods for specific and global assessment of the proteins for clinical purposes.

AMINO ACIDS

Amino acids were likely among the first organic molecules to emerge from the mix of methane, hydrogen, ammonia, and water in earth's primordial atmosphere. Only 20 of the hundreds of known amino acids account for the vast majority of residues that make up human proteins. Their structure and molecular properties are summarized in Table 18.1. These 20, along with dozens of non–protein-forming amino acids, are critical to the form and function of the human body. Disrupted amino acid metabolism is not surprisingly associated with a multitude of pathologic processes.

Basic Biochemistry

Amino acids are organic compounds containing both an amino group ($-NH_2$) and a carboxyl group (–COOH), or another acidic group such as a sulfonate group ($-SO_3$). In a majority of biologically relevant amino acids, the amine moiety is primary ($-NH_2$), but some (e.g., sarcosine) are secondary (–NH–) amines, and others containing tertiary amines (e.g., proline) are referred to as imino (=N–) acids. Amino acids that occur in protein are referred to as α-amino acids (Fig. 18.1) because the amine moiety is linked to the α-carbon atom. The α carbon atom is directly adjacent to the carboxyl group, while carbon atoms successively distal to the carboxyl are referred to as the β- and γ-carbon, respectively. Not all biologic amino acids are α-amino acids. β-amino acids (β-alanine and taurine) as well as γ-amino acids (γ-aminobutyric acid, GABA) also play key biochemical roles. Some of these atypical amino acids are shown in Fig. 18.2.

With the exception of glycine, all α amino acids contain four distinct moieties asymmetrically arranged around the central "α" carbon. As a consequence, amino acids exist as mirror images (enantiomers) referred to as the *D* or *L* configuration. With few exceptions, the biologically relevant amino acids exist in the *L*

TABLE 18.1 Structure and Chemical Properties of the 20 Proteogenic Amino Acids

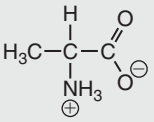
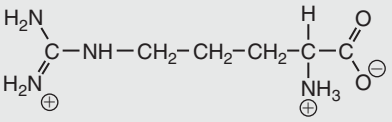
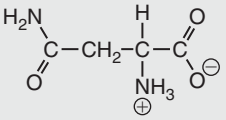
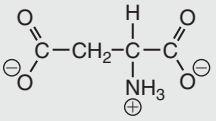
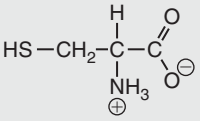
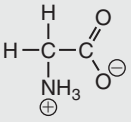
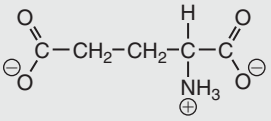
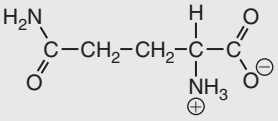
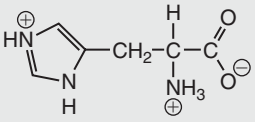
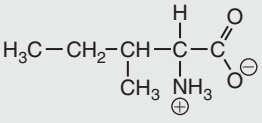
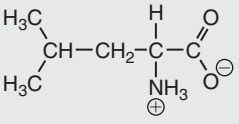
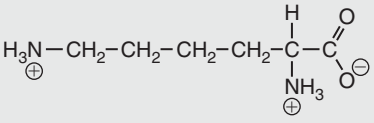
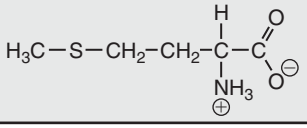
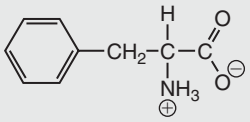
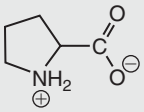
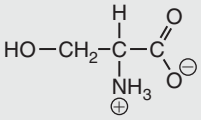
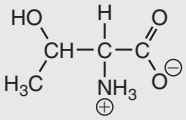
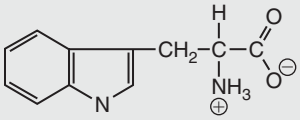
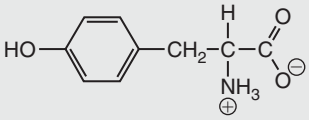
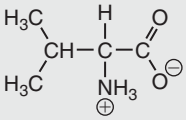
Amino Acid	MW (Da)	Structure (pH 7.0)	pK_1	pK_2	pK_3	pI	HI
Alanine (ALA, A)	89.09		2.4	9.7		6.0	1.8
Arginine (ARG, R)	174.20		2.2	9.0	12.5	10.8	−4.5
Asparagine (ASN, N)	132.12		2.0	8.8		5.4	−3.5
Aspartate (ASP, D)	133.10		2.1	9.8	3.9	2.9	−3.5
Cysteine (CYS, C)	121.16		1.7	10.8	8.3	5.1	2.5
Glycine (GLY, G)	75.07		2.3	9.6		6.0	−0.4
Glutamate (GLU, E)	147.13		2.2	9.7	4.3	3.2	−3.5
Glutamine (GLN, Q)	146.15		2.2	9.1		5.7	−3.5
Histidine (HIS, H)	155.16		1.8	9.2	6.0	7.6	−3.2
Isoleucine (ILE, I)	131.17		2.4	9.7		6.0	4.5
Leucine(LEU, L)	131.17		2.4	9.6		6.0	3.8
Lysine (LYS, K)	146.19		2.2	9.0	10.5	9.7	−3.9
Methionine (MET, M)	149.21		2.3	9.2		5.8	1.9

TABLE 18.1 Structure and Chemical Properties of the 20 Proteogenic Amino Acids—cont'd

Amino Acid	MW (Da)	Structure (pH 7.0)	pK_1	pK_2	pK_3	pI	HI
Phenylalanine (PHE, F)	165.19		1.8	9.1		5.5	2.8
Proline (PRO, P)	115.13		2.1	10.6		6.1	1.6
Serine (SER, S)	105.09		2.2	9.2		5.7	−0.8
Threonine (THR, T)	119.12		2.6	10.4		6.5	−0.7
Tryptophan (TRP, W)	201.22		2.5	9.4		5.9	−0.9
Tyrosine (TYR, Y)	181.19		2.2	9.2	10.5	5.7	−1.3
Valine (VAL, V)	117.17		2.3	9.6		6.0	4.2

Three-letter and single-letter definitions of the amino acids are included in parentheses.
HI, Hydropathy index; *MW,* molecular weight; *pk,* acid ionization constant; *pI,* isoelectric point.

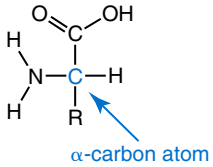


Fig. 18.1 Planar representation of an α-amino acid. "R" represents unique side chain specific to each amino acid.

configuration. Small quantities of *D* amino acids occur in physiological fluids but typically do not have specific functions. An exception is *D* serine, which represents 5% to 20% of total serine in cerebrospinal fluid and may serve as a neurotransmitter. Amino acids with the *D* configuration occur in some bacterial products, foods, and pharmaceuticals but are metabolized in the liver and kidney. Two amino acids, threonine and isoleucine, have a second asymmetric carbon that gives rise to stereoisomers referred to as *allothreonine* and *alloisoleucine.* The latter compound has utility in the diagnosis of maple syrup urine disease.

In addition to the 20 protein-forming amino acids, a number of rare amino acids are recovered from protein hydrolysates. For example, 4-hydroxyproline and 5-hydroxylysine are found in collagen. Desmosine and isodesmosine are recovered from elastin. These amino acids are formed by posttranslational mechanisms because no codon is responsible for their incorporation into growing polypeptides. Selenocysteine is a rare amino acid incorporated into just a few protein sequences (e.g., iodothyronine deiodinase). The structures of some of these rare amino acids are included in Fig. 18.2.

Acid–base properties of amino acids depend on the amino and carboxyl groups attached to the α carbon and on basic or acidic groups that may occur on side chains (R). At pH 7.4, the α-carboxyl group is ionized and carries a negative charge, while the α amino group is protonated and carries a positive charge. Molecules existing simultaneously as cations and anions are referred to as zwitterions (Fig. 18.3).

The pH at which ionizable groups exist equally as charged and uncharged forms is known as the pK. Amino acids thus have two or more pKs—one for the carboxyl, one for the amino group, and an additional one in the presence of an ionizable side chain. The isoelectric point (pI) is the pH at which an amino acid or other molecule has a net charge of 0. For a typical neutral amino acid such as glycine, the pI of 5.97 is midway between the pK of 2.34 for the carboxylic acid and the pK of 9.60 for the amino group. The pKs of amino acid side chains in proteins vary somewhat from those in free amino acids because of the influence of neighboring amino

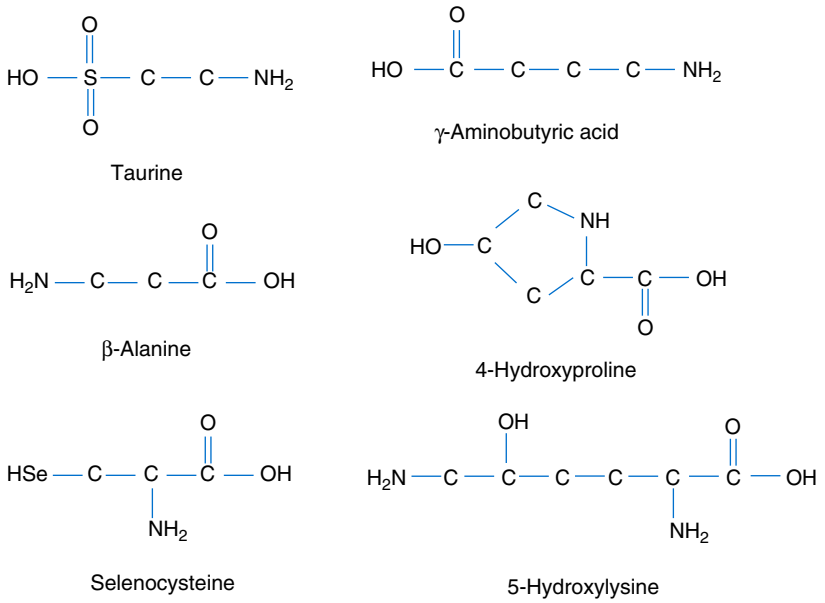


Fig. 18.2 Planar structures of rare or unusual, naturally occurring amino acids.

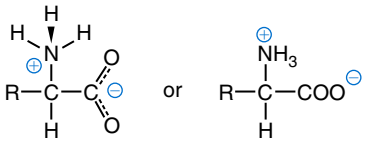


Fig. 18.3 Planar schematic of an α-amino acid as a zwitterion.

acids. The buffering capacity of ionizable groups is primarily in a pH range within ±1 unit of the pK for the respective groups. Amino acids and proteins therefore have a limited buffering capacity near physiologic pH. The imidazole side chain of histidine is an exception, with a pK near 6.0.

The structural diversity of side chains permits formation of proteins with a variety of structure and function. Side chain diversity is dictated not only by pK but by size and hydrophobicity as well. Amino acids with longer aliphatic or aromatic side chains such as isoleucine, leucine, and phenylalanine have greater hydrophobicity than shorter side chains, such as the methyl group found in alanine. Neutral amino acids with polar groups such as hydroxyl or amide groups in their side chains are more hydrophilic. Acidic amino acids have side chains with carboxylic acids, and basic amino acids have side chains with amino, guanidino, or imidazole groups. The thiol side chain (–SH) of cysteine oxidizes easily and may become linked to other molecules via a disulfide bond. In plasma, cysteine occurs as cystine (cysteine homodimer linked via a disulfide) or as a mixed disulfide with albumin or other proteins.

With some exceptions, amino acids are water soluble and stable in plasma. The most soluble amino acids have small side chains with polar or ionizable moieties such as glycine, alanine, arginine, serine, and threonine. Less soluble amino acids such as phenylalanine, tyrosine, leucine, and tryptophan tend to have larger, nonpolar aliphatic or alicyclic side chains. Amino acid solubility is rarely limiting in vivo, except in some metabolic disorders. Deposition of tyrosine crystals in the eye and skin is common in tyrosinemia. Likewise, cystine may crystallize in the renal parenchyma in patients with cystinuria.

Amino Acid Supply and Transport

In addition to serving as substrates for protein synthesis, amino acids participate in many metabolic pathways. In the healthy state, women require approximately 46 g/day and men approximately 56 g/day of dietary protein (0.8 g/kg body weight), and substantial increases in demand occur during growth and in many disease states. Dietary protein is digested by proteases in the stomach (e.g., pepsin) and small intestine (e.g., trypsin, chymotrypsin) to yield amino acids. Endogenous protein turnover serves as another source of free amino acids. Eight amino acids used for protein synthesis (isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, and valine) are not synthesized by humans and therefore are considered "essential" constituents of the diet. Meat, milk, eggs, and fish contain a full range of **essential amino acids**. Gelatin is deficient in tryptophan, and some plant sources of protein may be additionally deficient in lysine or methionine. Therefore diets based on a single source of plant protein may be deficient in some amino acids. When liver function is compromised, cysteine and tyrosine become essential because they are not converted from their usual precursors, methionine and phenylalanine. Arginine may be conditionally essential as well, because endogenous rates of synthesis may be insufficient to meet requirements in adults under metabolic stress or in growing children.

Requirements for dietary protein to maintain nitrogen balance increase in infancy and childhood when there are increased demands for growth. Daily requirements increase by up to 3.5 to 4 g protein/kg body weight for premature infants, for example. Protein demand is also increased in pregnancy, lactation, and states of protein loss or catabolic states (e.g., burn patients). Persistent negative nitrogen balance results in a number of undesirable phenotypic features.

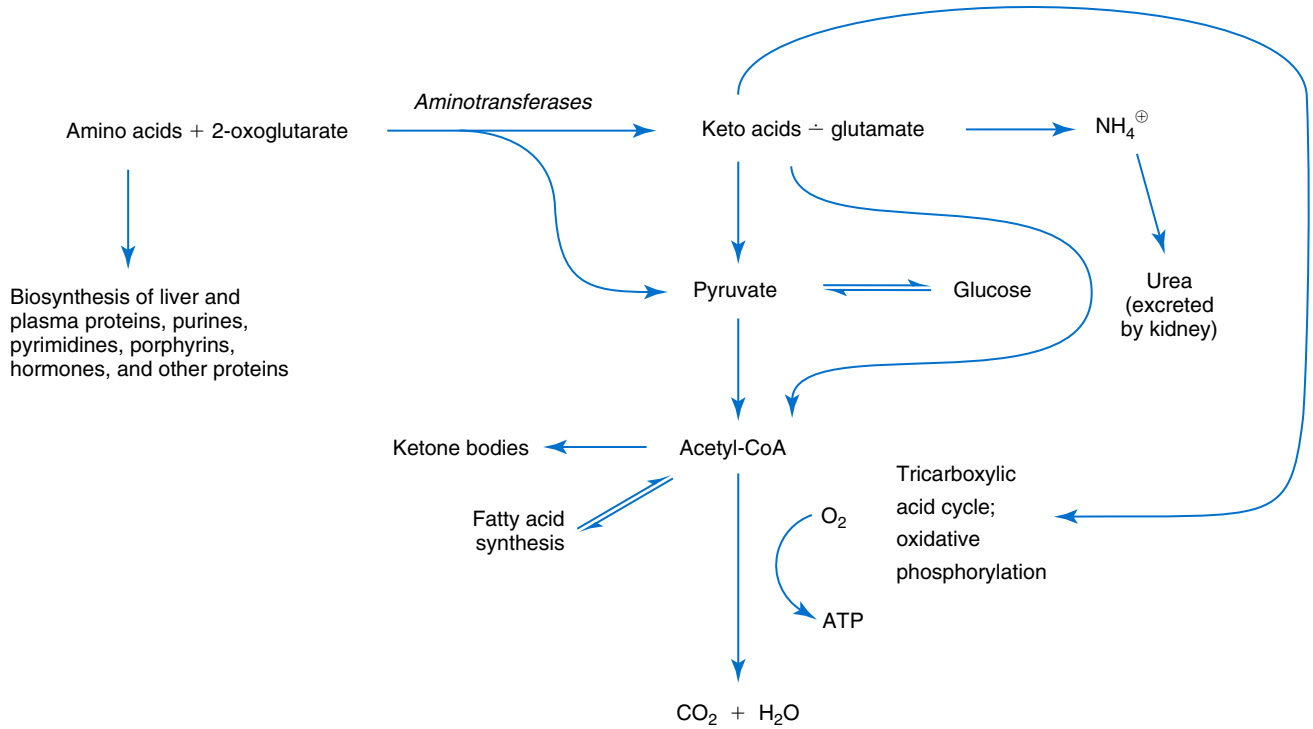


Fig. 18.4 A generalized scheme of amino acid metabolism in the liver. After transamination, amino acid carbon may be used in the Krebs cycle or transformed into respiratory fuels such as glucose and ketones. Waste nitrogen is disposed of via the urea cycle. *ATP,* Adenosine triphosphate; *CoA,* coenzyme A.

A diet severely deficient in protein and consisting primarily of high-starch foods can lead to **kwashiorkor**, a disorder characterized by decreased serum albumin, immune deficiency, edema, ascites, growth failure, apathy, and many other symptoms. **Marasmus** results when protein and energy sources such as carbohydrates are deficient, causing wasting of muscles and subcutaneous tissues. Albumin or prealbumin concentrations can be used to assess adequacy of the amino acid supply. The shorter biologic half-life of prealbumin compared with albumin (2 vs. 20 days) makes it a valuable marker for acute dietary assessment.

Homeostasis of cellular amino acid concentrations is dependent on supply, catabolism, and excretion. Supply and excretion are regulated by a series of transport systems with overlapping substrate specificity, strategic tissue expression, and polarized cellular distribution. Amino acids are derived from dietary protein precursors through the action of proteolytic enzymes in the stomach and small intestine that produce shorter oligopeptides and individual amino acids. Enteral absorption of di- and tripeptides is mediated by a single proton-coupled transport system termed PEPT1 (encoded by *SLC15A1*). Transport of individual amino acids across the intestinal and renal epithelium as well as the blood–brain barrier is far more specialized.

A diverse array of transport systems consisting of multiple polypeptide chains has evolved for both inter- and intracellular amino acid transport. Some transporters require energy, while others rely on established concentration gradients. The substrate specificity of these systems is often broad. For example, distinct systems are responsible for transporting branched-chain, cationic, anionic, or hydrophobic amino acids. Multiple different systems often cooperate to achieve amino acid absorption across the gut or renal epithelia. Defects in these transport systems are responsible for a number of disease states including cystinuria, cystinosis, lysinuric protein intolerance, or HHH (hyperornithinemia, hyperammonemia, homocitrullinuria) syndrome.

Amino Acid Metabolism

Amino acids serve as scaffolds for the synthesis of many hormones, nucleotides, lipids, signaling molecules, and metabolic intermediates that play a role in energy production. As portrayed in Fig. 18.4, the transformation of amino acid carbon to energetic intermediate typically begins with transamination. Excess nitrogen is ultimately excreted as urea. Resulting α-ketoacids may enter the Krebs cycle; undergo conversion to ketone bodies, fatty acids, or glucose; or be completely oxidized to CO_2, depending on cellular energy demands. A vast array of enzyme networks has evolved to orchestrate demand for amino acids.

Glucose, fatty acids, and ketones are primary respiratory substrates in humans. These substrates generate adenosine triphosphate (ATP) via the mitochondrial Krebs cycle. Amino acids play two key roles in energy provision. First, amino acids are converted to Krebs cycle intermediates to maintain the activity of the cycle in a process called anaplerosis. Glutamine and glutamate, for example, are converted to α-ketoglutarate. Fumarate may be derived from asparagine and aspartate, and succinate is derived from methionine, threonine, and valine. Second, amino acids may be mobilized

to generate fuels for a variety of organ systems. Five amino acids (LEU, ILE, LYS, PHE, and TYR) may be converted to ketones. All of the amino acids except for leucine may be used to produce glucose. In times of high energy demand and limiting fuel sources, therefore, flux of amino acid carbon through proper pathways becomes an important source of respiratory fuel.

Excess tissue nitrogen is disposed of as urea, which contains two moles of nitrogen per mole. Urea production is limited to the liver, so selected amino acids, primarily glutamine and alanine, serve to shuttle excess nitrogen to the liver. Nitrogen in the form of ammonium ion is first converted to carbamoyl phosphate, which is transferred to ornithine to form citrulline. Aspartate and citrulline are condensed to form argininosuccinate, which, in turn, is cleaved to arginine and fumarate. Arginase hydrolyzes arginine to urea and ornithine to allow the cycle to repeat. Urea is usually viewed simply as waste, but it is also the primary contributor to the high osmolality in the renal medulla and enables maximal urinary concentrating ability.

Amino acids are precursors for many hormones and signaling molecules. Tyrosine provides a scaffold for thyroxine, dopamine, and adrenaline synthesis. Tryptophan is a precursor of serotonin and melatonin. The potent vasodilator nitric oxide (NO) is produced from arginine. Glycine, aspartate, glutamine, and serine contribute atoms to purine and pyrimidine synthesis. Glycine and arginine are precursors for creatine synthesis.

Creatine synthesis and many other biochemical processes rely on a series of single-carbon transfer reactions mediated by serine, glycine, histidine, and methionine. Transfer of fully oxidized carbonyl carbon (=C=O) to molecules such as propionyl coenzyme A (CoA) (to form methylmalonyl CoA) is mediated by biotin. Glycine, serine, and histidine contribute less oxidized carbon units such as methylidyne (=CH–), and methylene ($–CH_2–$) groups that enable purine and pyrimidine synthesis via folate derivatives. Folate also mediates transfer of methyl ($–CH_3$) groups to homocysteine to form methionine. The resulting methionine is, in turn, activated to *S*-adenosylmethionine, becoming a methyl donor to a vast array of substrates including DNA, RNA, histones, choline, and catecholamines. The importance of folate and single-carbon metabolism to cell growth and division cannot be overstated. Folate deficiency in a developing embryo can lead to death or severe neurologic birth defects. The use of folate antimetabolites such as methotrexate has been a mainstay in the treatment of cancer for many decades.

Amino Acid Concentrations

Plasma amino acid concentrations collectively span a 1000-fold dynamic range from very low micromolar quantities (e.g., β-alanine, cystathionine) to near 1 mmol/L (e.g., glutamine, glycine). With protein intake of 1 to 2 g/kg, daily variation of amino acid concentrations approaches 30% in healthy adults. Concentrations of both essential and nonessential amino acids vary in a coordinated way, suggesting that mechanisms beyond diet and enteral extraction are responsible. Amino acid concentrations tend to peak between 12 and 8 p.m., with a nadir between midnight and 4 a.m. After an ingested protein bolus, dietary amino acids rise and tend to return to preprandial levels in 3 to 6 hours. Determination of "fasting" amino acid concentrations, therefore, requires extended periods of dietary abstinence.

Most amino acids in blood undergo glomerular filtration but are efficiently reabsorbed in proximal renal tubules by previously described saturable transport systems. Increased renal excretion of amino acids (**aminoaciduria**) results from the filtration of excessive plasma concentrations, generalized tubular impairment, or heritable defects in amino acid transport systems. Glycine tends to be most abundant in normal urine, followed by histidine, glutamine, and serine. Increased concentrations of proteogenic amino acids in plasma tend to precipitate only mildly elevated excretion because of efficient reabsorption. Other amino acids that accumulate in plasma secondary to metabolic errors (e.g., argininosuccinate, homocitrulline) demonstrate pronounced excretion because of the absence of specific tubular mechanisms enabling reclamation from the filtrate.

With the exception of glutamine, cerebral spinal fluid (CSF) amino acid concentrations are typically less than 10% of those found in plasma. This high plasma to CSF gradient suggests active net brain to blood transport across the blood–brain barrier. Glutamine concentrations in CSF are generally equal to those in plasma, suggesting a bidirectional facilitative transport process. Insofar as CSF concentrations reflect synaptic concentrations, the regulation of neurotransmitter amino acid concentrations is critical for normal neural action potential propagation. Glutamate is the most abundant amino acid in the brain and is the primary excitatory transmitter. Glycine and GABA are the predominant inhibitory transmitters. Lumbar puncture to access the CSF amino acid pool must be done with great care to avoid overestimation of central amino acid concentrations secondary to contamination with peripheral blood.

Analysis of Amino Acids

For decades, the standard method of amino acid analysis was cation-exchange chromatography with postcolumn spectrophotometric or fluorescent detection of amine derivatives. Derivatizing agents have included dansyl chloride, *o*-phthaladehyde, and ninhydrin. The ninhydrin approach developed in the 1950s was initially applied to determination of amino acid content of protein hydrolysates and then adapted for profiling of free amino acids in deproteinized body fluids. Other systems commercialized by Beckman (Brea, CA) and Hitachi (Tokyo, Japan) were large floor models and required as long as 2 to 3 hours to quantitate 30 to 50 physiologic amino acids in a single patient specimen. These systems have given way to smaller bench-top systems using ninhydrin (Biochrom, Cambridge, United Kingdom) or fluorescent (quinolyl-*N*-hydroxysuccinimidyl) amine derivatives (Waters, Milford, MA) that still require 90 to 120 minutes for full sample analysis. In addition to long

cycle times, these methods are subject to interference from co-eluting amines, leading to overestimation of some amino acid concentrations.

Mass spectrometry (MS) is gaining favor as the method of choice for amino acid profiling. Newborn screening programs quantitate amino acid butyl esters derived from dried blood spots using flow-injection MS protocols. These methods do not use chromatographic separation and so cannot distinguish between isomeric or isobaric amino acids such as leucine, isoleucine, alloisoleucine, and hydroxyproline. Liquid chromatography–tandem mass spectrometry (LC-MS/MS) methods capable of distinguishing some isomers and isobars in plasma and other body fluids have also been developed. Some of these use amine-targeted derivatives, others target the carboxyl group, and some others employ no chemical derivatization. Advantages of MS techniques include improved analytic specificity, a 3- to 4-order-of-magnitude dynamic range, and rapid (20-minute) throughput. MS methods may also be optimized for profiling multiple molecular species in addition to amino acids. Such approaches promise to improve the scope of metabolic disorders detectable with a single patient specimen in a single analytic run.

PEPTIDES

This section describes the basic biochemistry of peptides. In general, the term *peptide* applies to relatively short polymers of amino acids with molecular weights less than 6000 Da (<~50 amino acid residues). The chemistry of the peptide bond and the physical characteristics of the peptide backbone are discussed in this section, along with a number of clinically relevant peptide systems.

Peptide Bond

A **peptide bond**, also referred to as an amide bond, is formed between the α-nitrogen atom of one amino acid and the carbonyl carbon of a second (Fig. 18.5). So-called isopeptide bonds refer to amide bonds between side chain amines or carbonyl carbons, rather than α-amine or α-carbonyl. In glutathione, for example, the γ-carboxyl group of glutamic acid is linked to the α-amino group of cysteine. During translation, peptide bonds are formed from the amino (N) to the carboxyl (C) terminus by removal of water (also referred to as dehydration or condensation) and catalyzed by RNA (referred to as a ribozyme) that forms part of the ribosome. Peptides are also synthesized in vitro for therapeutic and experimental purposes. Cleavage of peptide bonds may be nonspecifically achieved by acid hydrolysis or accomplished specifically by a host of proteolytic enzymes with affinity for bonds between specific amino acid residues. These protease systems are described later in this chapter.

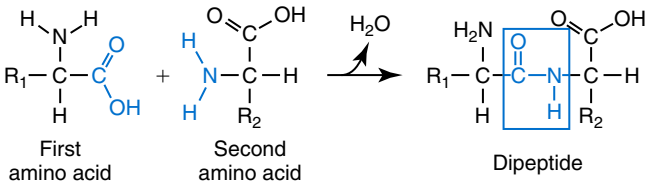


Fig. 18.5 Formation of peptide bond between α-amino and α-carbonyl of adjacent amino acids.

Peptide Heterogeneity and Analysis

The assessment of circulating peptide concentrations has a number of limitations. In the absence of enzymatic activity, peptide measurements have been historically limited to immunologic techniques. Antibodies may recognize linear sequence epitopes or discontinuous, conformational epitopes. These epitopes typically involve 10 to 20 amino acids binding exposed areas of 600 to 1000 Å. Measurements of short peptides (<20 to 30 amino acids) are therefore limited to single-site, competitive assays that lack the analytic specificity of two-site (sandwich) immunoassays. The molecular specificity issues may be addressed using MS as an alternative. Small peptides may be ionized via electrospray (ESI) or matrix-assisted laser desorption (MALDI) and interfaced to tandem quadrupole mass analyzers. Refer to Chapter 13 for additional information on mass spectrometers.

Absolute analytic specificity is not always ideal when applied to biologic peptide systems. Peptide populations may consist of species with a variable number of amino acid residues with sometimes unknown biologic potency. Hepcidin, for example, is an iron transport regulatory peptide that circulates principally as a 25–amino acid peptide but also as shorter peptides of 22 and 20 amino acids with diminished biologic activity. Likewise, dozens of truncated forms of the mature 32–amino acid B-type natriuretic peptide (BNP), ranging from 24 to 31 amino acids, are detectable in heart failure patients. Some of these truncated forms are present in vivo, and others likely develop in vitro following sample acquisition. Thus narrowly targeted MS assays may exclude active peptide species and run the risk of underestimating bioactive peptide. Cross-reactive immunoassays, on the other hand, may stoichiometrically detect both active and inactive peptide, thus running the risk of overestimating bioactive peptide concentrations. Examples of several important biologic peptide systems and their analytic considerations follow.

Clinically Relevant Peptide Systems

Pro-Opiomelanocortin System

The pro-opiomelanocortin *(POMC)* gene on chromosome 2 is expressed primarily in the pituitary gland, arcuate nucleus of the hypothalamus, and skin melanocytes. The gene produces a 241–amino acid prohormone that can yield as many as 10 distinct biologically active peptides, depending on patterns of cleavage in specific tissue types. The POMC peptides have diverse effects on glucose and electrolyte homeostasis (via adrenocorticotropic hormone [ACTH]), body mass and appetite (via lipotropins and melanocortins), pigmentation (also via melanocortins), and pain (via endorphins). Clinical exploitation of this complex peptide system is currently limited to the impact of ACTH on the adrenal gland and subsequent feedback by cortisol. Measurement of circulating ACTH concentrations is used to clarify the mechanism of

adrenal disease. For example, Cushing disease may result from autonomous adrenal function or may be fueled by ectopic ACTH production. Likewise, Addison disease may result from adrenal or pituitary failure.

Natriuretic Peptides

This peptide family consists of atrial natriuretic peptide (ANP), BNP (formerly brain natriuretic peptide), and C-type natriuretic peptide (CNP). ANP and BNP are highly expressed in cardiac tissue relative to CNP, which is expressed at low levels in a broad variety of tissue types. The mature forms of these peptides contain a 17–amino acid loop stabilized by an intramolecular disulfide bond. Each peptide acts via a specific guanylate cyclase-coupled receptor to promote sodium and water excretion, blunt activation of the renin–angiotensin system, and decrease vascular resistance. Circulating concentrations of ANP and BNP (but not CNP) increase rapidly in response to increased cardiac filling pressures that are characteristic of heart failure. Clinical measurement of BNP has become a widely used tool to detect heart failure and monitor its progression. For additional information on the clinical utility of these peptides, refer to Chapter 34.

B-type natriuretic peptide is synthesized as a 108–amino acid precursor (pro-BNP) that is cleaved upon cellular release to the active 32–amino acid peptide and an N-terminal fragment (NT-proBNP), which lacks biologic activity. Two-site immunoassays for both BNP and NT-proBNP are commercially available and widely used. NT-proBNP circulates at a higher concentration than BNP by virtue of its longer biologic half-life (~60 minutes vs. ~20 minutes). Recent evidence suggests that the BNP detected immunologically in heart failure is not the bioactive form of BNP. These studies suggest that the immunoreactive BNP in the plasma of heart failure patients may be attributed to higher molecular weight forms such as proBNP that exhibit a fraction of the bioactivity of the mature peptide. This new analytic information may prompt a reevaluation of the role of BNP in the pathophysiology, diagnosis, and treatment of heart failure.

Hepcidin

Hepcidin was initially described as an antimicrobial peptide. The role of hepcidin in iron metabolism was noted by Nicolas et al. in 2001. The mature 25–amino acid molecule is derived from a 60–amino acid precursor expressed from three exons of the *HAMP* gene on chromosome 19. The tightly looped structure of hepcidin is stabilized by four intramolecular disulfide bonds. The biologic activity of hepcidin is mediated via its interaction with ferroportin, the transport protein that mediates iron transport from duodenal enterocytes and macrophages. Hepcidin binding promotes the internalization and degradation of ferroportin, thus inhibiting mobilization of iron stores. Physiologic states such as chronic inflammation are characterized by microcytic anemia with paradoxically adequate iron stores known as anemia of chronic disease. Increased hepcidin expression is at the pathologic root of this condition.

Despite the important role of hepcidin in iron metabolism, its clinical use remains infrequent partly because of difficulties associated with its measurement. Antibodies toward hepcidin have been difficult to develop because of its small, compact size, and because it is highly conserved across multiple species. Immunoassays have largely been limited to single-site, competitive formats that cross-react significantly with shorter versions of the molecule (22-, 20-mers). This can lead to overestimation of circulating hepcidin compared with MS techniques. Improvements in the molecular specificity and harmonization of hepcidin determination promise to enhance its clinical utility. For additional information on the clinical utility and measurement of hepcidin, refer to Chapter 28.

Angiotensins

Renin is secreted by the afferent arterioles of the kidney in response to decreased flow, pressure, and sodium delivery to the renal juxtaglomerular apparatus. Renin acts on circulating angiotensinogen to initiate the formation of vasoactive peptides that act to reestablish glomerular flow. The decapeptide known as angiotensin I is cleaved from the 452–amino acid angiotensinogen molecule. Angiotensin-converting enzyme (ACE) cleaves two C-terminal residues from angiotensin I to form the octapeptide, angiotensin II, which promotes contraction of vascular smooth muscle and stimulates proximal tubular sodium reabsorption to increase blood pressure. ACE inhibitors (e.g., captopril, enalapril, quinapril) are important pharmacologic tools used to treat hypertension. For additional information, refer to Chapters 34 and 35.

Although the renin–angiotensin–aldosterone axis is an important therapeutic target, it is an infrequent target for diagnostic laboratory studies. Plasma renin activity is assessed to explore the possibility of renovascular hypertension. In this condition, unilateral restriction of renal blood flow results in the inappropriate release of renin and severe hypertension. Renin activity in the plasma is normally very low (<10 ng/mL per hour) and is typically measured by assessing the production of angiotensin I from endogenous angiotensinogen after a long incubation period (>12 hours). Angiotensin I generation is most commonly monitored via a competitive immunoassay with the potential to cross-react with shorter peptides. MS approaches that address peptide specificity and stability may mitigate these analytic limitations.

Endothelins

Endothelins (ETs) are peptides with 21 amino acids derived from the vascular endothelium (ET-1), intestinal and renal tissue (ET-2), and neural tissue (ET-3). ET-1 is produced from a 203–amino acid precursor (preproendothelin) and smaller, 30– to 40–amino acid "big" ET molecules that are inactive. ET-1 is a potent vasoconstrictor and may mediate pathology associated with diabetic nephropathy and hypertension. The current clinical utility of ET measurement is limited.

Vasopressin

Vasopressin (arginine vasopressin [AVP]), also known as antidiuretic hormone (ADH), is a nonapeptide stored in and secreted from the posterior pituitary gland. For additional information, refer to Chapter 40. Its primary target organ is the distal convoluted tubule and collecting duct, where it acts to promote water reabsorption. ADH circulates at very low concentrations (<40 pmol/L) and has a very short half-life (15 to 20 minutes), making routine diagnostic measurement impractical. Diabetes insipidus (DI) may result from faulty secretion (central DI) or from end-organ resistance (nephrogenic DI). Head injury, tumors, and some medications may also induce pathologic secretion of ADH, resulting in fluid overload, referred to as the syndrome of inappropriate antidiuretic hormone secretion (SIADH). A synthetic analog referred to as DDAVP (1-desamino, 8-D-AVP) is used therapeutically to treat DI and some forms of coagulopathy. DDAVP stimulates release of von Willebrand factor from endothelial cells and extends the half-life of circulating factor VIII, thereby mediating improvements of circulating hemostatic factors associated with various forms of von Willebrand disease and hemophilia A.

Glutathione

Glutathione consists of a glutamate residue linked to cysteine via its γ-carboxyl rather than the α-carboxyl group, followed by a conventional peptide bond between cysteine and glycine. This ubiquitous tripeptide is the most abundant intracellular thiol and circulates in the blood at micromolar concentrations. The cellular ratio of reduced glutathione (GSH) to oxidized glutathione (GSSG) ranges from 10 to 100. Intracellular glutathione performs a variety of important functions. It plays a role in maintaining the proper ratio of oxidized to reduced forms of metabolically important thiols such as coenzyme A. It also provides reducing equivalents that detoxify reactive oxygen species such as peroxides. Through the activity of glutathione-S-transferase, glutathione also serves to detoxify other xenobiotic compounds via formation of a thioether derivative, which is then excreted. Amines and peptides are transported across the plasma membrane via the γ-glutamyl moiety of glutathione, a reaction catalyzed by γ-glutamyltranspeptidase. Determination of circulating GSH and GSSG is not routinely called for in clinical practice, as the site of action is intracellular. Inborn errors of glutathione metabolism such as glutathione synthetase deficiency are detected via the accumulation of components of the γ-glutamyl cycle, such as pyroglutamic acid (5-oxoproline).

PROTEINS

The structural diversity of proteins may be described using the following features:

1. Primary structure is the linear sequence of amino acids in a peptide or protein. Posttranslational modifications of amino acids contribute to increased diversity.
2. Secondary structure describes the nature of the peptide backbone dictated by peptide bond angles described earlier and stabilized by hydrogen bonds. Examples of secondary structure include α-helix, β-sheet. Random coils refer to segments of peptide that lack defined secondary structure.
3. Tertiary structure refers to the folding of the polypeptide chain and elements of secondary structure into a compact three-dimensional (3D) shape. Folding is a complex process driven by intramolecular and solvent interactions. Hydrophobic groups tend to fold into the interior with less exposure to solvent, while charged and polar side chains tend to be located on the surface. 3D structure is stabilized by hydrogen bonds, hydrophobic interactions, and covalent disulfide bonds. Loss of tertiary structure (denaturation) is induced by temperature, organic solvents, and detergents. Limited denaturation can be reversible, but extensive unfolding and denaturation of proteins often lead to irreversible aggregation.
4. Quaternary structure refers to the incorporation of two or more polypeptide chains or subunits into a larger multimeric unit. Examples range from the relatively simple creatine kinase, a heterodimer of M and B subunits, to branched chain α-ketoacid dehydrogenase, which is a heteromeric complex of 42 (12 E1, 24 E2, and 6 E3) subunits.
5. Ligands and prosthetic groups provide additional functional and structural elements. Examples include metals in metalloenzymes, heme in hemoglobin and cytochromes, and lipids in lipoproteins. Proteins without their associated ligands are often referred to as apoproteins (e.g., apotransferrin without iron, apolipoproteins without lipid).

Physical Properties of Proteins

Differing physical properties serve as the basis of methods to separate proteins. Some important characteristics include the following:

1. *Differential solubility.* The solubility of proteins is affected by pH, ionic strength, temperature, and the characteristics of the solvent. Changing solvent pH affects the net charge of a protein. Changing ionic strength affects the hydration and solubility of proteins. The addition of organic solvents and polyethylene glycol is also useful for differential precipitation.
2. *Molecular size.* Separation of small and large molecules is commonly achieved by filtration, ultracentrifugation, and electrophoresis. These techniques may be used under conditions when proteins and peptides are in native globular states or under denaturing conditions. The addition of reducing agents allows separation of disulfide-linked components.
3. *Molecular mass.* Advances in MS allow the determination of masses of peptides and proteins with increasing accuracy.
4. *Electrical charge.* Ion-exchange chromatography, **isoelectric focusing**, and electrophoresis separate peptides and proteins based on charge.
5. *Surface adsorption.* The affinity of peptides and proteins for a variety of physical surfaces may also be used as the basis for separation. Reverse-phase chromatography, for

example, exploits the interaction of hydrophobic molecular moieties with hydrophobic surfaces (C8 or C18 alkyl chains) when the ratio of water to organic solvent is high, but not when organic content is increased.

6. *Affinity chromatography.* Specific ligands, antibodies, and other recognition molecules have been used to separate peptides or proteins selectively.

Protein Formation

Folding

Proteins are synthesized by ribosomes starting from the 5′-end of mRNA. Triplet codons in mRNA are matched with complementary sequence in transfer RNA carrying specific amino acids. Protein synthesis begins with an AUG codon encoding methionine, and the polypeptide chain is synthesized from the N-terminus.

Instructions for folding are largely contained in the primary amino acid sequence of the growing polypeptide chain. The rate of elongation may have a significant impact on folding. Pauses in translation enhance formation of secondary structural elements. Folding begins as the chain is elongated and assisted by a family of proteins referred to as chaperones. Protein folding is an error-prone process, and many molecular chaperones work to refold, prevent aggregation, or degrade misfolded proteins.

Despite such protective mechanisms, several age-related, genetic, and infectious diseases appear connected to abnormal protein folding and aggregation. Prion diseases are infectious diseases in which the transmissible protein agent may catalyze misfolding of endogenous proteins. In Alzheimer disease, deposits of amyloid may contribute to pathogenesis. Polyglutamine diseases result from genetic expansion of repeat units encoding glutamine and are associated with Huntington disease and other neurodegenerative disorders. Amyotrophic lateral sclerosis (Lou Gehrig disease) results from aggregation of transactive DNA-binding proteins. Several inherited disorders also result from mutated, misfolded proteins. In α_1- antitrypsin (AAT) deficiency, liver injury results from aggregation and accumulation of misfolded protein. Cystic fibrosis is most commonly caused by mutations in the cystic fibrosis transmembrane conductance regulator (CFTR) that result in rapid degradation of incompletely folded protein. Small molecule therapeutics capable of modulating protein folding have shown some promise in mitigating disease caused by abnormal protein aggregation.

Targeting

Proteins that are secreted or located in vesicles and membranes usually contain a hydrophobic N-terminal signal peptide about 15 to 30 amino acids in length. These signal peptides mediate interaction with the endoplasmic reticulum (ER) and are typically removed even before synthesis of the entire protein chain is completed. Membrane retention may be achieved via an uncleaved signal sequence, by one or more hydrophobic transmembrane domains, or by lipid modifications such as N-myristoylation. Newly synthesized proteins ultimately reside in a number of membranous or soluble compartments, including the nucleus, lysosome, peroxisome, mitochondrion, or plasma membrane. Plasma membrane sorting is further complicated in polarized epithelial cells where proteins may be targeted to basolateral or apical environments.

Posttranslational Modifications

Fatty Acylation

The activity and location of proteins may be modulated by the covalent attachment of fatty acyl chains. Attachment of myristate via a glycine residue promotes membrane association. Palmitoylation of cysteine residues controls localization to membrane microdomains known as lipid rafts and caveolae. Examples of palmitoylated proteins include caveolin-1 and some members of the SRC protein kinase family. Ghrelin, a potent growth hormone secretagogue, is modified by covalent attachment of an octanoyl moiety. Only octanoylated ghrelin promotes growth hormone release.

Phosphorylation

Reversible phosphorylation may impact as many as one-third of all human cellular proteins. The human genome encodes approximately 1000 kinases responsible for phosphorylation, and about 500 phosphatases responsible for removal of covalent phosphate groups. O-phosphorylation occurs at serine, threonine, and tyrosine residues. In general, serine and threonine phosphorylation acutely modifies enzyme activity and subcellular localization. Tyrosine phosphorylation, on the other hand, regulates a plethora of signaling pathways.

Prenylation

Isoprenoids are hydrophobic moieties formed as byproducts of the cholesterol synthetic pathway. These groups modify more than 300 members of the human proteome. Isoprenylation regulates membrane and molecular association of a number of proteins important for signal transduction, vesicular trafficking, cytoskeletal function, and the integrity of the nuclear membrane.

Glycosylphosphatidylinositol Anchor

The glycosylphosphatidylinositol (GPI) anchor is a glycoglycero-phospholipid construct that mediates membrane attachment for a variety of proteins. GPI-modified proteins associate with the membrane via interdigitation of two fatty acyl chains with a single membrane leaflet. The purpose of the GPI anchor remains unclear, although lipid-anchored proteins may diffuse more rapidly in the lateral plane of membranes than transmembrane proteins. Defects in GPI synthesis are responsible for paroxysmal nocturnal hemoglobinuria.

γ-Carboxylation

Glutamic acid is a five-carbon α-amino dicarboxylic acid. Vitamin K enables the addition of a second carboxyl group to the γ-carbon of glutamate. A cluster of γ-carboxylated glutamyl residues is referred to as a "gla" domain. This modification mediates calcium-dependent membrane association and is required for full functional activity. Many proteins of the coagulation cascade contain the gla domain, including thrombin, factors VII, IX, and X, and protein C, S, and

Z. Warfarin exerts its anticoagulant effect by abrogating the activities of these vitamin K–dependent factors.

Glycosylation

Secreted proteins and the extracellular domains of membrane proteins are commonly modified with carbohydrate. Carbohydrate plays an important role in folding, secretion, and stability of the modified polypeptide. Sugar chains are added in the ER and Golgi apparatus via O-linkage to serine and threonine residues or N-linkage to the amide nitrogen of asparagine residues. O-linked sugar modifications are typically simple and consist of one to four residues. N-linked sugars are far more complex. Approximately 200 human gene products are involved in glycosylation, and more than 100 genetic defects in this process have now been documented.

Sulfation

Sulfation of proteins may occur on tyrosine residues. Nearly 50 secreted and membrane proteins carry this irreversible modification that occurs in the Golgi apparatus. Examples of sulfated proteins include coagulation factors V, VIII, and IX, fibrinogen, thyroglobulin, and α-fetoprotein. Sulfation may modify molecular recognition events. A naturally occurring TYR → PHE at position 1680 of factor VIII, for example, prevents sulfation and weakens its interaction with von Willebrand factor, causing a mild form of hemophilia.

Hydroxylation

Hydroxylation is the most prevalent posttranslational modification of human proteins. Hydroxylation most commonly occurs at the 4 position of the proline ring in collagen molecules. Hydroxylation of proline residues is thought to alter polypeptide flexibility, which stabilizes the structure of collagen and other connective tissue proteins (e.g., elastin).

Nitrosylation

NO is a volatile free radical produced from arginine via three human nitric oxide synthase (NOS) enzyme systems. NO exerts its primary biologic effects (e.g., vasodilation) via guanylate cyclase–coupled receptors, but it also forms covalent nitrosothiol (S–N=O) derivatives with protein cysteine residues. Thousands of proteins with a diverse array of functions are reversibly modified in this way. The role of protein nitrosylation in health and disease remains to be fully clarified.

PROTEIN CATABOLISM

The steady-state concentration of any protein reflects its rate of synthesis and its rate of degradation. The degradative process is more than a passive, nonspecific mechanism for disposing of unwanted cellular material. It is highly specific and tightly controlled. As supporting evidence, consider that the human genome contains more than 500 proteases. These proteases belong to one of four families based on the catalytic mechanism for hydrolyzing peptide bonds. Representative members of the four protease families are detailed in Table 18.2.

Protease Families

Serine Proteases

The serine proteases are the most abundant family in humans and are so named because serine serves as the nucleophilic residue at the active site of the enzyme. Serine protease inhibitors (serpins), including AAT, antithrombin, and protein C, inactivate a broad variety of serine proteases. Some members of the serpin family, including angiotensinogen and thyroxine-binding globulin, do not possess known activity against specific proteases.

Cysteinyl Proteases

This group of proteases uses a cysteine thiol in nucleophilic attack on the peptide bond. In humans, cysteine proteases mediate apoptosis via a series of enzymes referred to as caspases. Other notable cysteine proteases include members of the cathepsin family and interleukin-1β-converting enzyme (ICE). Cystatins are endogenous inhibitors of cysteine protease activities.

Aspartyl Proteases

Peptide bond lysis by aspartyl proteases is achieved through the coordination of a water molecule between two highly conserved aspartate residues. Human aspartyl protease enzymes include members of the pepsin, cathepsin, and renin families. The target for HIV protease inhibitors (e.g., indinavir, ritonavir) is also an aspartyl protease.

Metalloproteases

Members of this protease family, commonly referred to as matrix metalloproteinases (MMPs), include approximately 25 Zn^{2+} dependent enzymes subclassified by their reactivity to collagen, gelatin, and other extracellular matrix proteins. MMPs play a multitude of roles in wound healing and repair, pathogen defense, cancer metastasis, and rheumatoid arthritis. Control of MMP activity is exercised by a group of four proteins referred to as tissue inhibitors of matrix metalloproteases (TIMPs).

Ubiquitin-Proteasome System

Ubiquitin is a highly conserved 76–amino acid polypeptide containing seven lysine residues encoded by four human genes. Proteins may be modified by a single ubiquitin moiety, or multiple molecules may be added to the original ubiquitin. Limited ubiquitination of target proteins generally modifies their subcellular location or intermolecular interactions. Polyubiquitination targets the protein for destruction. Proteolysis occurs in the 26S proteasome, consisting of a 20S catalytic and 19S regulatory subunit constructed from greater than 60 polypeptides.

Proteins in Human Serum and Plasma

The circulating proteome is a complex mixture of thousands of gene products. The 12 most abundant proteins represent more than 95% of total protein mass. Albumin alone represents more than 50% of the total mass of protein and an even higher proportion of the number of molecules, so that albumin is the main contributor to colloid osmotic pressure (oncotic pressure). Table 18.3 lists the 30 most abundant

TABLE 18.2 Selected Members of the Four Protease Families

Protease/Family	Gene Loci	Tissue Expression	Subcellular Localization	Substrate, Function, Pathology
Serine				
Corin	4p13	Broad	Plasma membrane	Natriuretic peptides (heart failure)
Trypsin	7q34	Pancreas	Secreted	Promiscuous (cleavage of LYS-X, ARG-X)
Chymotrypsin	16q23	Pancreas	Secreted	Promiscuous (cleavage of TRP-X, PHE-X, TYR-X)
Neutrophil elastase	19p13	Myeloid cells	Cytoplasm, secreted	Promiscuous (cleavage of VAL-X, ALA-X)
Factor IX	Xq27	Liver	Secreted, plasma membrane	Conversion of factor X to Xa (hemophilia B)
Activated protein C	2q14	Liver		Factor V_a, factor VIII_a (coagulopathy)
PSA (kallikrein)	19q13	Prostate	Secreted	Semen liquefaction; marker of prostate mass and cancer
C1s	12p13	Broad	Secreted	Complement C1r, C2, and C4 (angioedema)
Cysteinyl				
Caspase-3	4q34	Broad	Cytosol, nucleus, mitochondria	ASP-X-X-ASP (apoptosis)
Cathepsin C	8p23	Broad	Lysosome	Amyloid precursor (Alzheimer disease)
Aspartyl				
Renin	1q32	Ovary, broad	Secreted	Angiotensinogen (hypertension)
Pepsin	6p21	GI tract, lung	Secreted	Cleavage at adjacent hydrophobic residues (PHE-VAL, ALA-LEU, LEU-TYR)
Presenilin	14q24	Broad	Endoplasmic reticulum/Golgi	Amyloid precursor (Alzheimer disease)
Metallo				
ADAMTS13	9q34	Liver, erythroid precursors, broad	Secreted	von Willebrand factor (thrombotic thrombocytopenic purpura)
MMP-1	11q22	Muscle	Secreted	Collagen (tissue remodeling, embryogenesis, metastasis)
Angiotensin I-converting enzyme	17q23	Testes, broad	Secreted	Angiotensin I (hypertension)

proteins by mass and molecular abundance. An exhaustive list of the contents of the circulating proteome would exceed 12,000 entries.

The range of protein concentrations measured in clinical assays spans more than 10 orders of magnitude and thus poses a significant analytic challenge. In decreasing order of abundance, the source of circulating proteins include (1) proteins secreted directly into plasma, (2) proteins associated with the cell membrane and shed into the circulation, (3) secretory proteins in exocrine secretions, (4) high-abundance cytoplasmic proteins, (5) low-abundance cytoplasmic proteins, (6) transmembrane proteins, and (7) organellar proteins that must traverse more than one membrane to exit cells. Many of these serve as useful markers of physiology and disease.

Circulating concentrations of proteins depend not only on rates of production and efficiency of entry to the circulation but also on rates of clearance. Proteins and peptides substantially smaller than albumin are cleared from the circulation by glomerular filtration, unless they are bound to larger carrier molecules. Other proteins and bioactive peptides are cleared by receptor-mediated uptake or degraded by peptidases.

For all circulating proteins, the choice of measurement using plasma or serum is not without consequence. Plasma refers to the fluid portion of blood in the presence of an anticoagulant after the cells are removed by centrifugation. Common anticoagulants include heparin, ethylenediamine tetraacetic acid (EDTA), or solutions containing citrate. Small molecular weight additives such as EDTA may introduce osmotic effects on plasma volume, and citrate solutions introduce some specimen dilution. Serum is the fluid component of blood after it is allowed to clot. It differs from plasma in several respects. An approximate 4% decrease in total protein content is related mainly to removal of fibrinogen during coagulation. Intact clotting factors are consumed during the clotting process, and in their place, proteolytic fragments and the contents of platelet granules produced by the clotting process may be recovered in serum.

Abundant Components of the Circulating Proteome

Prealbumin (Transthyretin)

Prealbumin (molecular weight, 35 kDa) transports 10% of circulating triiodothyronine (T_3) and thyroxine (T_4). Prealbumin concentrations are often used as an indicator of adequacy of protein nutrition because of its relatively short half-life (~2 days), and high content of essential amino acids.

TABLE 18.3 High-Abundance Plasma Proteins

	RANKED BY MASS ABUNDANCE (MG/L)		RANKED BY MOLECULAR ABUNDANCE (μMOL/L)	
Rank	**Protein**	**Concentration**	**Protein**	**Concentration**
1	Albumin	35,000–52,000	Albumin	500–800
2	Immunoglobulin G	7000–16,000	Immunoglobulin G	40–120
3	Transferrin	2000–3600	Apolipoprotein A-I	30–70
4	Immunoglobulin A	700–4000	Apolipoprotein A-II	30–60
5	α_2-Macroglobulin	1300–3000	Transferrin	25–45
6	Fibrinogen	2000–4000	α_1-Antitrypsin	18–40
7	α_1-Antitrypsin	900–2000	Haptoglobin	6–40
8	Apolipoprotein A-I	910–1940	α_1-Acid glycoprotein	15–30
9	C3	900–1800	α_2HS-glycoprotein	9–30
10	IgM	400–2300	Immunoglobulin A	5–30
11	Haptoglobin	300–2000	Hemopexin	9–20
12	Apolipoprotein B	600–1550	Apolipoprotein C-III	6–20
13	α_1-Acid glycoprotein	500–1200	Fibrinogen	5–18
14	α_2HS-glycoprotein	400–1300	Gc-globulin	8–14
15	Hemopexin	500–1150	Apolipoprotein C-I	6–12
16	Gc-globulin (vitamin D–BP)	400–700	C3	5–10
17	Factor H	240–740	α_1-Antichymotrypsin	4–9
18	α_1-Antichymotrypsin	300–600	Apolipoprotein D	2–10
19	Inter-α-trypsin inhibitor	200–700	Prealbumin	4–8
20	Apolipoprotein A-II	260–510	β_2-Glycoprotein I	3–6
21	C4b-binding protein	200–530	Apolipoprotein A-IV	3–6
22	Ceruloplasmin	200–500	Apolipoprotein C-II	2–7
23	Factor B	180–460	Serum amyloid A4	3–6
24	Prealbumin	200–400	Inter-α-trypsin inhibitor	3–5
25	Gelsolin	200–400	Antithrombin III	3–5
26	Fibronectin	300	α_1B-glycoprotein	3–5
27	C1 inhibitor	190–370	Gelsolin	3–5
28	C4	100–400	Ceruloplasmin	2–5
29	Plasminogen	150–350	Factor H	2–5
30	Antithrombin	170–300	Factor B	2–5

Data from Hortin, G. L., Sviridov, D., & Anderson, L. (2008). High-abundance polypeptides of the human plasma proteome comprising the top 4 logs of polypeptide abundance. *Clinical Chemistry, 54*, 1608–1616.

When compared with albumin, prealbumin is a minor component of plasma but a proportionately greater component of CSF. Prealbumin is most commonly assessed using immunonephelometric or immunoturbidimetric methods. See Chapter 15 for additional information on these techniques.

Albumin

The name *albumin* (L. *albus,* meaning white) originated from the white precipitate formed during the boiling of acidic urine from patients with proteinuria. Albumin has a molecular weight of approximately 66 kDa, circulates with a half-life of 15 to 19 days, and is the most abundant plasma protein from the fetal period onward. It plays a major role in the maintenance of colloid osmotic pressure and serves as a transporter for a diverse range of substances such as fatty acids, bilirubin, and drugs. Most clinical laboratories assay albumin in plasma or serum samples by dye-binding methods, which rely on a shift in the absorption spectrum of dyes such as bromcresol green (BCG) or purple (BCP) upon albumin binding. BCP generally is slightly more specific for albumin and yields lower values than BCG, particularly for patients with kidney failure.

α_1-Antitrypsin

AAT is a serpin and is also known as α_1-proteinase inhibitor (Pi). AAT is synthesized by hepatocytes as a single polypeptide chain with 394 amino acid residues and 3 *N*-linked oligosaccharides, yielding a total molecular weight of approximately 51 kDa. AAT deficiency affects 1 in 3000 to 5000 in the United States, resulting in liver disease and pulmonary disease. Assessment of circulating AAT concentrations is usually performed by immunoturbidimetry or immunonephelometry. Concentrations typically range from 70 to 200 mg/dL (14 to 50 μmol/L) in healthy adults but are higher in patients with inflammatory disorders, malignancy, or trauma, and in women who are pregnant, on estrogen therapy, or taking oral contraceptives.

Ceruloplasmin

Ceruloplasmin (Cp) has a molecular weight of 132 kDa and contains about 95% of serum copper. Failure to incorporate copper into Cp results in an unstable protein and is the cause of Wilson disease.

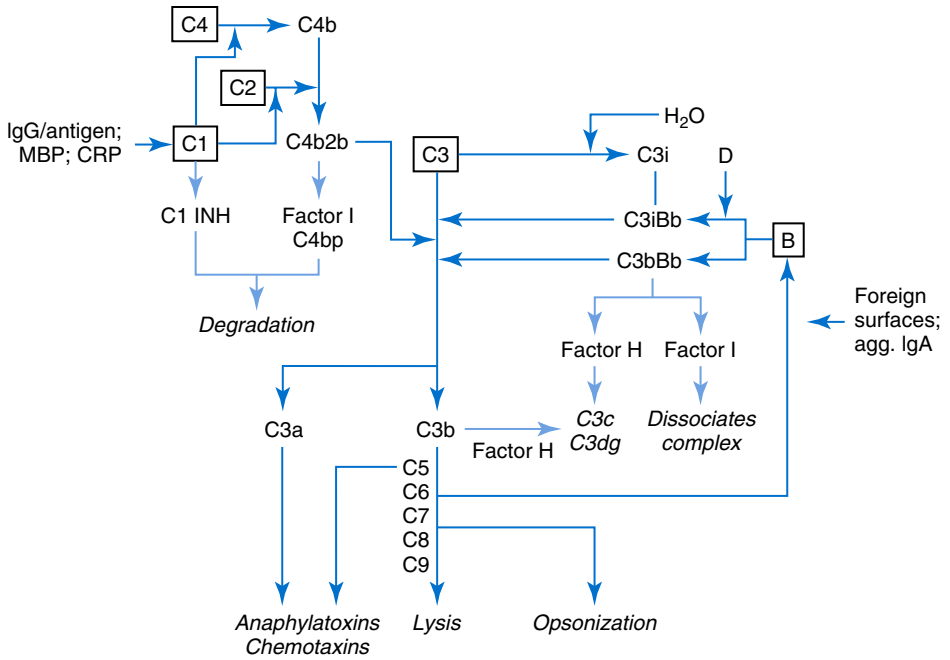


Fig. 18.6 The complement cascade. Activation via the classical pathway is shown on the *left* and via the alternative pathway on the *right*. Both pathways converge at the level of C3 convertase. Direct activation of C3 by neutrophil and plasma proteases also may occur. *CRP,* C-reactive protein; *IgA,* immunoglobulin A; *MBP,* mannan-binding protein. (Courtesy J.W. Whicher, with modifications.)

Haptoglobin

Haptoglobin (Hp) scavenges hemoglobin in the vascular space. The normal plasma half-life of Hp is about 5.5 days. Hemolysis of specimens after blood collection does not decrease Hp. Hp measurement, typically by immunoturbidimetry or immunonephelometry, is therefore part of the assessment of possible transfusion reactions or other causes of hemolysis.

Transferrin

Transferrin (originally named *siderophilin*) is the principal plasma transport protein for iron (Fe^{3+}). Transferrin has a molecular weight of 79.6 kDa, including 5.5% carbohydrate, and circulates with a half-life of 8 to 10 days. Clinical indications for direct measurement of transferrin are few. Indirect assessment of transferrin concentration may be inferred by total iron-binding capacity (TIBC). Transferrin glycan structure is widely used to detect congenital disorders of glycosylation. These altered transferrin molecules are generically referred to as carbohydrate-deficient transferrin (CDT). A desialylated version of transferrin, termed tau-transferrin or β_2-transferrin, is a substantial component of CSF but not serum or plasma.

β_2-Microglobulin

β_2-Microglobulin (BMG) is a small, nonglycosylated 99 residue protein with a molecular weight of 11.8 kDa. It is the noncovalently bound light chain subunit of class I major histocompatibility complex and has a molecular weight of ~12 kDa. Its small size allows efficient glomerular filtration, resulting in a plasma half-life of approximately 100 minutes. High plasma concentrations occur in renal failure, inflammation, and neoplasms, especially those associated with B lymphocytes.

C-Reactive Protein

C-Reactive protein (CRP) consists of five non–covalently-associated 23 kDa subunits and aids in nonspecific host defense against infectious organisms by activating the classical complement pathway. CRP is one of the strongest acute-phase reactants, with plasma concentrations rising up to 1000-fold after myocardial infarction, stress, trauma, infection, inflammation, surgery, or neoplastic proliferation. Concentrations are generally higher in bacterial than viral infection.

Epidemiologic studies demonstrate that mildly increased CRP concentrations are associated with an elevated risk of cardiovascular disease. The use of CRP for these purposes requires assays with detection limits below 0.3 mg/L that are referred to as *high-sensitivity* CRP assays.

Complement

The **complement system** consists of more than 20 proteins that recruit effector phagocytes to trigger destruction of foreign organisms. A schematic of the cascade is presented in Fig. 18.6. Activation of the cascade may proceed by the classical pathway, alternative, or lectin pathway. During activation, many complement components are cleaved. Larger cleavage products are designated by a lowercase *b* and the smaller ones by a lowercase *a*. The larger fragments usually contain a binding site for membranes, immune complexes, and protein association, or in some cases, yield new protease activities that activate subsequent component(s) of the

TABLE 18.4 Inherited Deficiencies of Complement Components

Component	Frequency of Deficiency	Associated Disorders
Ficolins 1-3	Rare?	Recurrent infection
MBP	5%	Infection in infancy; less effect on adults
MASP	Rare?	Recurrent infection (e.g., pneumococcal); inflammation
C1q[a]	Relatively rare	SLE; DLE; GN
C1r, C1s	Rare	SLE; DLE; infection
C2	≥0.0003% (homozygous)	Recurrent infection, vasculitis; SLE, DLE (no antinuclear antibody); half of affected individuals are asymptomatic
C3	Rare	Severe and recurrent bacterial infection, especially with encapsulated, pyogenic bacteria
C4a	13% (heterozygous)	SLE, DLE
C4b	13% (heterozygous)	IgA nephropathy; infection
Combined C4	35% one null; 8%–10% 2 nulls; ≈1% 3 nulls; <0.1% 4 null alleles	Total deficiency: SLE, GN, DLE (many are anti-dsDNA negative but anti-Ro/SSA positive)
C5-C9	Rare	Severe or recurrent infection with Neisseria species
Properdin	Rare	*X*-linked; neisserial infection
Factor D	Rare	Recurrent infection
Factor H or I	Rare	Hypercatabolism of C3; recurrent bacterial infection; HUS in factor H deficiency
C1 inhibitor	0.002%	Hereditary angioedema (autosomal dominant)
Decay accelerating factor (DAF, CD 55)	Rare	PNH related to decreased DAF and CD59 on cell surfaces

[a]Both quantitative and qualitative (functional) deficiencies reported.
DLE, Discoid lupus erythematosus; *GN,* glomerulonephritis; *HUS,* hemolytic-uremic syndrome; *MBP,* mannan-binding protein; *PNH,* paroxysmal nocturnal hematuria; *SLE,* systemic lupus erythematosus (or SLE-like disease).
Data from Colten, H. R., & Rosen, F. S. (1992). Complement deficiencies. *Annual Review of Immunology, 10,* 809–834; and Unsworth, D. J. (2008). Complement deficiency and disease. *Journal of Clinical Pathology, 61,* 1013–1017.

cascade. Smaller fragments typically serve as anaphylatoxic or chemotactic peptides. All three pathways converge at the C3 convertase, which triggers the formation of the C5 convertase and formation of a membrane attack complex (MAC) consisting of components C5 to C9. The MAC forms an ion channel in the membrane of the foreign pathogen triggering lysis and destruction of the target organism. Clinical disorders of inherited complement deficiency are listed in Table 18.4.

Immunoglobulins

Immunoglobulins (antibodies) are generated against foreign immunogens to initiate clearance of the foreign molecule or organism. Human immunoglobulins (commonly depicted in the shape of a Y, Fig. 18.7) consist of two identical heavy (H) chains and two identical light (L) chains. Each of the four chains has one variable and one or more constant domains. The variable domains contain the antigen-binding regions and the constant domains of the heavy chains contain sites for complement activation and receptor binding. Cleavage of immunoglobulins with pepsin or papain yields antigen-binding fragments (F_{ab}) and constant region fragments (F_c). Variations in the constant domains of heavy chains (F_c region) result in the classes and subclasses into which immunoglobulins are grouped: IgM, IgG (four subclasses), IgA (two subclasses), IgD, and IgE, respectively. Light chains, which are produced independently and in slight excess of heavy chains, are of two types—kappa (κ) and lambda (λ).

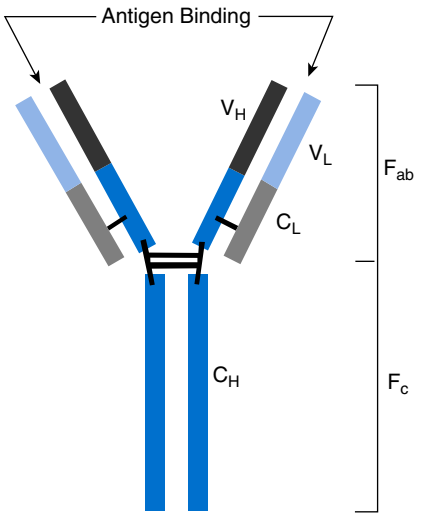


Fig. 18.7 Schematic of basic antibody structure depicting the arrangement of heavy *(H)* and light *(L)* chains, as well as respective variable *(V)* and constant *(C)* regions. Selective proteolytic digestion yields the F_{ab} fragment that contains antigen-binding sites and the F_c region.

Individual Immunoglobulins and Light Chains

Immunoglobulin G. Immunoglobulin G (IgG) is produced by mature B cells (plasma cells) and accounts for 70% to 75% of the total immunoglobulin in plasma. IgG has a molecular weight of approximately 150 kDa and migrates broadly in the γ- and β-regions on gel electrophoresis. IgG has four

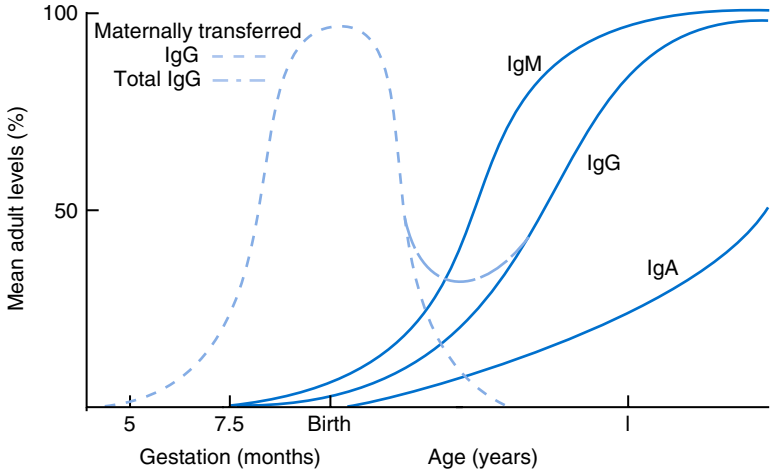


Fig. 18.8 Serum immunoglobulin concentrations as percent of mean adult concentrations before birth and for the first year of life. *IgA*, Immunoglobulin A; *IgG*, immunoglobulin G; *IgM*, immunoglobulin M.

subclasses: IgG_1, IgG_2, IgG_3, and IgG_4. IgG_1 is the principal IgG to cross the placenta, and neonatal concentrations are similar to maternal concentrations (Fig. 18.8). Neonates have low production of IgG as the result of immaturity of their immune systems, and IgG concentrations fall through infancy as maternally acquired antibody is cleared.

Immunoglobulin M. Immunoglobulin M (IgM) is produced at earlier stages of B-cell development. In the immature immune systems of neonates, IgM is the major immunoglobulin synthesized. In adult serum, it is the third most abundant immunoglobulin, usually accounting for 5% to 10% of total circulating immunoglobulins. Membrane-bound IgM may be found as a component of the B-cell receptor, while circulating IgM is a pentamer containing five monomers linked via disulfides to the small J (joining) chain. The high molecular weight of pentameric IgM (970 kDa; approximately 10% carbohydrate) prevents its ready passage into extravascular spaces. IgM is not transported across the placenta and therefore is not involved in hemolytic disease of neonates. It activates complement more efficiently than IgG.

Immunoglobulin A. Immunoglobulin A (IgA) has a molecular weight of 160 kDa, including about 10% carbohydrate. IgA accounts for about 10% to 15% of serum immunoglobulin and has a half-life of 6 days. IgA is an important component of mucosal immunity. Secretory IgA is found in tears, sweat, saliva, milk, and colostrum, as well as gastrointestinal (GI) and bronchial secretions.

Immunoglobulin D. Immunoglobulin D (IgD) accounts for less than 1% of serum immunoglobulin. It is monomeric, contains about 12% carbohydrate, and has a molecular weight of 184 kDa. Its structure is similar to that of IgG. Similar to IgM, IgD is a surface receptor for antigen on B lymphocytes, but its primary function is unknown.

Immunoglobulin E. Immunoglobulin E (IgE) contains 15% carbohydrate and has a molecular weight of 188 kDa. IgE is so rapidly and firmly bound to specific IgE receptors on mast cells that only small amounts of it are normally present in serum. IgE binds to mast cells via its F_c region. When the antigen (allergen) cross-links two of the attached IgE molecules, the mast cell is stimulated to release histamine and other vasoactive amines that increase vascular permeability and smooth muscle contraction, mediating type 1 hypersensitivity reactions such as hay fever, asthma, urticaria, and eczema. IgE molecules specific for particular allergens are analyzed to identify the source of allergies. The total serum concentration of IgE may be increased in individuals with allergic disorders.

Free Immunoglobulin Light Chains. Light chains are usually synthesized in slight excess of heavy chains to ensure the assembly of intact immunoglobulins. Consequently, small amounts of free light chain, representing only about 0.1% of total immunoglobulin, are present in serum or plasma. Free κ-light chains (23 kDa) are cleared faster than free λ-light chains, so the plasma concentration of free λ-light chains is usually higher, except in renal failure. Immunoassays specific for free light chains are often used to detect plasma cell disorders.

Clinical Utility Of Immunoglobulin Measurement

Serum normally contains a diverse, polyclonal mixture of antibodies. Disease may be associated with a decrease or an increase in normal polyclonal immunoglobulins, or an increase in one or more monoclonal immunoglobulins. These disease states are detailed below.

Immunoglobulin Deficiency. The diagnosis of major deficiencies in immunoglobulin production is important to avoid infection. Severe combined immune deficiency (SCID) is a B-cell disorder affecting 1 in 100,000 newborns, resulting in broad-spectrum immunoglobulin deficiency. More common primary deficiencies involve only one or two immunoglobulin classes or subclasses. IgA deficiency occurs in about 1 in 500 whites but much less frequently in Asian populations. IgA deficiency may lead to false-negative assays for celiac disease detection, and some affected individuals are at risk for anaphylaxis if they receive blood products containing IgA. Selective deficiency of IgG subclasses is not rare, but it is unclear whether it is an important risk for infection.

Infants have transient physiologic deficiency of IgG, with a nadir at about 3 months of age (Fig. 18.8). Prolonged or severe deficiency may be associated with increased infection rates.

TABLE 18.5 Monoclonal Immunoglobulins (Paraproteins) in Multiple Myeloma

Plasma Paraprotein	Incidence[a] (%)	Mean Age of Occurrence[a] (year)	Incidence of Free Light Chain in Urine (%)	Comments
IgG	50	65	60	Patients more susceptible to infection; paraproteins reach highest concentrations
IgA	25	65	70	Tend to have hypercalcemia and amyloidosis
Free light chain only	20	56	100	Often renal failure; bone lesions; amyloidosis; poor prognosis
IgD	2	57	100	90% λ type; often have extraosseous lesion, amyloidosis, renal failure; poor prognosis
IgM	1	—	100	May or may not have hyperviscosity syndrome
IgE	0.1	—	Most	—
Biclonal	1	—	—	—
None detected	<1	—	0	Usually reduction of normal immunoglobulins; increased plasma cells in bone marrow biopsy

[a]Approximate.
Ig, Immunoglobulin.

Concentrations of maternal IgG, transferred across the placenta, rise rapidly in the fetus during the last half of pregnancy, but then drop over a few months after birth. Two groups of neonates are at risk for clinically significant IgG deficiency: premature infants who begin life with less maternal IgG and infants with delayed initiation of IgG synthesis. Monitoring of IgG concentrations can identify this problem. Rising IgM and normal salivary IgA concentrations at 6 weeks of age suggest a favorable prognosis.

Polyclonal Hyperimmunoglobulinemia. Polyclonal increases in plasma immunoglobulins are the normal response to infection. IgG predominates in autoimmune responses. IgA is increased in skin, gut, respiratory, and renal infections, and IgM is increased in primary viral infections and bloodstream infection. Chronic bacterial infection may cause increased concentrations of all immunoglobulins. Measurements of total and allergen-specific IgE are used in the management of asthma and other allergic conditions, especially in children.

Monoclonal Immunoglobulins (Paraproteins). A single clone of plasma cells produces immunoglobulin molecules with a single defined amino acid sequence. If the clone expands greatly, it may produce a discrete band on electrophoresis, often referred to as an M-spike or M-protein. These monoclonal immunoglobulins, termed **paraproteins**, may be polymers, monomers, individual immunoglobulin chains such as free light chains or heavy chains, or fragments of immunoglobulins. Clinical, epidemiologic, and biochemical characteristics of monoclonal paraprotein diseases are summarized in Table 18.5. Up to 25% of paraproteins are benign and have been termed **monoclonal gammopathy of undetermined significance (MGUS)**. The incidence of MGUS increases with age, with 1% incidence for people 50 to 70 years of age and 3% incidence for people older than 70. The occurrence of MGUS is associated with increased risk of progression to **multiple myeloma**.

The primary clinical interest in identifying paraproteins is to detect or monitor proliferative disorders of B cells. Many patients with paraproteins have nonspecific presentations such as anemia or infection. Identification of paraproteins in serum usually is based on serum protein electrophoresis and immunofixation electrophoresis (IFE; described in a later section). Urine protein electrophoresis and urine IFE are helpful mainly in identifying patients with free immunoglobulin light chains. Urinary free light chains are referred to as ***Bence-Jones proteins.*** From the laboratory standpoint, paraproteins are often an unpredictable source of interference with many assays. Paraproteins may aggregate or precipitate, causing interference in a variety of photometric reactions and in light-scattering hematology analyzers.

Acute Phase Response

Systemic inflammation in response to infection, tissue injury, or inflammatory disease triggers changes in hepatic production of multiple plasma proteins, as indicated in Box 18.1. This process has been termed the **acute-phase response** (APR). It is a nonspecific reaction to inflammation, analogous to the increase in temperature or leukocyte count. Plasma concentrations of individual acute-phase proteins rise at different rates, and all reach maxima within 2 to 5 days after an acute insult.

Methods for Analyzing Proteins

Determination of Total Protein

Plasma normally contains about 6.5 to 8.5 g/dL protein, and serum about 4% less. Determination of total protein in biologic fluids in some respects represents a greater challenge than analysis of a specific protein because variable protein composition of biologic fluids leads to variable carbohydrate composition, charge, and physical characteristics of proteins in the mixture. Many methods have been developed to measure the total protein content of biologic fluids, and several are reviewed here.

Kjeldahl Method. This method is rarely used in clinical laboratories but is of historical importance and is sometimes used as a reference method. In this method, protein nitrogen is converted to ammonium ion, which is subsequently measured via titration. Nonprotein nitrogen, such as that from urea and amino acids, also is measured, so a protein

BOX 18.1 Acute Phase Response: Changes in Plasma Protein Concentrations

Positive Acute Phase Response
C-reactive protein (extreme)
Serum amyloid A (extreme)
α_1-Acid glycoprotein
α_1-Antitrypsin
α_1-Antichymotrypsin
Antithrombin III
C3, C4, and C9
C1 inhibitor
C4b-binding protein
Ceruloplasmin
Factor B
Ferritin
Fibrinogen
Haptoglobin
Hemopexin
Lipopolysaccharide-binding protein
Mannan-binding protein (lectin)
Plasminogen
Procalcitonin

Negative Acute Phase Response
Albumin
Apolipoprotein A-I
Apolipoprotein B
α_2-HS glycoprotein
Insulin-like growth factor I
Prealbumin
Retinol-binding protein
Thyroxine-binding globulin
Transferrin

Data from Craig, W. Y., Ledue, T. B., & Ritchie, R. F. (2001). *Plasma proteins: Clinical utility and interpretation.* Scarborough, ME: Foundation for Blood Research; Gabay, C., Kushner, I. (1999). Acute-phase proteins and other systemic responses to inflammation. *New England Journal of Medicine, 340,* 448–454; and Vollmer, T., Piper, C., Kleesiek, K., Dreier, J. (2009). Lipopolysaccharide-binding protein: a new biomarker for infectious endocarditis? *Clinical Chemistry, 55*:295–302.

precipitation step may be included. The Kjeldahl method was one of the first methods used for reproducible total protein measurement, but it is time consuming and impractical for routine use. This method has been used to assign values to reference materials for the biuret method.

Biuret Method. Under strongly alkaline conditions, Cu^{2+} ions form multivalent complexes with peptide bonds in proteins. Binding shifts the absorption spectrum of Cu^{2+} ions to shorter wavelengths, leading to a color change from blue to violet that has been termed the *biuret reaction*. Absorbance attributable to protein is measured spectrophotometrically at 540 nm. In general, the biuret reaction reacts equally with all proteins and peptides longer than two amino acids.

Direct Optical Methods. Absorbance of ultraviolet (UV) light between 200 to 225 nm and 270 to 290 nm has been used to measure protein concentrations and is commonly applied to monitor chromatographic separations of peptides and proteins. Absorbance at 280 nm depends primarily on the tryptophan and tyrosine content of a protein. This technique works best for purified proteins. From 200 to 225 nm, peptide bonds are chiefly responsible for UV absorbance. Absorptivity by proteins at these shorter wavelengths is 10 to 30 times greater than at 280 nm. Many low molecular weight compounds such as urea also have absorbance at wavelengths below 220 nm. Accurate measurement of proteins by this method may require removal of low molecular weight molecules before absorbance measurements are performed.

Dye-Binding Methods. Dye-binding methods depend on shifts in the absorbance spectra of dyes when they bind to proteins. Variable binding of dyes to different proteins is a limitation. Calibration with a protein mixture is particularly difficult to define consistently. Calibration with a pure protein such as albumin may not simulate binding to the complex mixture of proteins in serum or plasma. Coomassie blue binds to polypeptide chains under acidic conditions, resulting in increased absorbance at 595 nm. Some dyes offer sensitivity at lower protein concentrations. Pyrogallol red, for instance, is commonly used to detect total protein in urine and CSF.

Lowry (Folin-Ciocalteu) Method. The detection limit of the Lowry method is about 100 times lower than that of the biuret method. In this technique, specimens are mixed with an alkaline copper solution followed by the addition of the Folin-Ciocalteu reagent. The components of this reagent, phosphotungstic acid and phosphomolybdic acid, are reduced when complexed with peptide, resulting in increased absorbance between 650 and 750 nm.

Refractometry. Refractometry is a method used to rapidly estimate protein at high concentrations. Accuracy decreases at protein concentrations below 3.5 g/dL, where salts, glucose, and other low molecular weight compounds have a larger proportional effect on refractive index. Refractometry is used more often in clinical laboratories to assess the concentration of solutes in urine specimens than for determining total protein.

Light-scattering Methods. Many different reagents have been used to aggregate protein for turbidimetric or nephelometric assays, including trichloroacetic acid, sulfosalicylic acid, benzethonium chloride, and benzalkonium salts under alkaline conditions. Precipitation methods for total protein assays depend on forming a suspension of uniform, insoluble protein particles, which scatter incident light. Albumin and globulins often give different reactivities in precipitation methods.

Variables Affecting Measured Total Protein Concentrations. Biuret methods are commonly calibrated with bovine or human albumin. Protein mixtures with specific albumin-to-globulin ratios are recommended for calibrating other methods. Using these calibration schemes, the total protein concentration of plasma obtained from healthy ambulatory adults is typically 6.5 to 8.5 g/dL. Serum usually contains a protein concentration about 0.3 g/dL less because of the content of fibrinogen and other proteins removed during clotting. Hemodilution and dehydration, along with a change in position (standing vs. recumbent), alter plasma protein concentration. Assuming a recumbent position decreases total protein concentration by 0.3 to 0.5 g/dL. This reflects fluid

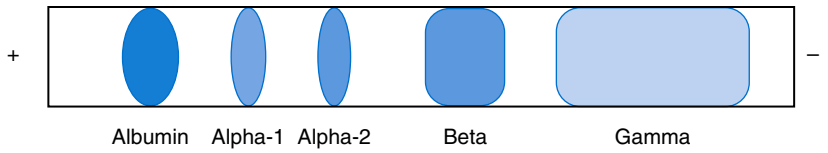


Fig. 18.9 Schematic of protein migration and regional annotation in serum protein electrophoresis. The anode is depicted on the left and cathode on the right. *Blue shading* represents the relative protein concentration (*darker* is higher) in each region in a normal individual.

redistribution, resulting in dilution of a constant amount of plasma protein in a larger volume.

Nephelometry and Turbidimetry

Nephelometric and turbidimetric methods rely on light scattering by antigen–antibody complexes. For additional information on these two techniques and the difference between them, refer to Chapter 15. Limits of detection of approximately 10 mg/L can be achieved with these methods using antibodies in solution. Binding antibodies to particles of latex or other materials enhances light scattering and can lower limits of detection by 10- to 100-fold. Nephelometric and turbidimetric assays commonly offer within-run coefficients of variation (CVs) of less than 5%, except as limits of detection are approached. Turbidimetric methods can be applied on most automated chemistry analyzers capable of performing photometric methods. Nephelometry requires instrumentation capable of measuring light scattering at an angle to the incident light.

Electrophoresis

Electrophoresis is used to separate proteins by charge. Electrophoretic techniques commonly performed in clinical laboratories include nondenaturing electrophoresis on cellulose acetate strips or agarose gels, capillary electrophoresis (CE), immunofixation, and "Western blotting." Protein separation in one or two dimensions with polyacrylamide gel is a powerful separation technique frequently used for research.

Serum Protein Electrophoresis. An idealized representation of serum protein electrophoresis is shown in Fig. 18.9. Serum is generally used for electrophoresis of proteins on agarose gels to avoid the fibrinogen in plasma that migrates at the β-γ interface. Highly abundant proteins within gels are visualized with a variety of dyes, including amido black, Ponceau S, acid violet, and Coomassie blue. Quantitation of individual bands can be done using densitometry if the total protein concentration of the sample is known. CE works differently by detecting proteins using the absorbance of proteins at wavelengths below 220 nm.

The major clinical application of serum protein electrophoresis is the detection of monoclonal immunoglobulins (paraproteins) to assist in the diagnosis and monitoring of multiple myeloma and related disorders. Most monoclonal immunoglobulins are observed in the β-region or γ-region. Quantitation of monoclonal components serves as a means of monitoring disease progression and response to therapy. Protein electrophoresis is also informative in other clinical circumstances. Changes in the α_1-region are typically related to AAT. Changes in the α_2-region usually relate to changes in Hp and α_2 macroglobulin (AMG). Bands in the β-region are related to transferrin, C3, and low-density lipoprotein (apolipoprotein B). An increase between β- and γ-bands, so-called bridging of β- and γ-bands, suggests an increase in IgA, as seen with cirrhosis, respiratory tract or skin infection, and rheumatoid arthritis. Finally, increases or decreases in the γ-region suggest changes in immunoglobulins. Fig. 18.10 provides multiple examples of serum electrophoretic patterns in normal and pathologic specimens

Immunofixation Electrophoresis. IFE employs antisera targeted to specific immunoglobulins prior to protein staining. IFE serves as a more sensitive method to detect monoclonal expansion of immunoglobulins and also serves to identify the type of heavy and light chain produced. Examples of IFE from two patients with monoclonal gammopathies are shown in Fig. 18.11.

Western Blotting. For Western blotting, proteins separated on a gel are transferred by electroblotting onto a membrane made of nitrocellulose or polyvinylidene fluoride. Proteins bound on the membrane are identified with specific antibodies conjugated to enzymes such as peroxidase or alkaline phosphatase that act on photometric, fluorescent, or chemiluminescent signaling molecules.

Mass Spectrometry. Multiple types of MS instrumentation provide qualitative or quantitative information about proteins contained in a complex mixture of other molecules. Ionization of peptides and proteins may be accomplished by electrospray or MALDI sources. Electrospray is better suited to analyzing small peptides than larger intact proteins. After ionization, proteins can be separated by quadrupoles, ion traps, time of flight, and other types of mass analyzers. Use of tandem MS with an intermediate fragmentation step between two stages of MS separation offers high sensitivity and specificity for the quantitative analysis of peptides, as it does for most small molecules.

Advantages of using MS for quantitative analysis include the ability to analyze components without developing specific antibodies, and the ability to multiplex a large number of measurements. MS can provide information about post-translational modifications that is difficult to assess by immunoassays and chromatographic or electrophoretic techniques. Examples of clinical applications include identification of genetic variants of prealbumin and CDT. The use of MS is likely to increase for accurate determination of protein concentrations, as recently applied for standardization of hemoglobin A_{1c}, insulin, and C-peptide.

Normal (Adult)

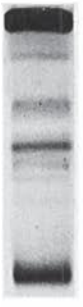
Protein	Concentration (mg/dL)
TP	6800-8300
Alb	3500-5000
AAT	100-200
AAG	50-150
Hp	30-215
AMG	125-140
TRF	200-350
C3	70-150
C4	10-40
IgA	40-390
IgM	25-210
IgG	525-1650
CRP	<2

Normal (Pediatric)

Protein	Concentration (mg/dL)
TP	6900
Alb	4390
AAT	240
AAG	59
Hp	65
AMG	490
TRF	300
C3	127
C4	27
IgA	180
IgM	140
IgG	870
CRP	<1

Chronic Renal Disease

Protein	Concentration (mg/dL)
TP	2300↓
Alb	1110↓
AAT	260↙
AAG	72↙
Hp	101
AMG	180
TRF	81
C3	71↙
C4	14↙
IgA	67↙
IgM	47↙
IgG	200↓
CRP	<1



IgG Monoclonal Gammopathy (Benign)

Protein	Concentration (mg/dL)
TP	6900↗
Alb	4380↙
AAT	200
AAG	50
Hp	75
AMG	220
TRF	270
C3	122
C4	24
IgA	70↙
IgM	170↙
IgG	1330↗
CRP	<1

IgA Monclonal Gammopathy (Multiple Myeloma)

Protein	Concentration (mg/dL)
TP	9100↑
Alb	2170↓
AAT	250
AAG	63
Hp	97
AMG	170
TRF	150
C3	90
C4	20
IgA	5800↑
IgM	24↓
IgG	200↓
CRP	<1

Nephrotic Syndrome

Protein	Concentration (mg/dL)
TP	2900↓
Alb	680↓
AAT	160↙
AAG	35↙
Hp	370*
AMG	460↑
TRF	101↓
C3	125↙
C4	22↙
IgA	250↙
IgM	93↙
IgG	440↓
CRP	<1

Inflammation (Acute)

Protein	Concentration (mg/dL)
TP	5700↙
Alb	2470↓
AAT	400↑
AAG	170↑
Hp	340↑
AMG	210
TRF	71↓
C3	120↗
C4	17↗
IgA	270
IgM	137
IgG	1440
CRP	9.8↑

Systemic Lupus Erythematosus

Protein	Concentration (mg/dL)
TP	7800
Alb	3390↙
AAT	230
AAG	43
Hp	111↙
AMG	240
TRF	310
C3	94↙
C4	12↙
IgA	650↑
IgM	170
IgG	2480↑
CRP	7.8↑

Rheumatoid Arthritis (Adult)

Protein	Concentration (mg/dL)
TP	6300
Alb	2840↙
AAT	400↑
AAG	150↑
Hp	290↑
AMG	148
TRF	220↙
C3	90↗
C4	13↗
IgA	260↗
IgM	880↑
IgG	930↗
CRP	6.1↑

Iron Deficiency

Protein	Concentration (mg/dL)
TP	6800
Alb	4770
AAT	280
AAG	44
Hp	101
AMG	220
TRF	530↑
C3	136
C4	22
IgA	150
IgM	82
IgG	880
CRP	<1

Chronic Hepatic Disease

Protein	Concentration (mg/dL)
TP	6300↙
Alb	2240↓
AAT	97↙
AAG	19↙
Hp	<1↙
AMG	290
TRF	129↓
C3	53↙
C4	4↙
IgA	480↑
IgM	620↗
IgG	2370↑
CRP	<1↑

Chronic Hemolysis and Iron Deficiency

Protein	Concentration (mg/dL)
TP	6300
Alb	4010
AAT	190
AAG	43
Hp	<1↓
AMG	400
TRF	390↑
C3	134
C4	14
IgA	180
IgM	170
IgG	700
CRP	<1

Fig. 18.10 Electrophoretic patterns typical of normal conditions and of some pathologic conditions (agarose gel). The *upward-* and *downward-pointing arrows* indicate increase and decrease from the reference interval, respectively. *Right-* and *left-slanting arrows* indicate variation from normal to an increase or from normal to a decrease from the reference interval, respectively. *AAG*, Alpha-1 acid glycoprotein; *AAT*, alpha-1 antitrypsin; *Alb*, albumin; *AMG*, alpha-2 macroglobulin; *C3*, complement component 3; *C4*, complement component 4; *CRP*, C-reactive protein; *Hp*, haptoglobin; *Ig*, immunoglobulin; *TP*, total protein concentration; *TRF*, transferrin.

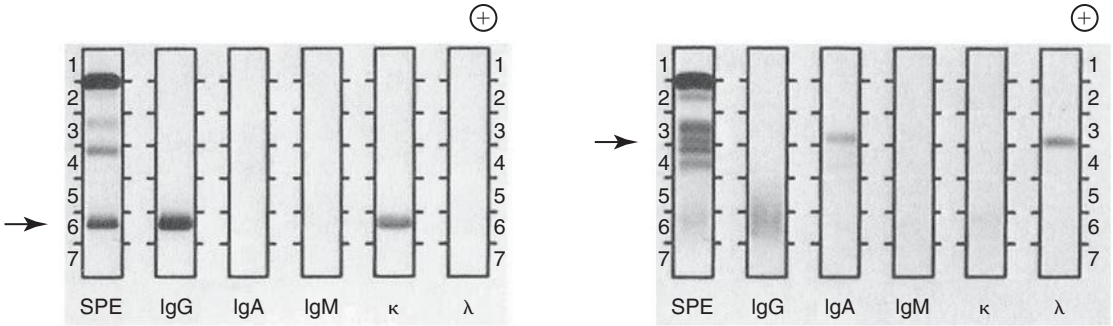


Fig. 18.11 Immunofixation electrophoresis. *Left*, Patient specimen with an immunoglobulin G (IgG; κ) monoclonal protein. *Right*, Patient specimen with an immunoglobulin A (IgA; λ) monoclonal protein. The *arrow* indicates the position of monoclonal protein. *SPE*, Serum protein electrophoresis.

Proteins in Other Body Fluids

Complex mixtures of proteins are present in all biologic fluids; analysis of a variety of other specimens is diagnostically useful, including analyses of urine, CSF, pleural and peritoneal fluids, amniotic fluid, saliva, and feces.

Saliva

Saliva is an ultrafiltrate of plasma that contains a unique set of proteins derived from the salivary glands. Protein composition varies with the site and method of sampling. In addition to the well-known presence of amylase, proteomic approaches have detected sequence from hundreds of different proteins in saliva. Proteins involved in host defense, such as immunoglobulins, lysozyme, and lactoferrin, are also particularly abundant. Efforts are underway to exploit the salivary proteome to detect and characterize susceptibility to dental caries, periodontal disease, head and neck cancers, diabetes, and cystic fibrosis. Interrogation of saliva for the presence of secretory IgA is common in the diagnostic workup of hypogammaglobulinemia.

Cerebrospinal Fluid

CSF is the extracellular fluid around the brain and spinal column. CSF usually has total protein concentrations about 100-fold lower than plasma and a different protein composition because most proteins have limited passage across the blood–brain barrier. CSF is typically obtained by a spinal tap in the lumbar region. The blood–brain or blood–CSF barrier limits exchange of large compounds, so low to intermediate molecular weight plasma proteins such as prealbumin, albumin, and transferrin normally predominate in CSF. Analyses of total protein and specific proteins in CSF are employed to detect increased permeability of the blood–CSF barrier, increased intrathecal synthesis, or increased release of proteins from neural and glial tissue. Conditions such as viral meningitis, encephalitis, increased intracranial pressure, trauma, and hemorrhage may all compromise the blood–brain barrier, resulting in increased CSF protein. Protein concentrations associated with these conditions are displayed in Table 18.6. Increased intrathecal synthesis of immunoglobulins, particularly IgG, occurs in demyelinating diseases of the CNS, especially multiple sclerosis.

Total Protein In Cerebrospinal Fluid. The total protein concentration in CSF is an indicator of blood–CSF permeability. The protein concentration of CSF usually is more than 100-fold lower than plasma, so methods with greater sensitivity or increased specimen volume are required for CSF applications. Methods commonly used by clinical laboratories to measure total CSF protein include (1) pyrogallol red, (2) benzethonium chloride, (3) reverse-biuret, and (4) biuret. The reverse-biuret method measures free copper remaining after formation of biuret complexes by reduction and complexation with a chelating dye. Analyzers often use the same method, possibly with some adjustment of specimen volume, for CSF and urine protein. The usual reference interval for CSF total protein in adults is 15 to 45 mg/dL. Total CSF protein concentrations are considerably higher in neonates and in healthy elderly adults. In CSF from premature and full-term neonates, concentrations up to 400 mg/dL are observed. In term newborns, a progressive decline in CSF protein is seen over the first few weeks of life, with values approaching adult concentrations by 4 months of age.

Assessment Of Specific Cerebrospinal Fluid Proteins. The electrophoretic pattern of normal CSF has two striking features—a prealbumin band and two transferrin bands—one at β_2 in addition to the usual β_1 position. As mentioned previously, β_2-transferrin has decreased charge because its glyosyl chains lack terminal sialic acid residues. Both β_2-transferrin and another relatively abundant CSF protein, prostaglandin D synthase or β-trace, have been used to determine whether clear fluids from nasal or ear passages represent leakage of CSF—so-called CSF rhinorrhea and otorrhea. CSF immunoglobulin concentrations and oligoclonal immunoglobulin bands on electrophoretic separations of CSF are helpful in the diagnosis of multiple sclerosis and other inflammatory diseases of

TABLE 18.6 Cerebrospinal Fluid Total Protein in Various Diseases

Clinical Condition	Appearance and Cells × 106/L	Total Protein, mg/dL
Normal	Clear, colorless; 0–5 lymphocytes	15–45[a]
Increased Admixture of Proteins From Blood		
Increased Capillary Permeability		
• Bacterial meningitis	Turbid, opalescent, purulent, usually >500 polymorphs	80–500
• Cryptococcal meningitis	Clear or turbid; 50–150 polymorphs or lymphocytes	25–200
• Leptospiral meningitis	Clear to slight haze; polymorphs early, then 5–100 lymphocytes	50–100
• Viral meningitis	Clear or slight haze, colorless; usually ≤500 lymphocytes	30–100
• Encephalitis	Clear or slight haze, colorless; usually ≤500 lymphocytes	15–100
• Poliomyelitis	Clear, colorless; ≤500 lymphocytes	10–300
• Brain tumor	Usually clear; 0–80 lymphocytes	15–200 (usually normal)
Mechanical Obstruction		
• Spinal cord tumor[b]	Clear, colorless, or yellow	100–2,000
Hemorrhage		
Cerebral hemorrhage	Colorless, yellow, or bloody; blood cells	30–150
Local Immunoglobulin Production		
• Neurosyphilis	Clear, colorless; 10–100 lymphocytes	50–150
• Multiple sclerosis[c]	Clear, colorless; 0–10 lymphocytes	25–50
Both Increased Capillary Permeability and Local Immunoglobulin Production		
• Tuberculous meningitis	Colorless, fibrin clot, or slightly turbid; 50–500 lymphocytes	50–300 (occasionally ≤1000)
• Brain abscess	Clear or slightly turbid	20–120
After myelography (inflammatory reaction)		Slight increase

[a]Premature infants: ≤400 mg/dL; children: 30–100 mg/dL; older adults: ≤60 mg/dL.
[b]Froin syndrome: Lumbar fluid values are much higher than cisternal fluid values.
[c]Similar values may occur in certain other chronic inflammatory conditions of the nervous system.

the CNS. Myelin basic protein concentrations may be an indicator of myelin turnover in multiple sclerosis. A number of other proteins such as S100B and neuron-specific enolase are potential markers of traumatic or ischemic brain injury. Concentrations of tau and β-amyloid isoforms may be useful in the diagnosis and prognosis of Alzheimer disease.

Peritoneal and Pleural Fluids

Pathologic accumulations of fluid in the peritoneal and pleural cavities or elsewhere vary greatly in protein content. These fluids may be ultrafiltrates with low-protein concentrations relative to plasma and scant amounts of large proteins (transudates). Alternatively, fluids may have protein concentrations approaching those of plasma and significant amounts of large proteins such as immunoglobulins and α_2 macroglobulin in response to local inflammation and increased vascular permeability (exudates). Distinction between transudates and exudates assists in diagnosing the cause of fluid accumulation. The major cause of pleural transudates is congestive heart failure. Exudates occur with infection, pleuritis, pulmonary embolism, and cancer.

Fecal Material

The use of fecal material for protein analysis is relatively infrequent but indicated in some specific clinical circumstances. Assessment of protein content in feces has a number of limitations. Results of fecal protein content are often normalized to fecal weight. Thus watery stool specimens generally provide decreased estimates of protein concentration. Extraction is likewise dependent on stool consistency and homogeneity, leading to considerable within-subject and between-subject variability.

Protein loss in the GI tract may be assessed using an assay of fecal AAT. The amount of fecal AAT excreted over time is

determined as a function of the serum AAT concentration. Inflammatory bowel disease (IBD) includes Crohn disease and ulcerative colitis. A number of fecal protein markers have been employed as tools for assessment of IBD, Crohn disease, and ulcerative colitis. Fecal products secreted by white blood cells, such as lactoferrin and calprotectin, have been used as a measure of disease activity in IBD. Exocrine pancreatic disease may be assessed by measurement of proteins such as fecal elastase that are derived from pancreatic secretions.

POINTS TO REMEMBER

Amino acids are more than the backbone of peptides and proteins.
- Amino acids can exist in multiple charged states depending on pH.
- L-Amino acids are the most important biologically active isomers.
- Amino acid metabolites serve as respiratory fuel.
- Amino acid transport is achieved by numerous membrane transport systems.
- Analysis of amino acids allows detection of inborn errors of metabolism.

Peptides
- Larger peptides are processed to form multiple smaller bioactive peptides.
- Peptide bonds form between α-amino and α-carbonyl of successive amino acids.
- Immunoassays often detect mixtures of peptides with variable biologic activity.
- B-Type natriuretic peptide is useful in the diagnosis of heart failure.
- Hepcidin inhibits mobilization of storage iron.

Proteins
- The primary structure of proteins is the linear sequence of amino acid residues.
- Proteins may be separated by size, electric charge, and differential solubility.
- Protein folding is assisted by molecular chaperones.
- Proteins are posttranslationally modified to alter location and function.
- Protein catabolism is a specific and a highly controlled process.
- The human plasma proteome consists of more than 12,000 entries.

REVIEW QUESTIONS

1. Which of the following amino acids participates in formation of new polypeptide chains?
   a. Desmosine
   b. Ornithine
   c. Histidine
   d. Taurine
2. Which of the following amino acids is the most hydrophobic at physiologic pH?
   a. Glutamate
   b. Lysine
   c. Isoleucine
   d. Threonine
3. The pH at which a molecule such as a protein has a net charge of zero is referred to as the:
   a. isoelectric point.
   b. dissociation constant.
   c. isosbestic point.
   d. solubility point.
4. A paraprotein is a:
   a. membrane-bound protein.
   b. denatured/unfolded protein.
   c. protein-prosthetic group.
   d. monoclonal immunoglobulin.
5. Which of the following is a ketogenic amino acid?
   a. Glycine
   b. Leucine
   c. Glutamate
   d. Proline
6. Which of the following is a positive acute-phase reactant?
   a. Transferrin
   b. Albumin
   c. Ferritin
   d. Prealbumin
7. Covalent modification of proteins with ubiquitin leads to:
   a. membrane localization.
   b. catabolism/destruction.
   c. enhanced activity.
   d. nuclear localization.
8. The plasma protein that serves to transport a large number of compounds including bilirubin, calcium, drugs, and free fatty acids is:
   a. prealbumin.
   b. ceruloplasmin.
   c. haptoglobin.
   d. albumin.

9. α_1 Antitrypsin inhibits which class of protease?
   a. Metalloproteases
   b. Aspartyl proteases
   c. Cysteinyl proteases
   d. Serine proteases

10. The most abundant circulating immunoglobulin in adult serum is:
   a. IgA.
   b. IgD.
   c. IgG.
   d. IgM.

SUGGESTED READINGS

Broer, S. (2008). Amino acid transport across mammalian intestinal and renal epithelia. *Physiological Reviews*, *88*, 249–286.

Cohen, P. (2000). The regulation of protein function by multisite phosphorylation–a 25 year update. *Trends in Biochemical Science*, *25*, 596–601.

Dietzen, D. J., Weindel, A. L., Carayannopoulos, M. O., Landt, M., Normansell, E. T., Reimschisel, T. E., & Smith, C. H. (2008). Rapid comprehensive amino acid analysis by liquid chromatography/tandem mass spectrometry: comparison to cation exchange with post-column ninhydrin detection. *Rapid Communications in Mass Spectrometry*, *22*, 3481–3488.

Fares, F. (2006). The role of O-linked and N-linked oligosaccharides on the structure-function of glycoprotein hormones: development of agonists and antagonists. *Biochimica et Biophysica Acta*, *1760*, 560–567.

Gabay, C., & Kushner, I. (1999). Acute-phase proteins and other systemic responses to inflammation. *The New England Journal of Medicine*, *340*, 448–454.

Ganz, T. (2003). Hepcidin, a key regulator of iron metabolism and mediator of anemia of inflammation. *Blood*, *102*, 783–788.

Hochstrasser, M. (2009). Origin and function of ubiquitin-like proteins. *Nature*, *458*, 422–429.

Hoofnagle, A. N. (2010). Peptide lost and found: internal standards and the mass spectrometric quantification of peptides. *Clinical Chemistry*, *56*, 1515–1517.

Kramer, G., Boehringer, D., Ban, N., & Bukau, B. (2009). The ribosome as a platform for co-translational processing, folding and targeting of newly synthesized proteins. *Nature Structural & Molecular Biology*, *16*, 589–597.

Kringelum, J. V., Nielsen, M., Padkjaer, S. B., & Lund, O. (2013). Structural analysis of B-cell epitopes in antibody:protein complexes. *Molecular Immunology*, *53*, 24–34.

Kyle, R. A., & Rajkumar, S. V. (2009). Criteria for diagnosis, staging, risk stratification and response assessment of multiple myeloma. *Leukemia*, *23*, 3–9.

Lingappa, V. R., & Blobel, G. (1980). Early events in the biosynthesis of secretory and membrane proteins: the signal hypothesis. *Recent Progress Hormone Research*, *36*, 451–475.

Meister, A., & Anderson, M. E. (1983). Glutathione. *Annual Review of Biochemistry*, *52*, 711–760.

Miller, S. L. (1953). A production of amino acids under possible primitive earth conditions. *Science*, *117*, 528–529.

Pandey, K. N. (2005). Biology of natriuretic peptides and their receptors. *Peptides*, *26*, 901–932.

Ramakrishnan, C., & Ramachandran, G. N. (1965). Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units. *Biophysical Journal*, *5*, 909–933.

Rodnina, M. V., Beringer, M., & Bieling, P. (2005). Ten remarks on peptide bond formation on the ribosome. *Biochemical Society Transactions*, *33*, 493–498.

Stiehm, E. R. (2008). The four most common pediatric immunodeficiencies. *Journal Immunotoxicol*, *5*, 227–234.

Wurtman, R. J., Rose, C. M., Chou, C., & Larin, F. F. (1968). Daily rhythms in the concentrations of various amino acids in human plasma. *The New England Journal of Medicine*, *279*, 171–175.

Serum Enzymes

Mauro Panteghini, Renze Bais

OBJECTIVES

1. Define the following and give an example of each type:
   - Hydrolase
   - Phosphotransferase
   - Lyase
   - Transferase
   - Oxidoreductase
2. List the factors that affect enzyme activity and appearance in blood.
3. For each of the following enzymes, state and describe the biochemistry, physiological actions (if known), tissue distribution, clinical significance, method of laboratory analysis (including analytical performance specifications), and possible analytical interferences:
   - 5′-Nucleotidase
   - Aspartate aminotransferase
   - Alanine aminotransferase
   - Creatine kinase (CK)
   - Alkaline phosphatase (ALP)
   - γ-Glutamyltransferase
   - Amylase
   - Lactate dehydrogenase (LDH)
   - Lipase
   - Tartrate-resistant acid phosphatase
   - Serum cholinesterase
4. List the isoenzymes, the clinical significance of the isoenzymes (if any), and methods of laboratory analysis used to assess the isoenzymes of each of the following:
   - ALP
   - LDH
   - Amylase
   - Tartrate-resistant acid phosphatase
   - CK
5. Assess and solve case studies involving serum enzymes and isoenzymes and describe the laboratory methods for their analysis.

KEYWORDS AND DEFINITIONS

α-Amylase An enzyme that catalyzes the hydrolysis of 1,4-alpha-glycosidic linkages in starch, glycogen, and related polysaccharides and oligosaccharides.

Acid phosphatase All phosphatases with optimal activity below pH 7.0 that catalyze the cleavage of orthophosphate from orthophosphoric monoesters; most of the activity in serum is of a tartrate-resistant type.

Acute pancreatitis A sudden inflammation of the pancreas that is usually accompanied with severe upper abdominal pain.

Aldolase A lyase that catalyzes cleavage of fructose-1,6-diphosphate into dihydroxyacetone-phosphate and glyceraldehyde 3-phosphate in the glycolytic breakdown of glucose to pyruvate.

Alkaline phosphatase A hydrolase that catalyzes the alkaline hydrolysis of a large variety of naturally occurring and synthetic substrates.

Aminotransferases A subclass of enzymes of the transferase class that catalyze the transfer of an amino group from a donor (generally an amino acid) to an acceptor (generally a 2-oxo acid). Most of these enzymes are pyridoxal phosphate proteins. Alanine and aspartate aminotransferase are examples that are of significant clinical utility.

Apoenzyme The protein component of an enzyme.

Cholecystitis A painful inflammation of the gallbladder.

Cholinesterase An enzyme of the hydrolase class that catalyzes the cleavage of the acyl group from various esters of choline, including acetylcholine, and some related compounds.

Coenzyme An organic nonprotein molecule that binds with the protein molecule (apoenzyme) to form the active enzyme (holoenzyme).

Creatine kinase A dimeric transferase enzyme that catalyzes the reversible phosphorylation of creatine by adenosine triphosphate (ATP). Creatine kinase (CK) has four isoenzymes: CK-MM, CK-MB, CK-BB, and mitochondrial CK.

γ-Glutamyltransferase A transferase enzyme that reversibly catalyzes the transfer of a glutamyl group from a glutamyl-peptide and an amino acid to a peptide and a glutamyl-amino acid.

Holoenzyme Active enzyme formed by combination of a coenzyme and an apoenzyme.

Isoenzyme A molecular form that originates at the level of the genes that encode the structures of the enzyme proteins in question.

Isoform An enzyme molecular form that has been posttranslationally modified.

Lactate dehydrogenase An oxidoreductase enzyme that reversibly catalyzes the reduction of pyruvate to (L)-lactate, using NADH (reduced form of nicotinamide adenine dinucleotide) as an electron donor.

Lipase A hydrolase that hydrolyzes glycerol esters of long-chain fatty acid.

5′-Nucleotidase A phosphatase that acts only on nucleoside-5′-phosphates, such as adenosine-5′-phosphate (AMP), releasing inorganic phosphate.

Paget disease A chronic disorder that results in enlarged and deformed bones (also known as osteitis deformans, osteodystrophia deformans).

Prosthetic group A tightly bound, nonpeptide structure required for the activity of an enzyme.

The basic principles of enzyme and rate analyses are discussed in Chapter 14. Individual enzymes of clinical utility are discussed in this chapter. To better clarify their clinical application, the individual enzymes are discussed relative to the organ in which they are clinically most important. However, overlap may occur for this classification, as the same enzyme may be used for investigating disease in different organs.

BASIC CONCEPTS

Injury to tissue releases cellular substances such as enzymes that can be used as plasma markers of tissue damage. For a substance to serve as a biochemical marker of damage to a specific organ or tissue, it must arise predominantly from the organ or tissue of interest. Some enzymes are found predominantly in specialized tissue (e.g., lipase [LIP] in the pancreas); others, more widely distributed, have tissue-specific **isoenzymes** or **isoforms** (e.g., the pancreatic isoenzyme of α-amylase) that can be evaluated to enhance tissue and organ specificity.

In general, laboratory specialists use changes in activity in the serum or plasma of enzymes that are predominantly intracellular and physiologically present in the blood at low concentrations only. Increases in the serum activities of these enzymes are used to infer the location and nature of pathological changes in tissues of the body. Therefore an understanding of the factors that affect the rate of release of enzymes from their cells of origin and the rate at which they are cleared from the circulation is necessary to correctly interpret changes in activity that occur with disease.

Knowledge of the intracellular locations of enzymes can assist in determining the nature and severity of a pathological process if suitable enzymes are assayed in the blood. For instance, a mild inflammation of the liver, such as a mild attack of viral hepatitis, is likely to increase only the permeability of the liver cell membrane while allowing cytoplasmic enzymes to leak out into the blood, but a severe attack causing cell necrosis also disrupts the mitochondrial membrane, and both cytoplasmic and mitochondrial enzymes are detected in the blood. The timing of the enzyme's diagnostic window is another important aspect to be considered when these markers are used to evaluate acute injury. The diagnostic window for an injury marker is the interval of time after an episode of injury during which plasma concentrations of the marker are increased, thereby demonstrating the occurrence of injury.

POINTS TO REMEMBER

The pattern of appearance of an enzyme in blood after an acute injury depends on:

- The intracellular location and whether molecules are bound or free
- Molecular weight (because heavier molecules diffuse at a slower rate)
- Local blood and lymphatic flow
- The rate of elimination from blood

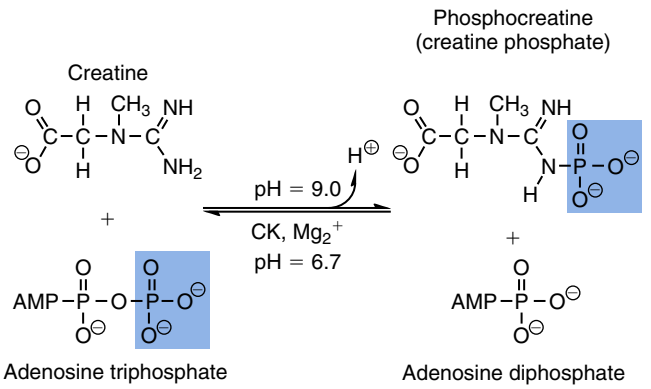
The main enzymes of established clinical value, together with their tissues of origin and their major clinical applications, are listed in Table 19.1.

MUSCLE ENZYMES

Enzymes in this category include **CK** and **aldolase** (ALD).

Creatine Kinase

Creatine kinase (CK) (EC 2.7.3.2; adenosine triphosphate [ATP]: creatine *N*-phosphotransferase) is a dimeric enzyme (82 kDa) that catalyzes the reversible phosphorylation of creatine (Cr) by ATP.



Physiologically, when muscle contracts, ATP is converted to adenosine diphosphate (ADP), and CK catalyzes the rephosphorylation of ADP to ATP using creatine phosphate (CrP) as the phosphorylation reservoir.

Optimal pH values for the forward (Cr + ATP → ADP + CrP) and reverse (CrP + ADP → ATP + Cr) reactions are 9.0 and 6.7, respectively. Mg^{2+} is an obligate activating ion that forms complexes with ATP and ADP. The optimal

TABLE 19.1 Distribution of Clinically Important Enzymes

Enzyme	Principal Sources of Enzyme in Blood	Principal Clinical Applications
Alanine aminotransferase	Liver	Hepatic parenchymal disease
Alkaline phosphatase	Liver, bone, intestinal mucosa, placenta	Hepatobiliary disease, bone disease
Amylase (pancreatic isoenzyme)	Pancreas	Pancreatic disease
Aspartate aminotransferase	Heart, liver, skeletal muscle, erythrocytes	Hepatic parenchymal disease
Creatine kinase	Skeletal muscle, heart	Muscle disease
γ-Glutamyltransferase	Liver, pancreas, kidney	Hepatobiliary disease
Lactate dehydrogenase	Heart, erythrocytes, lymph nodes, skeletal muscle, liver	Hemolytic and megaloblastic anemias, leukemia and lymphoma, oncology
Lipase	Pancreas	Pancreatic disease

TABLE 19.2 Approximate Concentrations of Tissue Creatine Kinase Activity (Expressed as Multiples of Creatine Kinase Activity Concentrations in Serum) and Cytoplasmic Isoenzyme Composition

		ISOENZYMES (%)		
Tissue	Relative CK Activity	CK-BB	CK-MB	CK-MM
Skeletal muscle (type I, slow twitch, or red fibers)	50,000	<1	3	97
Skeletal muscle (type II, fast twitch, or white fibers)	50,000	<1	1	99
Heart	10,000	<1	22	78
Brain	5000	100	0	0
Gastrointestinal tract smooth muscle	5000	96	1	3
Urinary bladder smooth muscle	4000	92	6	2

CK, Creatine kinase.

concentration range for Mg^{2+} is narrow, and excess Mg^{2+} is inhibitory. Many metal ions, such as Mn^{2+}, Ca^{2+}, Zn^{2+}, and Cu^{2+}, inhibit enzyme activity, as do iodoacetate and other sulfhydryl-binding reagents.

The enzyme in serum is relatively unstable, activity being lost because of sulfhydryl group oxidation at the active site of the enzyme. Activity can be partially restored by incubating the enzyme preparation with sulfhydryl compounds, such as *N*-acetylcysteine, dithiothreitol (Cleland reagent), or glutathione.

Biochemistry

CK activity is greatest in striated muscle and heart tissue, which contain some 2500 and 550 U/g of protein, respectively. Other tissues, such as (1) brain, (2) gastrointestinal tract, and (3) urinary bladder, contain significantly less activity (Table 19.2).

CK is a dimer composed of two subunits (B and M), each with a molecular weight of about 40 kDa. These subunits are the products of loci on chromosomes 14 and 19, respectively. Because the active form of the enzyme is a dimer, only three different pairs of subunits can exist: BB (or CK-1), MB (or CK-2), and MM (or CK-3). The distribution of these isoenzymes in the various tissues of humans is shown in Table 19.2. All three of these isoenzyme species are found in the cytosol of the cell or are associated with myofibrillar structures. However, there exists a fourth form that differs from the others both immunologically and by electrophoretic mobility. This isoenzyme (CK-Mt) is located between the inner and outer membranes of mitochondria, and it constitutes, in the heart for example, up to 15% of total CK activity. The gene for CK-Mt is located on chromosome 15.

CK activity in serum may also be found in macromolecular form—the so-called macro-CK. It exists in two forms: types 1 and 2. Type 1 is a complex of CK, typically CK-BB, and an immunoglobulin, often IgG. It is not of pathological significance, but it can be the cause of increased CK results, leading to diagnostic confusion and unnecessary further investigation. Prevalence has been estimated between 0.8% and 2.3%, and >80% of individuals positive for macro-CK type 1 (immunoglobulin bound) are female. Macro-CK type 2 is oligomeric CK-Mt, with a reported prevalence of between 0.5% and 2.6% in hospitalized patients. It is found predominantly in adults who are severely ill with malignancy or liver disease, and in children who have notable tissue distress. The appearance of this macro-CK in serum is usually associated with a poor prognosis. Macro-CKs can interfere with the assay of CK-MB by some immunoinhibition methods.

Clinical Significance

CK determination is the preferred laboratory test in cases of suspected muscle damage. Serum CK is increased in nearly all patients when (1) injury, (2) inflammation, or (3) necrosis of skeletal (or heart) muscle occurs.

Increased serum CK activity may be the only sign of subclinical neuromuscular disorders. About 30% to 44% of asymptomatic subjects with persistent hyper-CKemia have myopathy. Serum CK activity is greatly elevated in all types of

muscular dystrophy. In progressive muscular dystrophy (particularly Duchenne sex-linked muscular dystrophy), enzyme activity in serum is highest in infancy and childhood, and may be increased long before the disease is clinically apparent. Serum CK activity characteristically falls as patients get older and as the mass of functioning muscle diminishes with progression of the disease. About 50% to 80% of asymptomatic female carriers of Duchenne dystrophy show a three- to sixfold increase in CK activity. High values of CK (up to 50-fold the upper reference limit [URL] in active disease) are noted in viral myositis, polymyositis, and other inflammatory myopathies. However, in neurogenic muscle diseases, such as (1) myasthenia gravis, (2) multiple sclerosis, (3) poliomyelitis, and (4) parkinsonism, serum enzyme activity is not increased.

In acute rhabdomyolysis caused by crush injury, with severe muscle destruction, serum CK activities exceeding 200 times the URL may be found. In this condition, a very high serum CK, mirroring myoglobinuria and its hem-induced mechanism of renal injury, has been associated with the risk of development of acute renal failure. If the CK remains below 5000 U/L (about 30 times the URL) during the first 3 days after the insult, the probability of developing acute renal failure appears to be low. Serum CK can also be mildly increased by other direct trauma to muscle, including intramuscular injection and surgical intervention. Finally, a number of drugs can increase serum CK activities. The drugs principally responsible are (1) statins, (2) fibrates, (3) antiretrovirals, and (4) angiotensin II receptor antagonists. The clinical spectrum of statin induced myotoxicity includes (1) asymptomatic rise in serum CK activity, (2) myalgia, (3) myositis, and (4) rhabdomyolysis (0.02%). Routine monitoring of CK in asymptomatic patients is not recommended; however, CK must be assessed in patients presenting with muscle pain and weakness, and statin treatment stopped if values increase more than fivefold the URL.

Changes in serum CK and its MB isoenzyme after acute myocardial infarction have been used for diagnosis for many years. However, it is now more clinically appropriate to use more cardiac-specific troponin I or T (see Chapter 34).

During normal childbirth, a sixfold increase in maternal serum CK activity occurs. Surgical intervention during labor further increases the activity of CK in serum.

Determination of Creatine Kinase Activity

Currently all commercial assays for total CK are based on the reverse reaction (Cr ← CrP), which proceeds about six times faster than the forward reaction. CK catalyzes the conversion of CrP to Cr with concomitant phosphorylation of ADP to ATP. The ATP produced is measured by hexokinase (HK)/glucose-6-phosphate dehydrogenase (G6PD)—coupled reactions that ultimately convert $NADP^+$ to NADPH, which is monitored spectrophotometrically at 340 nm. The assay is optimized by adding (1) *N*-acetylcysteine to activate CK, (2) ethylenediaminetetraacetic acid (EDTA) to bind Ca^{2+} and to increase the stability of the reaction mixture, and (3) adenosine pentaphosphate (Ap_5A) in addition to AMP to inhibit adenylate kinase (AK). The International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) developed a reference procedure based on this reaction principle for the measurement of CK at 37°C.

Specimens for CK analysis include serum and plasma heparin. Anticoagulants other than heparin should not be used in collection tubes, because they inhibit CK activity. CK activity in serum is relatively unstable and is rapidly lost during storage. Average stabilities are less than 8 hours at room temperature, 48 hours at 4°C, and 1 month at −20°C. Therefore the serum specimen should be chilled to 4°C if the sample is not analyzed immediately, and stored at −80°C if analysis is delayed for longer than 30 days. A moderate degree of hemolysis is tolerated because erythrocytes contain no CK activity. However, severely hemolyzed specimens (free hemoglobin concentration >20 g/L) are unsatisfactory because enzymes and intermediates (AK, ATP, and glucose-6-phosphate) liberated from the erythrocytes may affect the lag phase and side reactions occurring in the assay system.

For a single laboratory, the analytical specifications for desirable performance of CK assays, derived from biological variation of the enzyme, are as follows: imprecision (CV) ≤7.3%, bias no more than ±8.7%, and total error lower than ±20.6%.

Serum CK activity is subject to a number of physiological variations. It is influenced by (1) sex, (2) age, (3) race, (4) muscle mass, and (5) physical activity. Men have higher values than women, and blacks have higher values than nonblacks. In white subjects, the reference interval was found to be 46 to 171 U/L for men and 34 to 145 U/L for women when measured with an assay traceable to the IFCC 37°C reference procedure. Newborns generally have higher CK activity, resulting from skeletal muscle trauma during birth. Exercise, particularly if unaccustomed, can increase serum CK activity, which may ascend as high as 10 times the URL within 24 hours of activity.

Separation and Quantification of Creatine Kinase Isoenzymes

Electrophoretic methods are useful for the separation of all CK isoenzymes, but they are no longer routinely used. The isoenzyme bands are visualized by incubating the support (e.g., agarose or cellulose acetate) with a concentrated CK assay mixture using the reverse reaction. NADPH formed in this reaction is then detected by observing the bluish-white fluorescence after excitation by long-wave (360 nm) ultraviolet light. The mobility of CK isoenzymes at pH 8.6 toward the anode is BB > MB > MM, with the MM remaining cathodic to the point of application. The discriminating power of electrophoresis also allows the detection of abnormal bands (e.g., macro-CKs).

Immunoassays are applicable to the direct measurement of CK-MB. Immunoinhibition techniques measuring the catalytic activity of the B subunit of CK dimer were first introduced, but now their use should be discouraged for the lack of analytical specificity and sensitivity. Currently the most common approach is to measure concentrations of the CK-MB protein ("mass") by using "sandwich" immunoassays with

monoclonal antibodies specifically recognizing the MB dimer. Mass assays are sensitive with a limit of detection for CK-MB of less than 1 μg/L. With CK-MB mass assays, the URL for males is 5.0 μg/L, with values for females being less than male values.

Aldolase

Aldolase (ALD) (EC 4.1.2.13; D-fructose-1,6-diphosphate D-glyceraldehyde-3-phosphate-lyase) catalyzes the splitting of D-fructose-1,6-diphosphate to D-glyceraldehyde-3-phosphate and dihydroxyacetone-phosphate, an important reaction in the glycolytic breakdown of glucose to pyruvate. ALD is probably present in all cells, although it occurs in particularly large quantities in the muscles, liver, and brain.

Serum ALD determinations have been proposed for diagnosing diseases of skeletal muscle. In general, however, measurement of ALD activity in the serum of subjects with suspected muscle disease does not add information to that available more readily from measurement of CK. Accordingly, ALD measurement within clinical laboratories should be discouraged.

LIVER ENZYMES

Enzymes in this category include (1) alanine **aminotransferase** (ALT), (2) aspartate aminotransferase (AST), (3) alkaline phosphatase (ALP), (4) γ-glutamyltransferase (GGT), and (5) 5′-nucleotidase (NTP). The aminotransferases, ALP and γ-GGT are widely used and available on automated analyzers. They have long been mistakenly considered as a part of "liver function tests." They are not tests of liver function, but the habit sometimes persists.

The most common alterations in liver enzyme activities encountered in clinical practice can be divided into two major pathophysiology subgroups: (1) hepatocellular damage (increased aminotransferase activities) and (2) cholestasis (increased ALP and γ-GGT activities), although certain liver diseases may display a mixed biochemical picture (Fig. 19.1).

Aminotransferases

The aminotransferases constitute a group of enzymes that catalyze the interconversion of amino acids to 2-oxo-acids by transfer of amino groups. AST (EC 2.6.1.1; L-aspartate: 2-oxoglutarate aminotransferase) and ALT (EC 2.6.1.2; L-alanine: 2-oxoglutarate aminotransferase) are the aminotransferases of clinical interest.

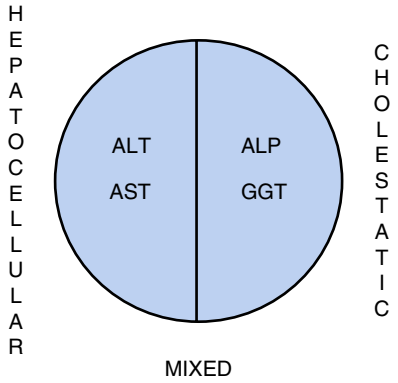
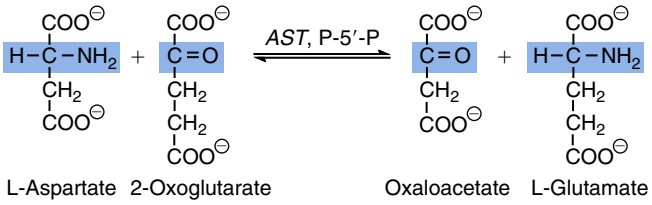
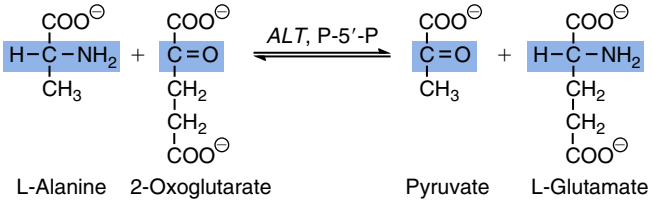


Fig 19.1 Liver enzymology patterns. *ALP*, Alkaline phosphatase; *ALT*, alanine aminotransferase; *AST*, aspartate aminotransferase; *GGT*, γ-glutamyltransferase.

The 2-oxoglutarate/L-glutamate couple serves as one amino group acceptor and donor pair in all amino-transfer reactions; the specificity of the individual enzymes derives from the particular amino acid that serves as the other donor of an amino group. Thus AST catalyzes the following reaction:



ALT catalyzes the analogous reaction as follows:



The reactions are reversible, but the equilibria of the AST and ALT reactions favor formation of aspartate and alanine, respectively.

Pyridoxal-5′-phosphate (P-5′-P) and its amino analogue, pyridoxamine-5′-phosphate, function as **coenzymes** in in vivo amino-transfer reactions. The P-5′-P is bound to the inactive **apoenzyme** and serves as a true **prosthetic group**. P-5′-P bound to the apoenzyme accepts the amino group from the first substrate, aspartate or alanine, to form enzyme-bound pyridoxamine-5′-phosphate and the first reaction product, oxaloacetate or pyruvate, respectively. The coenzyme in amino form then transfers its amino group to the second substrate, 2-oxoglutarate, to form the second product, glutamate. P-5′-P is thus regenerated.

Both coenzyme-deficient apoenzymes and **holoenzymes** may be present in serum. Thus the addition of P-5′-P under measurement conditions that allow recombination with the enzymes usually produces an increase in aminotransferase activity. For clinical assays, in accordance with the principle that all factors affecting the rate of reaction must be optimized and controlled, the addition of P-5′-P in aminotransferase assays is recommended to ensure that all enzymatic activity is measured.

Biochemistry

AST is found in the (1) heart, (2) liver, (3) skeletal muscle, and (4) kidney, whereas ALT is found primarily only in the liver and kidney (Table 19.3). ALT is exclusively cytoplasmic, but both mitochondrial and cytoplasmic forms of AST are found in cells. These are genetically distinct AST isoenzymes. About 5% to 10% of the activity of total AST in serum from healthy individuals is of mitochondrial origin.

TABLE 19.3 Aminotransferase Activities in Human Tissues, Relative to Serum as Unity

	Aspartate Aminotransferase	Alanine Aminotransferase
Heart	7800	450
Liver	7100	2850
Skeletal muscle	5000	300
Kidneys	4500	1200
Pancreas	1400	130
Spleen	700	80
Lungs	500	45
Erythrocytes	15	7
Serum	1	1

Clinical Significance

Liver disease is the most important cause of increased aminotransferase activity in serum and represents the indication for the measurement requests. Although serum activities of both AST and ALT become increased whenever disease processes affect liver cell integrity, ALT is the more liver-specific enzyme. Serum increases of ALT activity are rarely observed in conditions other than parenchymal liver disease. Moreover, increases of ALT activity persist longer than those of AST activity. Thus the incremental benefit of routine determination of AST, in addition to ALT, is limited. Ideally laboratories reporting abnormal ALT results (e.g., >URL) should offer AST as a reflex test and calculate the AST-to-ALT ratio because it provides useful diagnostic and prognostic information (see below).

In most types of liver disease, ALT activity is higher than that of AST; exceptions may be seen in (1) alcoholic hepatitis, (2) hepatic cirrhosis, and (3) liver neoplasia. In viral hepatitis and other forms of liver disease associated with acute hepatic necrosis, serum AST and ALT activities are increased even before the clinical signs and symptoms of disease (e.g., jaundice) appear. Activities for both enzymes may reach values as high as 100 times the URL, although 10-fold to 40-fold increases are most frequently encountered. The most efficient ALT threshold for diagnosing acute liver injury lies at seven times the URL (sensitivity and specificity >95%). Peak values of aminotransferase activity occur between the 7th and 12th days; activities then gradually decrease, reaching physiological concentrations by the 3rd to 5th week if recovery is uneventful. Peak activities bear no relationship to prognosis and may fall with worsening of the patient's condition due to lack of further functional hepatocytes to continue enzyme release.

Persistence of increased ALT for longer than 6 months after an episode of acute hepatitis is used to diagnose chronic hepatitis. ALT has been reported persistently normal in 15% to 50% of patients with chronic hepatitis C, but the likelihood of continuously normal ALT decreases with an increasing number of measurements. Therefore, in patients with acute hepatitis C, ALT should be measured periodically over the next 1 to 2 years to determine if it becomes and stays normal.

An AST-to-ALT ratio greater than or equal to 1 has approximately 90% positive predictive value for diagnosing the presence of advanced fibrosis in patients with chronic liver disease. Furthermore, the amount of increase in the ratio can reflect the grade of fibrosis in these patients.

Nonalcoholic fatty liver disease (NAFLD) is the most common cause of aminotransferase increases, other than viral and alcoholic hepatitis. NAFLD includes a spectrum of liver pathology, from simple steatosis to nonalcoholic steatohepatitis (NASH), in which inflammatory changes and focal necrosis may progress to (1) liver fibrosis, (2) cirrhosis, and (3) hepatic failure. NAFLD is now considered to be an additional feature of the "metabolic syndrome" with serum aminotransferase elevation associated with (1) higher body mass index (BMI), (2) increased waist circumference, (3) elevated serum triglycerides, (4) elevated fasting insulin, and (5) lower HDL cholesterol—these features are characteristics of this syndrome.

Two- to fivefold increases of aminotransferases may occur in patients with primary or metastatic carcinoma of the liver, with AST usually being higher than ALT. Slight or moderate elevations of AST and ALT activities have been observed after administration of various medications, such as (1) nonsteroidal antiinflammatory drugs, (2) antibiotics, (3) antiepileptic drugs, (4) statins, or (5) opiates. Over-the-counter medications and herbal preparations are also implicated. In patients with increased aminotransferases, negative viral markers, and a negative history for drugs or alcohol ingestion, the diagnostic workup should include less common causes of chronic hepatic injury, such as (1) autoimmune hepatitis, (2) hemochromatosis, (3) Wilson disease, (4) primary biliary cirrhosis, (5) sclerosing cholangitis, (6) celiac disease, and (7) α_1-antitrypsin deficiency.

As might be expected from the high AST concentration in muscle, AST activity also is increased in muscular dystrophy and dermatomyositis, although it is normal in muscular diseases of neurogenic origin. Finally, significant AST increases are noted in hemolytic disease.

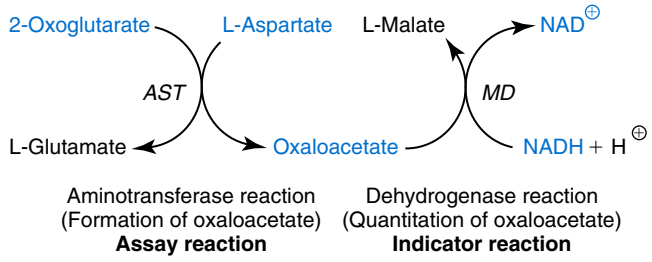
Several studies have described AST linked to immunoglobulins, or macro-AST. Typical findings include a persistent increase in serum AST activity with normal ALT concentrations in an asymptomatic subject, with the absence of any demonstrable pathology in organs rich in AST. Increased AST activity might reflect decreased clearance of the abnormal complex from plasma. Macro-AST has no known clinical relevance. However, identification is important to avoid unnecessary diagnostic procedures in these subjects. The demonstration of macro-AST in serum is obtained by differential precipitation with polyethylene glycol (PEG) 6000 (see the Amylase section later in this chapter).

POINTS TO REMEMBER

- Routine testing of AST in addition to ALT is not recommended, because ALT is the more liver-specific enzyme and the incremental benefit of routine determination of AST, in addition to ALT, is very limited. Moreover, increases of ALT activity persist longer than AST.
- Laboratories should offer AST as a reflex test in samples with abnormal ALT results (e.g., higher than the upper reference limit [URL]), also reporting the AST-to-ALT ratio because it provides useful diagnostic and prognostic information.

Methods of Analysis

Continuous-monitoring methods are used to measure aminotransferase activity by coupling aminotransferase reactions to specific dehydrogenase reactions. The oxo-acids formed in the aminotransferase reaction are measured indirectly by enzymatic reduction to corresponding hydroxy acids, and the accompanying change in NADH concentration is monitored spectrophotometrically. Thus oxaloacetate, formed in the AST reaction, is reduced to malate in the presence of malate dehydrogenase (MD).



Pyruvate formed in the ALT reaction is reduced to lactate by **lactate dehydrogenase** (LDH). The substrate, NADH, and an auxiliary enzyme, MD or LDH, are present in sufficient quantities, so that the reaction rate is limited only by the amounts of AST and ALT, respectively. As the reactions proceed, NADH is oxidized to NAD^+. The disappearance of NADH is followed by measuring the decrease in absorbance at 340 nm. The change in absorbance per minute (ΔA/min) is proportional to the micromoles of NADH oxidized and in turn to micromoles of substrate transformed per minute. A preliminary incubation period is necessary to ensure that NADH-dependent reduction of endogenous oxo-acids in the sample is completed before 2-oxoglutarate is added to start the aminotransferase reaction. As already mentioned, supplementation with P-5′-P ensures that all aminotransferase activity of the sample is measured.

Primary IFCC reference procedures are available for the measurement of catalytic activity concentrations of AST and ALT. To ensure accuracy and comparability of aminotransferase results among laboratories, values assigned to the manufacturer's product calibrators and measurement results obtained with commercial measuring systems in daily routine practice should be traceable to these top-level reference measurement procedures. It should be remembered that the concept of the reference measurement system is valid only if the reference procedure and corresponding commercial procedures have identical, or at least very similar, selectivity for the measured enzyme. Thus it will not be possible to calibrate procedures for aminotransferases that do not incorporate P-5′-P using a procedure that does, such as the IFCC reference procedure, because the ratio of preformed holoenzyme to apoenzyme differs among specimens.

AST activity in serum is stable for up to 48 hours at 4°C. Conversely, ALT activity should be assayed on the day of sample collection because activity is lost at room temperature, at 4°C, and at −25°C. ALT stability is better maintained at −70°C. Hemolyzed specimens should not be used, especially when AST is measured, because of the large amount of this enzyme present in red cells.

Biological variation data suggest that imprecision (CV) ≤4.8%, bias lower than ±5.6%, and total error lower than ±13.4% for AST, and imprecision (CV) ≤4.7%, bias lower than ±7.4%, and total error lower than ±15.1% for ALT, respectively, are required for the clinical use of aminotransferase determinations. The lack of P-5′-P addition in the employed commercial reagents is the most frequent cause of unacceptably biased results of aminotransferases.

Using assays traceable to the IFCC reference procedures, the AST URL for adults is 35 U/L, with no significant sex-related differences. Conversely, a clear difference in ALT activities has been noted between men and women. Corresponding ALT URLs are 59 U/L and 41 U/L, respectively. Using a more stringent criterion of selection by enrolling only individuals with BMI less than 25 kg/m^2, the resulting URLs were 49 U/L for men and 33 U/L for women, respectively. ALT does not reveal a distinct age dependency during childhood, but serum AST activity in neonates and in children younger than 3 years old is twice that in adults.

Alkaline Phosphatase

Alkaline phosphatase (ALP) (EC 3.1.3.1; orthophosphoric-monoester phosphohydrolase [alkaline optimum]) catalyzes the alkaline hydrolysis of a large variety of naturally occurring and synthetic substrates. Divalent ions, such as Mg^{2+}, Co^{2+}, and Mn^{2+}, are activators of the enzyme, and Zn^{2+} is a constituent metal ion. Inhibitors of ALP activity include (1) phosphate, (2) borate, (3) oxalate, and (4) cyanide ions. The type of buffer present affects the rate of enzyme activity. Accordingly, buffers can be classified as (1) inert (carbonate and barbital), (2) inhibiting (glycine and propylamine), or (3) activating (2-amino-2-methyl-1-propanol [AMP], tris [hydroxymethyl] aminomethane [TRIS], and diethanolamine [DEA]).

Biochemistry

ALP activity is present in most organs of the body and is especially associated with cell surfaces located in the mucosa of the small intestine and the proximal convoluted tubules of the kidney, in bone (osteoblasts), liver, and placenta, anchored on the cell membrane by glycosylphosphatidylinositol ("ectoenzyme"). Although the exact metabolic function of the enzyme is not yet understood, it appears that ALP is associated with lipid transport in the intestine and with the calcification process in bone.

ALP exists in multiple forms, some of which are true isoenzymes, encoded at separate genetic loci. Bone, liver, and kidney ALP forms share a common primary structure coded for by the same genetic locus (as tissue-nonspecific ALP), but they differ in their unique carbohydrate content.

The ALP activity present in the sera of healthy adults originates mainly in the liver, with most of the rest coming

from the skeleton. The respective contributions of these two forms to the total activity are age dependent. Minimal amounts of intestinal ALP may also be present, particularly in the sera of individuals of blood group B or 0. Because intestinal ALP activity in serum may increase after a meal, ALP should be measured preferentially in fasting subjects.

Clinical Significance

Increases in serum ALP activity commonly originate from one or both of two sources: liver and bone. Consequently, serum ALP measurements are of interest in the investigation of hepatobiliary disease and bone disease associated with increased osteoblastic activity.

Hepatobiliary disease. The response of the liver to any form of biliary tree obstruction induces the synthesis of ALP by hepatocytes. The newly formed ectoenzyme is released from cell membrane by the action of bile salts and enters the circulation to increase the enzyme activity in serum. The increase tends to be more notable in extrahepatic obstruction (e.g., by stone, by cancer of the head of the pancreas) than in intrahepatic obstruction, and is greater the more complete the obstruction. Serum enzyme activities may reach 10 to 12 times the URL and usually return to baseline on surgical removal of the obstruction. A similar increase is seen in patients with advanced primary liver cancer or widespread secondary hepatic metastases. GGT activity is co-elevated with ALP in cases of obstructive intrahepatic or posthepatic biliary disease.

Liver diseases that principally affect parenchymal cells, such as infectious hepatitis, typically show only moderately (less than threefold) increased or even normal serum ALP activities. Increases may also be a consequence of a reaction to drug therapy, and ALT/ALP-based criteria to discriminate the type of liver injury in drug-induced hepatic toxicity have been recommended (Table 19.4). Intestinal ALP isoenzyme, an asialoglycoprotein normally cleared by the hepatic asialoglycoprotein receptors, is often elevated in patients with liver cirrhosis.

Bone disease. Bone ALP is produced by the osteoblast and has been demonstrated in matrix vesicles deposited as "buds" derived from the cell's membrane. The enzyme therefore is an excellent indicator of global bone formation activity. Genetic inability to produce tissue-nonspecific ALP, including bone isoform, a rare inherited disorder known as hypophosphatasia, results in severe bone disease and impaired bone growth.

Among the bone diseases, the highest concentrations of (bone) ALP are encountered in **Paget disease** (osteitis deformans) as a result of the action of osteoblastic cells as they try to rebuild bone that is being resorbed by uncontrolled activity of osteoclasts. Values from 10 to 25 times the URL are not unusual, and in broad terms, the increase reflects the extent of disease. In vitamin D deficiency (osteomalacia and rickets), enzyme concentrations two to four times the URL may be observed, and these fall to baseline on treatment. Hyperparathyroidism (primary and secondary) is associated with slight to moderate increases of bone ALP in serum, with the existence and degree of increase reflecting the presence and extent of skeletal involvement. Very high enzyme concentrations are present in patients with osteogenic bone cancer. Increased ALP indicates bone metastasis in 70% of prostate cancers, and its measurement has been included in the European Association of Urology guidelines for staging this neoplasia. Bone ALP can be slightly increased in osteoporosis, but individuals with osteoporosis are not clearly distinguished from age-matched controls. Transient increases of ALP may be found during the healing of bone fractures. Physiological bone growth increases bone ALP in serum, and this accounts for the fact that in the sera of growing children, the enzyme concentration is 1.5 to 7 times that in healthy adult serum, the maximum being reached earlier in girls than in boys.

Other conditions that increase alkaline phosphatase. An increase of up to two to three times URL is observed in women in the third trimester of pregnancy, with the additional enzyme being of placental origin. Reports have described a benign familial increase in serum ALP activity caused by increased concentrations of intestinal ALP. Transient, benign increases in serum ALP may be observed in infants and children, with changes often more than 10 times the URL. These changes seem to reflect a reduction in the removal of ALP from blood caused by transient modifications of enzyme glycosylation.

ALP forms essentially identical to the normal placental or germ cell isoenzymes appear in the sera of some patients with malignant disease. This appears to result from de-repression of the corresponding genes. The presence of these isoenzymes can be readily detected in the serum by their increased stability at 65°C.

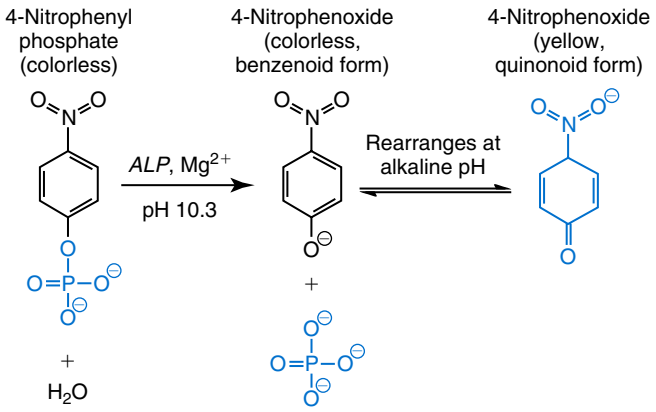
TABLE 19.4 Enzyme-Based Criteria for Drug-Induced Liver Toxicity (Enzyme Values Expressed as Multiples of the Upper Reference Limit in Serum)

Injury Type	ALT	ALP	ALT/ALP Ratio
Hepatocellular	>2 URL	<URL	≥5
Cholestatic	<URL	>2 URL	≤2
Mixed	>2 URL	>2 URL	2–5

ALP, Alkaline phosphatase; *ALT*, alanine aminotransferase; *URL*, upper reference limit.
Data from the Council of International Organizations of Medical Sciences, 1990.

Methods of Analysis for Total Alkaline Phosphatase Activity and Isoenzyme Content

The most popular of the chromogenic substrates for ALP is 4-nitrophenyl phosphate (usually abbreviated 4-NPP, or PNPP from the older name, *p*-nitrophenyl phosphate). This ester is colorless, but the final product is yellow at the pH of the reaction:



The enzymatic reaction is continuously monitored by observing the rate of formation of the 4-nitrophenoxide ions at 405 nm. This reaction forms the basis of current methods of ALP assay. The liberated phosphate group is transferred to water, and the rate of phosphatase action is enhanced if certain amino alcohols are used as phosphate-accepting buffers. Among these activators are compounds such as (1) AMP, (2) DEA, (3) TRIS, and (4) *N*-methyl-D-glucamine (MEG). The IFCC-recommended procedure uses 4-NPP as the substrate and AMP as the phosphate-acceptor buffer.

Serum or heparinized plasma, free of hemolysis, should be used. Complexing anticoagulants—such as citrate and EDTA—must be avoided because they bind cations, such as Mg^{2+} and Zn^{2+}, which are necessary cofactors for ALP activity. Freshly collected serum samples should be kept at room temperature and assayed as soon as possible, but preferably within 4 hours after collection. In sera stored at a refrigerated temperature, ALP activity increases slowly (2% per day). Frozen specimens should be thawed and kept at room temperature for 18 to 24 hours before measurement to achieve full enzyme reactivation.

Desirable analytical specifications for ALP determination are imprecision (CV) less than or equal to 3.0%, bias lower than ±5.5%, and total error lower than ±14%. As for other clinically important enzymes, the current emphasis is on the standardization and traceability of the commercial assays to the IFCC reference measurement system for ALP.

ALP activities in serum vary with age and, in a minor amount, with sex. Infants and peripubertal children show higher ALP activity (up to threefold) than healthy adults as a result of the leakage of bone ALP from osteoblasts during bone growth. However, activities in growing children are highly variable, and the decrease in ALP activity to typical adult ranges is known to differ from subject to subject, occurring on average 2 years earlier in females than in males. Using assays traceable to the IFCC reference procedure, the following ALP reference intervals have been established in adult individuals: 33 to 98 U/L for premenopausal women and 43 to 115 U/L for men.

Assays for ALP isoenzymes are needed when (1) the source of an increased ALP in serum is not obvious and should be clarified, (2) the main clinical question is concerned with detecting the presence of liver or bone involvement, or (3) it is important to ascertain any modifications in the activity of osteoblasts to monitor disease activity and the effects of appropriate therapies in the case of metabolic bone disorders. Criteria that have been used to differentiate the isoenzymes and other multiple forms of ALP include (1) electrophoretic mobility, (2) stability to denaturation by heat or chemicals, (3) response to the presence of selected inhibitors, (4) affinity for specific lectins, and (5) immunochemical characteristics.

After electrophoresis, ALP zones are made visible by incubating the gel in a solution of buffered substrate. The liver ALP typically moves most rapidly toward the anode. Bone ALP, which typically gives a more diffuse zone than the liver form, has slightly reduced anodal mobility—although the two zones usually overlap to some extent. Whereas intestinal ALP migrates more slowly than the bone enzyme, the placental isoenzyme commonly appears as a discrete band overlying the diffuse bone fraction. An additional band, which is frequently present in the serum of patients with various hepatic diseases, contains a high molecular weight form of ALP that is strongly negatively charged. Therefore it moves slowly in or may even fail to enter polyacrylamide gel, but it migrates more anodally than the main liver zone on nonsieving media, such as cellulose acetate. This form corresponds to the main liver form attached to the membrane moiety. Complexes between ALP and immunoglobulins, or macro-ALP, occur occasionally in serum, giving rise to abnormally migrating bands in the γ-globulin zone; however, they do not provide specific clinical information in the present state of knowledge. A short (15 minutes at 37°C) sample pretreatment with neuraminidase to remove a portion of the terminal sialic acid residues can be used to improve the electrophoretic separation between bone and liver ALPs, exploiting differences in the carbohydrate portions of the two forms of ALP.

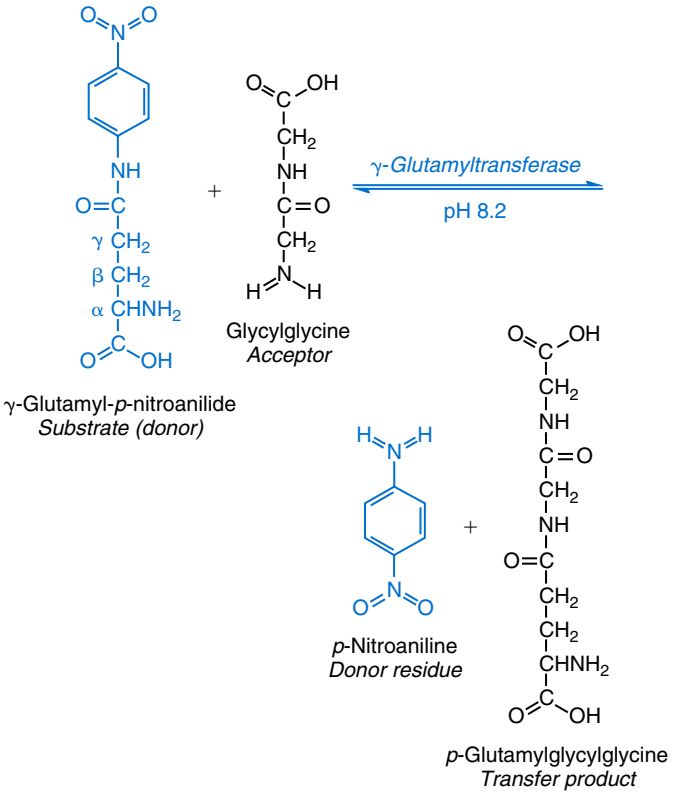
As an alternative to electrophoretic fractionation of ALP, measurement of γ-GGT, which is increased in liver disease but not in bone disease, may be a useful rapid tool to distinguish between the two diseases as the explanation for increased serum ALP.

Immunoassays for direct determination of bone ALP, which measure enzyme activity or mass concentration, are commercially available; cross-reactivity with the liver form by 6% to 20% has however been established, and their value has not been convincingly demonstrated, partly because measurements of total ALP provide the required clinical information in many situations.

γ-Glutamyltransferase

Peptidases are enzymes that catalyze the hydrolytic cleavage of peptides to form amino acids or smaller peptides. They constitute a broad group of enzymes of varied specificity, and some individual enzymes act as amino acid transferases and catalyze the transfer of amino acids from one peptide to another amino acid or peptide. **γ-Glutamyltransferase** (GGT) (EC 2.3.2.2; γ-glutamyl-peptide: amino acid γ-GGT) catalyzes the transfer of the γ-glutamyl group from peptides

and compounds to an acceptor. Substrates for GGT include (1) the γ-glutamyl acceptor, (2) an amino acid or peptide, or (3) even water, in which case simple hydrolysis takes place. The enzyme acts only on peptides or peptide-like compounds containing a terminal glutamate residue joined to the remainder of the compound through the terminal (-γ-) carboxyl. Glycylglycine is five times more effective as an acceptor than is glycine or the tripeptide (gly-gly-gly). An example of a reaction catalyzed by the GGT is shown here:



Biochemistry

GGT is present (in decreasing order of abundance) in the (1) proximal renal tubules, (2) liver, (3) pancreas, and (4) intestine. The enzyme is present in cytoplasm (microsomes), but the larger fraction is located in the cell membrane and may transport amino acids and peptides into the cell across the cell membrane in the form of γ-glutamyl peptides. GGT is critical for the maintenance of adequate intracellular concentrations of reduced glutathione, a major antioxidant agent.

Clinical Significance

Even though renal tissue has the highest concentration of GGT, the enzyme present in serum appears to originate primarily from the hepatobiliary system. Thus GGT is considered a sensitive indicator of the presence of hepatobiliary disease, even if its usefulness is limited by lack of specificity. Like ALP, GGT activity is highest in cases of intrahepatic or posthepatic biliary obstruction, reaching activities some 5 to 30 times the URL. High increases of GGT are also observed in patients with a primary or metastatic liver neoplasm. Moderate increases occur in infectious hepatitis. In acute and chronic pancreatitis and in some pancreatic malignancies (especially if associated with hepatobiliary obstruction), enzyme activity may be 5 to 15 times the URL.

Increased activities of GGT also are found in the sera of patients with alcoholic hepatitis and in the majority of sera from people who are heavy drinkers. GGT is also increased with increased body weight and obesity, and the effect of alcohol is more marked in these groups. Recent epidemiologic evidence has shown that serum GGT activity possesses an independent prognostic value for cardiovascular morbidity and mortality.

Increased concentrations of the enzyme are also found in the serum of subjects receiving anticonvulsant drugs such as phenytoin and phenobarbital. Such an increase in GGT activity in serum may reflect induction of new enzyme activity by the action of alcohol and drugs or their toxic effects on microsomal structures in liver cells.

Unlike ALP, serum GGT is not increased in conditions in which osteoblastic activity is increased, so the enzyme measurement can be useful in differentiating the source of ALP activity increase in serum, whether it is of bone or liver origin.

Methods of Analysis

Early GGT assays used L-γ-glutamyl-*p*-nitroanilide (GGPNA) as the substrate, with glycylglycine serving as the γ-glutamyl residue acceptor. The *p*-nitroaniline produced in the reaction is determined by its yellow color, which is monitored at 405 nm. However, GGPNA has limited solubility in the reaction mixture, and it is therefore difficult to obtain saturating substrate concentrations. To obviate this problem, derivatives of GGPNA, in which various groups have been introduced into the benzene ring to increase solubility in water, have been used. The most useful of these substrates is L-γ-glutamyl-3-carboxy-4-nitroanilide, which is readily soluble in water and is split by GGT at a rate comparable with that observed with GGPNA. In the IFCC reference measurement procedure for GGT, L-γ-glutamyl-3-carboxy-4-nitroanilide serves as the substrate, with glycylglycine serving as an acceptor. Buffering is provided by glycylglycine itself.

GGT activity is stable for at least 1 month at 4°C and for 1 year at −20°C. Nonhemolyzed serum is the preferred specimen, but EDTA plasma has also been used. Heparin may produce turbidity in the reaction mixture; citrate, oxalate, and fluoride depress GGT activity by 10% to 15%.

Desirable analytical specifications for GGT determination are imprecision (CV) ≤3.7%, bias lower than ±11.8%, and total error lower than ±22.2%.

In adults, the URL for GGT activity in serum is 40 U/L for females and 70 U/L for males, when measured with an assay traceable to the IFCC reference procedure. Reference limits are approximately twofold higher in people of African ancestry. In normal full-term neonates, GGT activity at birth is approximately six to seven times the adult reference interval. The activity then declines, reaching adult values by the age of 5 to 7 months.

5′-Nucleotidase

5′-Nucleotidase (NTP) (EC 3.1.3.5; 5′-ribonucleotide phosphohydrolase) is a phosphatase that acts only on nucleoside-5′-phosphates, such as adenosine-5′-phosphate (AMP) and adenylic acid, releasing inorganic phosphate.

NTP is a glycoprotein that is widely distributed throughout the tissues of the body and is principally localized in the cytoplasmic membrane of the cells in which it occurs. Despite its ubiquitous distribution, serum NTP activities appear to reflect hepatobiliary disease with considerable specificity. Particularly, NTP is increased in those hepatobiliary diseases in which there is an interference with the secretion of bile. This may be due to extrahepatic causes (a stone or tumor occluding the bile duct), or it may arise from intrahepatic conditions, such as cholestasis caused by malignant infiltration of the liver or biliary cirrhosis. When parenchymal cell damage is predominant, as in infectious hepatitis, serum NTP activity is only moderately increased.

The assay of NTP activity has been considered of value as an addition to measurement of ALP in patients with suspected hepatobiliary disease, and increased NTP activity is interpreted as evidence of a hepatic origin of increased ALP activity in serum. However, approximately half of individuals in whom liver ALP activity is increased in serum may simultaneously show a normal NTP. On the other hand, increased NTP in the serum of patients with normal liver ALP is very often associated with the presence of liver disease. Thus the frequent dissociation of the two enzyme activities supports the usefulness of determining both ALP and NTP to enhance diagnostic efficiency in patients with suspected liver disease. Different patterns of ALP and NTP are, however, not associated with specific hepatic conditions.

In a commercially available assay, serum NTP catalyzes the hydrolysis of inosine-5′-phosphate (IMP) to yield inosine, which is then converted to hypoxanthine by purine-nucleoside phosphorylase. Hypoxanthine is oxidized to urate with xanthine oxidase. Two moles of hydrogen peroxide are produced for each mole of hypoxanthine liberated and converted to uric acid. The formation rate of hydrogen peroxide is monitored at 510 nm by the oxidation of a chromogenic system. The effect of ALP on IMP is inhibited by β-glycerophosphate. Using this method, the reference interval for NTP activity at 37°C is from 3 to 9 U/L, with no sex-related differences.

NTP activity in serum or plasma heparin is stable for at least 4 days at 4°C and 4 months at −20°C. Desirable analytical specifications for NTP determination are imprecision (CV): ≤2.1%, bias lower than ±2.9%, and total error lower than ±8.3%.

PANCREATIC ENZYMES

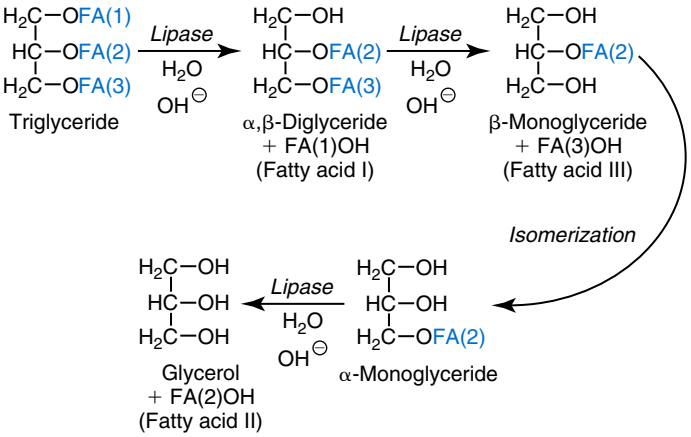
The most commonly used serum biomarkers for investigation of pancreatic disease, and more specifically **acute pancreatitis**, are LIP and P-type amylase (see Chapter 38).

Lipase

Human pancreatic **lipase** (LIP) (EC 3.1.1.3; triacylglycerol acylhydrolase) is a single-chain glycoprotein with a molecular weight of 48 kDa. For full catalytic activity and greatest specificity, the presence of bile salts and a cofactor called *colipase*, which is a small protein secreted by the pancreatic acinar cells, are required.

Biochemistry

Lipases are defined as enzymes that hydrolyze glycerol esters of long-chain fatty acids.



Only ester bonds at carbons 1 and 3 (α-positions) are attacked, and products of the reaction include 2 moles of fatty acids and 1 mole of 2-acylglycerol (β-monoglyceride) per mole of substrate. The latter is resistant to hydrolysis, but it can spontaneously isomerize to the α-form (3-acylglycerol). The isomerization permits the third fatty acid to be split off, but at a much slower rate.

LIP acts only when the substrate is present in an emulsified form at the interface between water and the substrate. The rate of LIP action depends on the surface area of the dispersed substrate. Bile acids ensure that the surface of the dispersed substrate remains free of other proteins, by lining the surface of the insoluble substrate and the aqueous medium.

Most LIP activity found in serum derives from the pancreatic acinar cells, but some is secreted by gastric and intestinal mucosa. LIP concentration in the pancreas is about 5000-fold greater than in other tissues, and the concentration gradient between the pancreas and serum is approximately 20,000-fold. LIP is a small enough molecule to be filtered through the glomerulus, but it is totally reabsorbed by the renal tubules, and it is not physiologically detected in urine.

Clinical Significance

LIP measurement in serum is the recommended laboratory test to diagnose acute pancreatitis. The clinical sensitivity is 80% to 100%, depending on the selected diagnostic cutoff, and the clinical specificity is 80% to 100%, depending on the mix of the patient population studied. After an attack of acute pancreatitis, serum LIP activity increases within 4 to 8 hours, peaks at about 24 hours, and decreases within 7 to 14 days (Fig. 19.2). Increases between 2 and 50 times the URL have been reported.

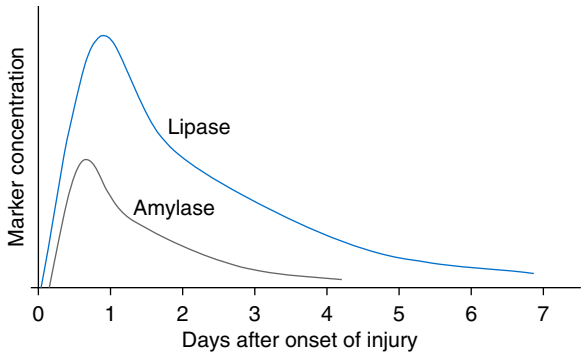


Fig 19.2 Time-dependent changes in serum amylase and lipase after acute pancreatitis.

Acute pancreatitis is sometimes difficult to diagnose because it must be differentiated from other acute intraabdominal disorders with similar clinical findings, such as (1) perforated gastric or duodenal ulcer, (2) intestinal obstruction, or (3) mesenteric vascular obstruction. A correct diagnosis of pancreatitis is vitally important, as the treatment for the other conditions mimicking pancreatitis typically involves surgery, while surgery is in general contraindicated in pancreatitis. In differential diagnosis, increase of serum LIP activity to greater than three times the URL, in the absence of renal failure, is a more specific diagnostic finding than increases in serum α-amylase activity. Furthermore, LIP concentrations remain increased longer than those of α-amylase do, which is another advantage over α-amylase measurement in patients with delayed presentation (see Fig. 19.2). Therefore it is recommended that LIP should replace α-amylase as the initial diagnostic test for acute pancreatitis in the emergency department; obtaining both serum α-amylase and LIP is not warranted.

In patients with a reduced glomerular filtration rate, serum LIP activity is increased. Thus care should be exercised in the interpretation of increased serum LIP values in the presence of renal disease.

Methods of Analysis

Many LIP methods have been described; they have used both triglyceride and nontriglyceride substrates and (1) titrimetric, (2) turbidimetric, (3) spectrophotometric, (4) fluorometric, and (5) immunologic techniques. In general, long-chain triglyceride (and some diglyceride) substrates have demonstrated correlation of results with the clinical state that is superior to that seen with methods using other substrates. Spectrophotometric methods have the advantage of minimum sample volume requirements, good precision, wide dynamic range, and ease of automation. Numerous substrates and complex auxiliary and indicator systems have been used.

In the enzymatic reaction rate diglyceride assay for LIP, the following sequence of indicator and auxiliary enzymes is used:

$$\text{1,2-Diacylglycerol} + H_2O \xrightarrow[\text{pH 8.7}]{\text{Pancreatic lipase, Colipase}} \text{2-Monoacylglycerol} + \text{Fatty acid}$$

$$\text{2-Monoacylglycerol} + H_2O \xrightarrow{\text{Monoglyceride lipase}} \text{Glycerol} + \text{Fatty acid}$$

$$\text{Glycerol} + \text{ATP} \xrightarrow{\text{Glycerol kinase}} \text{L-}\alpha\text{-glycerophospate} + \text{ADP}$$

$$\text{L-}\alpha\text{-glycerophospate} + O_2 \xrightarrow{\text{L-}\alpha\text{-glycerophospate kinase}} \text{Dihydroxyacetone phosphate} + H_2O_2$$

$$2\,H_2O_2 + \text{4-Aminoantipyrine} + \text{TOOS} \xrightarrow{\text{Peroxidase}} \text{Quinonediimine dye (colored)} + 2\,H_2O$$

TOOS is sodium *N*-ethyl-*N*-(2-hydroxyl-3-sulfopropyl)-m-toluidine, and its oxidation produces an intensely colored dye detectable at 550 nm.

In another method, a synthetic substrate (1,2-*O*-dilauryl-rac-glycero-3-glutaric acid-[4-methyl-resorufin]-ester) consisting of two glycerol ether bonds and one ester bond is used. LIP hydrolyzes the ester bond in an alkaline medium to an unstable dicarbonic acid ester that spontaneously hydrolyzes to yield glutaric acid and methylresorufin; this is a bluish-purple chromophore with peak absorption at 580 nm. The rate of methylresorufin formation is directly proportional to the LIP activity of the sample.

$$\text{1,2-Dioleoylglycerol} + H_2O \xrightarrow{\text{LIP}} \text{2-Monooleoylglycerol} + \text{oleic acid}$$

$$\text{2-Monooleoylglycerol} + H_2O \xrightarrow{\text{monoacylglycerol lipase}} \text{Glycerol} + \text{oleic acid}$$

$$\text{Glycerol} + \text{ATP} \xrightarrow{\text{glycerol kinase}} \text{ADP} + \text{glycerol-3-phosphate}$$

$$\text{ADP} + \text{glucose} \xrightarrow{\text{ADP-dependent hexokinase}} \text{Glucose-6-phosphate} + \text{AMP}$$

$$\text{Glucose-6-phosphate} + \text{NADP} \xrightarrow{\text{glucose-6-phosphate dehydrogenase}} \text{6-Phospho-glucono-1,5-lactone} + \text{NADPH}$$

LIP activity in serum is stable at room temperature for 1 week; sera may be stored for 3 weeks in the refrigerator and for several years if frozen.

Biological variation data suggest that imprecision (CV) less than or equal to 3.9%, bias lower than ±6.3%, and total error lower than ±12.7% for LIP measurements are required for clinical use.

Reference intervals for LIP activity are largely method dependent. For the enzymatic reaction rate diglyceride assay, the suggested URL is 45 U/L. The URL is 64 U/L for methylresorufin assay. There are no gender- or age-related differences.

Amylase

α-Amylases (AMY) (EC 3.2.1.1; 1,4-α-d glucan glucanohydrolase) are enzymes that catalyze the hydrolysis of 1,4-α-glucosidic linkages in polysaccharides. Both straight-chain (linear) polyglucans (amylose) and branched polyglucans (amylopectin and glycogen) are hydrolyzed, but at different rates. AMYs do not attack the α-1,6-linkages at the branch points. AMYs are calcium metalloenzymes, with the calcium essential for functional integrity. However, full activity is displayed only in the presence of various anions, with chloride and bromide being the most effective activators.

Biochemistry

AMYs have molecular weights varying from 54 to 62 kDa. The enzyme is thus small enough to pass through the glomeruli of the kidneys, and AMY is the only plasma enzyme normally found in urine. AMYs are present in many organs and tissues. The greatest concentration is noted in the salivary glands, which secrete a potent AMY (S-type) to initiate hydrolysis of starches while the food is still in the mouth and esophagus. In the pancreas, the enzyme (P-type) is synthesized by acinar cells and then is secreted into the intestinal tract by way of the pancreatic duct system. AMY activity is also found in extracts from the (1) ovaries, (2) fallopian tubes, (3) lungs, and (4) adipose tissue. Some tumors of the lung and ovary may also contain considerable AMY activity (usually S-type) and produce hyperamylasemia. Cases of AMY-producing multiple myeloma have also been described.

The enzyme present in serum and urine is predominantly of pancreatic (P-AMY) and salivary gland (S-AMY) origin. These isoenzymes are products of two closely linked loci on chromosome 1. In healthy adults, P-AMY represents approximately 40% to 50% of total AMY activity in serum. AMY isoenzymes also undergo posttranslational modification of (1) deamidation, (2) glycosylation, and (3) deglycosylation, to form a number of isoforms.

Clinical Significance

Blood AMY activity is physiologically low and greatly increases in acute pancreatitis and salivary gland inflammation. In acute pancreatitis, a rise in serum AMY activity occurs within 5 to 8 hours of symptom onset; activities typically return to baseline by the third or fourth day (see Fig. 19.2). The specificity of AMY for the diagnosis of acute pancreatitis is, however, low (20% to 60%, depending on the mix of the patient population studied) because increased values are also found in a number of acute intraabdominal disorders and in several extrapancreatic conditions (Table 19.5).

TABLE 19.5 Causes of Hyperamylasemia

Pancreatic disease	Pancreatitis, any cause (P-AMY↑)[a]
	Pancreatic trauma (P-AMY↑)
Intraabdominal diseases other than pancreatitis	Biliary tract disease (P-AMY↑)
	Intestinal obstruction (P-AMY↑)
	Mesenteric infarction (P-AMY↑)
	Perforated peptic ulcer (P-AMY↑)
	Gastritis, duodenitis (P-AMY↑)
	Ruptured aortic aneurysm
	Acute appendicitis (perforated)
	Peritonitis
	Trauma
Genitourinary disease	Ectopic, ruptured tubal pregnancy (S-AMY↑)
	Salpingitis (S-AMY↑)
	Ovarian malignancy (S-AMY↑)
	Renal insufficiency (Mixed)
Miscellaneous	Salivary gland lesions (S-AMY↑)
	Acute alcoholic abuse (S-AMY↑)
	Diabetic ketoacidosis (S-AMY↑)
	Macroamylasemia (S-AMY↑ or P-AMY↑)
	Septic shock (S-AMY↑)
	Cardiac surgery (S-AMY↑)
	Tumor (usually S-AMY↑)
	Drugs (usually S-AMY↑)

[a]Predominant isoenzyme type is shown in parentheses.
Mixed, Either or both isoenzymes may be present; *P-AMY,* pancreatic; *S-AMY,* salivary.

Lack of specificity of total AMY measurement has promoted the direct measurement of P-AMY instead of total enzyme activity for the differential diagnosis of patients with acute abdominal pain, making the measurement of total AMY obsolete. By applying the best decision limit (an activity equal to threefold the URL), the specificity of P-AMY for the diagnosis of acute pancreatitis is close to 90%. Sensitivity in late detection of this condition is also notably improved with P-AMY. P-AMY values remain increased in 80% of patients with uncomplicated pancreatitis 1 week after onset, when only 30% still show increased total AMY activity. This long-standing increase in P-AMY activity in serum also makes redundant the traditional measurement of total AMY in urine—a test performed to achieve better diagnostic sensitivity in the late phase of pancreatitis.

Various intraabdominal extrapancreatic events, such as acute **cholecystitis**, can lead to a significant increase in serum P-AMY activities. In renal insufficiency, serum P-AMY activity is increased in proportion to the extent of renal impairment.

In 1% of the population, macroamylases are present in sera and may cause hyperamylasemia; these are complexes of ordinary AMY (usually S-type) and IgG or IgA. These macroamylases cannot be filtered through the glomeruli of the kidneys because of their large size (>200 kDa molecular weight) and are thus retained in the plasma, where their presence may increase AMY activity some two- to eightfold above the URL, typically stable over time. No clinical symptoms are associated with this disorder.

POINTS TO REMEMBER

Pancreatic enzymes in acute pancreatitis:

- The diagnostic performance of pancreatic enzymes is greatly improved by restricting their use to a population with suspected disease.
- LIP measurement is superior to P-AMY in terms of diagnostic performance. Therefore it is recommended that LIP replaces P-AMY as an initial test for acute pancreatitis in emergency departments.
- Measuring both serum P-AMY and LIP is not warranted.
- The measurement of total AMY should be considered obsolete.

Methods of Analysis

The use of defined substrates in the AMY assay has improved the reaction stoichiometry and has led to more controlled and consistent hydrolysis conditions. Substrates used include (1) maltotetraose, (2) maltopentaose, and (3) 4-nitrophenyl (4-NP)-glycoside substrates prepared by bonding 4-NP to the reducing end of a defined oligosaccharide. If the oligosaccharide is maltoheptaose (G7), the substrate is then 4-NP-G7. AMY splits this substrate to produce free oligosaccharides (G5, G4, and G3) and 4-NP-G2, 4-NP-G3, and 4-NP-G4. Combined hydrolysis by AMY in the specimen and by the reagent α-glucosidase (EC 3.2.1.20; maltase) results in free NP production, which is detected by its absorbance at 405 nm. Problems arose with use of the 4-NP-glycoside assay with regard to the poor stability of the reconstituted assay mixture because of slow hydrolysis of the 4-NP-glycoside by α-glucosidase. This effect has been reduced by covalently linking a "blocking" group (i.e., a 4,6-ethylidene group; ethylidene-protected substrate [EPS]) to the nonreducing end of the molecule. The blocked substrate also shows a more advantageous hydrolysis pattern, thus increasing the liberation of 4-NP. As a result, cleavage of one α-glucosidic linkage by AMY results in the release of one molecule of 4-NP.

$$5\ \text{Ethylidene-4-NP-G}_7 + 5\ H_2O \xrightarrow{\alpha\text{-Amylase}} 2\ \text{Ethylidene-G}_5 + 2\ \text{4-NP-G}_2 + 2\ \text{Ethylidene-G}_4 + 2\ \text{4-NP-G}_3 + \text{Ethylidene-G}_3 + \text{4-NP-G}_4$$

$$2\ \text{4-NP-G}_2 + 2\ \text{4-NP-G}_3 + 10\ H_2O \xrightarrow{\alpha\text{-Glucosidase}} 4\ \text{4-NP} + 10\ \text{G}$$

Only methods based on selective S-AMY inhibition by monoclonal antibodies have shown sufficient (1) precision, (2) reliability, (3) practicability, and (4) analytical speed to be clinically useful for P-AMY determination. After the S-AMY activity is inhibited by the addition of antibodies, uninhibited P-AMY activity is measured using EPS-4-NP-G7 as a substrate. This assay is available in full automation today on clinical chemistry platforms with reagent costs similar to total AMY, which permits laboratories to abandon the latter.

Falsely increased P-AMY results have been reported in subjects with macroamylasemia, in whom immunoglobulin complexed to AMY forms; these complexes diminish or void the ability of monoclonal antibodies, included in the test, to efficiently inhibit S-AMY. In this condition, precipitation of the macrocomplex by a PEG 6000 solution (240 g/L) presents an easy and effective approach to confirm the suspicion of macroamylasemia. Residual P-AMY activity of less than 30% in the supernatant is indicative of macroamylasemia (Fig. 19.3).

Except for heparin, all common anticoagulants inhibit AMY activity because they chelate calcium. Therefore P-AMY assay should be performed only on serum or heparinized

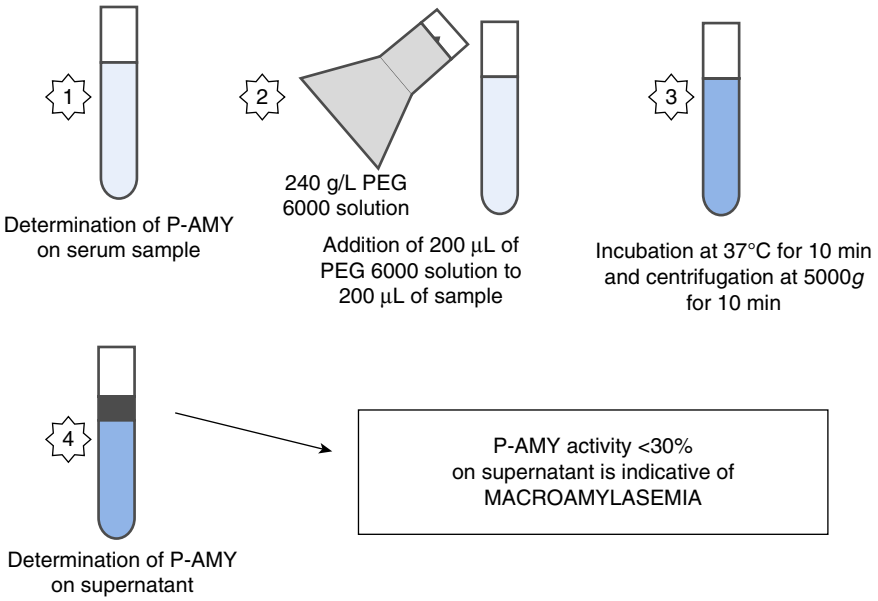


Fig 19.3 Demonstration of macroamylasemia by polyethylene glycol (PEG) 6000 solution. *P-AMY,* Pancreatic amylase.

plasma. AMY is quite stable; activity is fully retained during storage for 4 days at room temperature, for 2 weeks at 4°C, and for 1 year at −25°C.

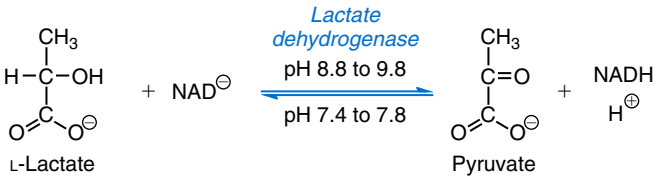
Biological variation data suggest that P-AMY measurements require an imprecision (CV) ≤3.2%, maximum bias of ±6.4%, and a total error lower than ±11.6% for good quality.

Using the immunoinhibition method, the reference interval for P-AMY activity in sera from adults is 13 to 53 U/L. Serum P-AMY activity is not demonstrable in most children younger than 6 months, but activity rises slowly thereafter to reach adult concentrations at 5 years of age, reflecting the postnatal development of exocrine pancreatic function. Therefore the use of this enzyme for diagnostic purposes in young children should be avoided.

OTHER CLINICALLY IMPORTANT ENZYMES

Lactate Dehydrogenase

LDH (EC 1.1.1.27; l-lactate: NAD^+ oxidoreductase) catalyzes the oxidation of l-lactate to pyruvate with the mediation of NAD^+ as a hydrogen acceptor.



As indicated, the reaction is reversible, and the reaction equilibrium strongly favors the reduction of pyruvate to lactate (P → L)—the "reverse reaction." Both pyruvate and lactate in excess inhibit enzyme activity, although the effect of pyruvate is greater. EDTA inhibits the enzyme perhaps by binding zinc ions, postulated as LDH activators.

Biochemistry

LDH has a molecular weight of 134 kDa and is composed of four peptide chains of two types, M (or A) and H (or B), each under separate genetic control. The subunit compositions of the five isoenzymes are LDH-1 (HHHH; H_4), LDH-2 (HHHM; H_3M), LDH-3 (HHMM; H_2M_2), LDH-4 (HMMM; HM_3), and LDH-5 (MMMM; M_4).

LD activity is present in many cells of the body and is invariably found only in the cytoplasm of the cell. Different tissues show different isoenzyme composition. In the heart, kidneys, and erythrocytes, isoenzymes LDH-1 and LDH-2 predominate, while LDH-4 and LDH-5 isoenzymes account for the enzyme activity in the liver and skeletal muscle.

Clinical Significance

Because of its wide tissue distribution, serum LDH increases occur in a variety of clinical conditions, including (1) myocardial infarction, (2) hepatitis, (3) hemolysis, and (4) disorders of the lung and muscle. Serum LDH measurement is, however, relevant only in hematology and oncology.

Hemolytic anemias significantly increase LDH concentrations in serum. Marked increases of LDH activity—up to 50 times the URL—have been observed in the megaloblastic anemias, usually resulting from the deficiency of folate or vitamin B_{12}. These increases rapidly return to normal after appropriate treatment. For monitoring purposes, LDH is relevant in predicting disease activity in leukemia, and the survival rate (probability of survival) and duration in Hodgkin disease and non-Hodgkin lymphoma.

Patients with malignant disease often show increased LDH activity in serum; up to 70% of patients with liver metastases, and 20% to 60% of patients with other nonhepatic metastases (e.g., lymph nodes) have increased LDH activity. Notably elevated LDH-1 is observed in germ cell tumors (≈60% of cases) such as seminoma of the testis and dysgerminoma of the ovary. LDH appears to be a useful predictor of outcome in patients with testicular tumor, melanoma, and small cell lung cancer.

Increases of LDH activity are observed in liver disease, but their clinical use appears limited and does not appear to add significantly to the aminotransferase activity investigation. LDH measured in pleural fluid (in combination with serum LDH) aids in distinguishing exudative from transudative effusions. Macro-LDH, which usually occurs as a result of the formation of an autoantibody-enzyme complex leading to a persistent increase in the amount of circulating enzyme, has been estimated to occur in fewer than 1 in 10,000 people.

Methods of Analysis

Methods for quantitation of LDH activity use kinetic spectrophotometry to measure the interconversion of the coenzyme NAD^+ and NADH at 340 nm. Procedures employing the L → P reaction are recommended because there is less dependence on the NAD^+ and lactate concentrations and less contamination of NAD^+ with inhibiting products.

Serum is the preferred specimen for measuring LDH activity. Plasma samples may be contaminated with platelets, which contain high concentrations of LDH. Serum should be separated from the clot as soon as possible after the specimen has been obtained. Hemolyzed serum must not be used because erythrocytes contain 4000 times more LDH activity than the serum. LDH-4 and LDH-5 isoenzymes are very sensitive to cold. Thus serum specimens should be stored at room temperature, at which no loss of activity occurs for at least 3 days.

The specifications for desirable performance of LDH assays, derived from biological variation of the enzyme, are as follows: imprecision (CV) ≤2.6%; bias lower than ±3.4%; and total error lower than ±7.7%.

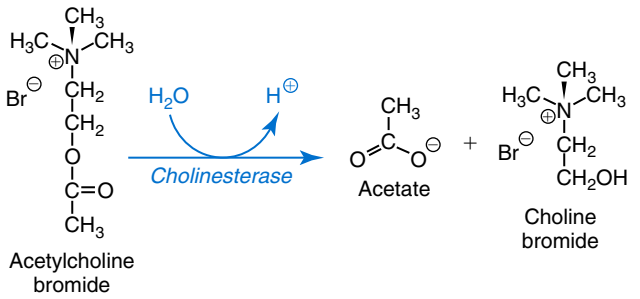
The reference interval for LDH activity in adults, applicable for assays traceable to the IFCC reference procedure, was found to be 125 to 220 U/L. LDH reference intervals are higher in children, with a gradual decrease noted over the whole childhood period.

Electrophoretic separation is the only procedure commercially available to demonstrate LDH isoenzymes. However, considering the described limited clinical use of the enzyme to hematology and oncology, LDH isoenzyme measurement is usually not routinely done.

Cholinesterase

Two related enzymes have the ability to hydrolyze acetylcholine. One is acetylcholinesterase (EC 3.1.1.7; acetylcholine acetylhydrolase), which is called *true cholinesterase* or *choline esterase I*. True **cholinesterase** is found in (1) erythrocytes, (2) the lungs and spleen, (3) nerve endings, and (4) the gray matter of the brain. It is responsible for the prompt hydrolysis of acetylcholine released at the nerve endings to mediate transmission of the neural impulse across the synapse. The second cholinesterase is acylcholine acylhydrolase (EC 3.1.1.8; acylcholine acylhydrolase), also called *pseudocholinesterase*, *serum cholinesterase* (CHE), *butyrylcholinesterase*, or *choline esterase II*. CHE is found in the (1) liver, (2) pancreas, (3) heart, (4) white matter of the brain, and (5) serum. Its exact biological role is unknown.

The type of reaction catalyzed by both cholinesterases is shown:



Biochemistry

Of clinical interest are the atypical (genetic) variants of the CHE, characterized by diminished activity against acetylcholine and other substrates, which are found in the sera of a small fraction of apparently healthy people. The gene controlling the synthesis of CHE can exist in many allelic forms. Four of the most common forms are designated as E^u, E^a, E^f, and E^s. These four allelic genes can be combined to form one normal and nine abnormal genotypes. The normal, most common genotype is designated as E^uE^u, or *UU* (*u* for *usual*). The gene E^a is referred to as the *atypical* gene; the sera of people homozygous for this gene ($E^aE^a = AA$) are only weakly active toward most substrates for CHE and demonstrate increased resistance to inhibition of enzyme activity by dibucaine. The E^f gene (*f* for *fluoride resistant*) gives rise to a weakly active enzyme but with increased resistance to fluoride inhibition. The E^s gene (*s* for *silent*) is associated with the absence of enzyme or the presence of a protein with minimal or no catalytic activity. Homozygous forms, *AA* and *FF*, are found in 0.3% to 0.5% of the white population; their incidence among blacks is even lower.

Clinical Significance

Measurements of CHE activity in serum are used (1) as a test of liver function, (2) as an indicator of possible insecticide poisoning, and (3) for the detection of patients with atypical forms of the enzyme who are at risk for prolonged responses to certain muscle relaxants used in surgical procedures.

In the absence of genetic causes or known inhibitors, any decrease in CHE activity reflects impaired synthesis of the enzyme by the liver. Among the organic phosphorous compounds that inhibit CHE activity are many insecticides, such as (1) parathion, (2) sarin, and (3) tetraethyl pyrophosphate. Workers in agriculture and in organic chemical industries may be subject to poisoning by inhalation of these materials or by direct contact with them.

Succinyldicholine (suxamethonium) and mivacurium, muscle relaxant drugs used in surgical procedures to aid in endotracheal intubation, are hydrolyzed by CHE, and their pharmacologic effect normally persists only long enough to meet the needs of the surgical procedure. In individuals with low enzyme activities or in those with a weakly active variant, destruction of the drug will not occur rapidly enough, and the subject may enter a period of prolonged paralysis of the respiratory muscles (apnea) requiring mechanical ventilation until the drug effects gradually wear off. Measurements of total CHE activity and determination of "dibucaine number" and "fluoride number" are needed to fully characterize CHE variants. The latter values indicate the percentage inhibition of enzyme activity in the presence of standard concentrations of dibucaine or fluoride. Mutation genotyping may confirm CHE gene abnormalities.

Methods of Analysis

Many of the contemporary methods use acylthiocholine esters as substrates. The iodide salts of (1) acetylthiocholine, (2) propionylthiocholine, (3) butyrylthiocholine, (4) benzoylthiocholine, and (5) succinylthiocholine all have been used. After the substrate hydrolysis, the thiocholine formed can be measured by reaction with chromogenic disulfide agents, such as 5,5′-dithio-bis(2-nitrobenzoate) (DTNB). The reaction of the thiocholine product with colorless DTNB forms colored 5-mercapto-2-nitro-benzoic acid, which is measured spectrophotometrically at 410 nm. Measuring CHE activity using succinyldithiocholine is the method of choice to diagnose succinylcholine sensitivity, purely based on the enzyme activity recorded in serum. At present, however, most automated instruments prefer the use of butyrylthiocholine as substrate for determining the CHE activity. Based on differences in sensitivity to inhibition by the local anesthetic dibucaine, a simple test was developed to classify the type of CHE as (1) usual, (2) intermediate, or (3) atypical ("dibucaine number"). Particularly, the usual CHE is inhibited by 80%, but atypical CHE is inhibited by only 20%. Subjects heterozygous for the normal and the atypical gene show about 60% inhibition of CHE. Recently, molecular methods to detect CHE variants have been developed and are being used increasingly in clinical laboratories.

Serum is the sample of choice. Enzyme activity in serum is stable for several weeks if the specimen is stored under refrigeration and for several years if stored at −20°C.

Desirable analytical specifications for CHE determination are imprecision (CV) less than or equal to 3.1%, bias lower than ±4.8%, and total error lower than ±9.8%.

Using the succinyldithiocholine/DTNB method, the reference interval for healthy adults with the usual CHE genotype was estimated to be 33 to 76 U/L for women and 40 to 78 U/L for men, respectively. The median activity in individuals with heterozygous genotype was 22 U/L (range, 5 to 35 U/L), and for atypical homozygotes 1.5 U/L (range, 1 to 4 U/L). The significant CHE decrease (≈30%) during pregnancy and early puerperium is explained by hemodilution.

Acid Phosphatase (Tartrate-Resistant 5b Isoform)

Under the name of **acid phosphatase** (ACP; EC 3.1.3.2; orthophosphoric-monoester phosphohydrolase [acid optimum]) are included all phosphatases with optimal activity below a pH of 7.0. ACP is present in lysosomes, which are organelles present in all cells with the possible exception of erythrocytes. Extralysosomal ACPs are also present in many cells. The greatest concentrations of ACP activity occur in the (1) prostate, (2) bone (osteoclasts), (3) spleen, (4) platelets, and (5) erythrocytes. The lysosomal and prostatic enzymes are strongly inhibited by dextrorotatory tartrate ions, but the erythrocyte and bone isoenzymes are not. Most of the physiologically low ACP activity of (unhemolyzed) serum is of a tartrate-resistant type (TR-ACP) and probably originates mainly in osteoclasts. Activities of this fraction are increased physiologically in growing children and pathologically in conditions of increased osteolysis and bone remodeling.

At least four ACP-determining genes have been identified and mapped. The erythrocyte ACP gene is located on chromosome 2, and a further gene on chromosome 19 encodes the TR-ACP expressed in osteoclasts and other tissue macrophages, such as alveolar macrophages and Kupffer cells (type 5 ACP). Isoenzyme 5 consists of two structurally related isoforms that differ by their carbohydrate content: TR-ACP 5a, which derives mainly from macrophages and dendritic cells, and type 5b, a more specific marker of osteoclastic activity. Genes encoding the tartrate-inhibited lysosomal and prostatic ACPs, mapped to chromosomes 11 and 13, respectively, exhibit considerable homology.

Serum TR-ACP is a potentially useful marker of conditions with a marked osteolytic component. Unlike blood concentrations of other markers of bone resorption (e.g., C-telopeptide of type I collagen), TR-ACP is not affected by renal dysfunction. However, TR-ACP appears to show relatively small dynamic changes in comparison with markers of bone resorption related to type I collagen metabolism. This may be attributable to the fact that the enzyme is released into the sealed osteoclast microenvironment, rather than directly into the circulation.

Immunoassays for serum TR-ACP have been developed that preferentially detect isoform 5b and are now commercially available. There is, however, the reality that both immunoassays are not entirely bone specific.

Serum should be immediately separated from erythrocytes and stabilized by the addition of 50 μL of acetic acid (5 mol/L) per milliliter of serum to lower the pH to 5.4, at which the enzyme is stable. Under these conditions, TR-ACP activity is maintained at room temperature for several hours, for up to 1 week if the serum is refrigerated, and for 4 months if stored at −20°C. Hemolyzed serum specimens are contaminated with the erythrocyte tartrate-resistant isoenzyme and should be rejected.

REVIEW QUESTIONS

1. The enzyme that demonstrates highest serum activity in Paget disease (osteitis deformans) and is also elevated in hepatobiliary disease is:
   a. alkaline phosphatase (ALP).
   b. creatine kinase (CK).
   c. amylase (AMY).
   d. α-glutamyltransferase (GGT).
2. Activity of which of the following isoenzymes of lactate dehydrogenase (LDH) is highest in the serum of individuals with seminoma of the testis?
   a. LDH-1
   b. LDH-2
   c. LDH-4
   d. LDH-5
3. The serum enzyme that demonstrates an increase in activity 4 to 8 hours after an attack of acute pancreatitis, peaks at 24 hours, and then returns to normal within a week is:
   a. AMY.
   b. lipase (LPS).
   c. ALP.
   d. serum cholinesterase (CHE).
4. Which of the following enzymes catalyzes the reaction of glutamate and pyruvate to form 2-oxoglutarate and an amino acid?
   a. ALP
   b. Aspartate aminotransferase (AST)
   c. Alanine aminotransferase (ALT)
   d. CK
5. In the laboratory measurement of ALP, a chromogenic assay forms the basis of almost all current methods used for ALP analysis. The substrate in this assay is:
   a. acid phosphatase.
   b. 4-nitrophenyl phosphate.
   c. *p*-nitroaniline.
   d. succinyldithiocholine.
6. Measurement of decreased activity of which of the following enzymes is used to determine possible insecticide poisoning?
   a. CK
   b. ALP
   c. ALT
   d. CHE

7. Children have higher ALP activity than healthy adults because:
   a. ALP leaks from osteoblasts during normal bone growth.
   b. developing hepatocytes produce excess ALP during normal growth.
   c. striated muscle contains the greatest activity of ALP, and children are more energetic than adults with concomitant release of excess ALP from muscle.
   d. the presence of ALP in the pancreas is elevated during childhood.
8. The isoenzyme of amylase that is synthesized by acinar cells of the pancreas and that remains elevated in most individuals for at least 1 week after the onset of pancreatitis is:
   a. S-AMY.
   b. macro-AMY.
   c. P-AMY.
   d. S-AMY + P-AMY.
9. Which of the following enzymes demonstrates an increase in activity with progressive Duchenne muscular dystrophy, followed by a decrease as muscle mass decreases?
   a. ALP
   b. ALT
   c. CHE
   d. CK
10. The oxidoreductase that is increased significantly during megaloblastic anemia is:
   a. LDH.
   b. CK.
   c. ALP.
   d. acid phosphatase.

SUGGESTED READINGS

Braga, F., Frusciante, E., Infusino, I., Aloisio, E., Guerra, E., Ceriotti, F., et al. (2017). Evaluation of the trueness of serum alkaline phosphatase measurement in a group of Italian laboratories. *Clinical Chemistry and Laboratory Medicine*, *55*, e47–e50.

Ceriotti, F., Panteghini, M., Guerra, E., Leoncini, R., & Cevenini, G. (2017). Traceable reference intervals for alkaline phosphatase in serum of pediatrics. *Biochimica Clinica, 41*, 166–174.

Dufour, D. R., Lott, J. A., Nolte, F. S., Gretch, D. R., Koff, R. S., & Seeff, L. B. (2000). Diagnosis and monitoring of hepatic injury. II. Recommendations for use of laboratory tests in screening, diagnosis, and monitoring. *Clinical Chemistry, 46,* 2050–2068.

Huijgen, H. J., Sanders, G. T., Koster, R. W., Vreeken, J., & Bossuyt, P. M. (1997). The clinical value of lactate dehydrogenase in serum: a quantitative review. *European Journal of Clinical Chemistry and Clinical Biochemistry*, *35*, 569–575.

Infusino, I., Frusciante, E., Braga, F., & Panteghini, M. (2017). Progress and impact of enzyme measurement standardization. *Clinical Chemistry and Laboratory Medicine, 55,* 334–340.

Kim, W. R., Flamm, S. L., Di Bisceglie, A. M., & Bodenheimer, H. C. (2008). Serum activity of alanine aminotransferase (ALT) as an indicator of health and disease. *Hepatology, 47,* 1363–1370.

Kristensen, S. R. (1994). Mechanisms of cell damage and enzyme release. *Danish Medical Bulletin, 41,* 423–433.

Kunutsor, S. K., Apekey, T. A., & Kahn, H. (2014). Liver enzymes and risk of cardiovascular disease in the general population: a meta-analysis of prospective cohort studies. *Atherosclerosis, 236,* 7–17.

McQueen, M. J. (1995). Clinical and analytical considerations in the utilization of cholinesterase measurements. *Clinica Chimica Acta, 237,* 91–105.

Morandi, L., Angelini, C., Prelle, A., Pini, A., Grassi, B., Bernardi, G., et al. (2006). High plasma creatine kinase: review of the literature and proposal for a diagnostic algorithm. *Journal of The Neurological Sciences*, *27*, 303–311.

Moss, D. W. (1992). Perspectives in alkaline phosphatase research. *Clinical Chemistry*, *38*, 2486–2492.

Moss, D. W., Raymond, F. D., & Wile, D. B. (1995). Clinical and biological aspects of acid phosphatase. *CRC Critical Reviews Clinical and Laboratory Science*, *32*, 431–467.

Pagani, F., & Panteghini, M. (2001). 5′-Nucleotidase in the detection of increased activity of the liver form of alkaline phosphatase in serum. *Clinical Chemistry*, *47*, 2046–2048.

Panteghini, M., Adeli, K., Ceriotti, F., Sandberg, S., & Horvath, A. R. (2017). American liver guidelines and cutoffs for "normal" ALT: a potential for overdiagnosis. *Clinical Chemistry*, *63*, 1196–1198.

Panteghini, M., Bonora, R., & Pagani, F. (1988). An alternative approach to the prevention of succinyldicholine-induced apnoea. *Journal of Clinical Chemistry and Clinical Biochemistry*, *26*, 85–90.

Panteghini, M., Ceriotti, F., Pagani, F., Secchiero, S., Zaninotto, M., & Franzini, C., & for the Italian Society of Clinical Biochemistry and Clinical Molecular Biology (SIBioC) Working Group on Enzymes. (2002). Recommendations for the routine use of pancreatic amylase measurement instead of total amylase for the diagnosis and monitoring of pancreatic pathology. *Clinical Chemistry and Laboratory Medicine*, *40*, 97–100.

Phillip, V., Steiner, J. M., & Algül, H. (2014). Early phase of acute pancreatitis: assessment and management. *World Journal of Gastrointestinal Pathophysiology, 5*, 158–168.

Pratt, D. S., & Kaplan, M. M. (2000). Evaluation of abnormal liver-enzyme results in asymptomatic patients. *The New England Journal of Medicine*, *342*, 1266–1271.

Tenner, S., Baillie, J., DeWitt, J., & Vege, S. S. (2013). American College of Gastroenterology guideline: management of acute pancreatitis. *The American Journal of Gastroenterology*, *108*, 1400–1415.

Whitfield, J. B. (2001). Gamma glutamyl transferase. *CRC Critical Reviews in Clinical Laboratory Sciences*, *38*, 263–355.

Tumor Markers

Catharine Sturgeon

OBJECTIVES

1. List the requirements of the "ideal" tumor marker.
2. Describe how to validate a "new" tumor marker.
3. Describe the major applications of tumor markers, providing specific examples as recommended in clinical guidelines.
4. Outline requirements of an optimal tumor marker service from test request to laboratory report.
5. Describe pitfalls and caveats in the use and interpretation of tumor markers.

KEYWORDS AND DEFINITIONS

Cancer Heterogeneous but related diseases characterized by uncontrolled division of cells.

Diagnostic accuracy The ability of a test to identify correctly patients who have malignancy.

Heterophilic antibodies Antibodies that may react with immunoassay reagents, yielding inappropriately high or low results.

High-dose "hooking" Interference leading to inappropriately low results for specimens of high analyte concentrations.

Prognosis Prediction of the likely outcome of disease with respect to risk of relapse or progression.

Screening Detection of unsuspected disease in asymptomatic subjects.

Serial monitoring Posttreatment tumor marker measurement enabling early detection of disease progression.

Therapy prediction Identification of patients likely to benefit from a particular treatment.

Tumor marker A substance present in abnormally high concentrations in body fluids or tissue from cancer patients.

Tumor markers are substances that may be present in abnormally high concentrations in body fluids or tissue from patients with cancer. Tumor markers can aid cancer management in a number of ways, including screening, diagnosis, prognostic assessment, **therapy prediction**, and/or posttreatment monitoring (Table 20.1). Reflecting the heterogeneous nature of cancer, tumor markers encompass a variety of molecular species. Many tumor markers are produced by a variety of different tumors, and few tumor markers are organ-specific or specific for a single type of malignancy.

Tumor markers range from simple molecules (e.g., catecholamines) through relatively well-characterized proteins (e.g., hormones) to much more heterogeneous glycoproteins and mucins (e.g., cancer antigen 125 [CA 125]), which may be defined by the antibodies used to measure them. Several important tumor markers (e.g., α-fetoprotein [AFP], carcinoembryonic antigen [CEA], and human chorionic gonadotropin [hCG]) are oncofetal antigens, which are present in the fetus in normal pregnancy but are expressed at high concentrations in tissue or body fluids of adults with some cancers. Tumor markers are present in cells, tissues, or body fluids and can be measured either qualitatively or quantitatively using chemical, immunologic, or molecular biologic methods. Although their structures and properties vary widely, the broad principles underpinning their evidence-based clinical use are common to all tumor markers.

This chapter provides a brief overview of cancer and tumor marker development followed by a comprehensive review of these principles. Clinically relevant tumor markers for a number of malignancies are then discussed in detail, considering current practice and future requirements.

CANCER—AN OVERVIEW

Cancer is the name given to a collection of heterogeneous but related diseases that can start in most parts of the body and are characterized by uncontrolled division of cells, which ultimately may spread into surrounding tissues. Causative factors include exposure to carcinogens, which may be physical (e.g., radiation), chemical (e.g., polycyclic hydrocarbons), or biological (e.g., viral). Excess weight, physical inactivity, and poor nutrition may also contribute to the development of some cancers.

TABLE 20.1 Current Clinical Applications of Tumor Markers

Clinical Application	Requirements	Examples
Screening for cancer	An acceptable test for a disease that poses an important health problem, for which the natural history is well understood, the test identifies treatable early stage and early intervention with effective treatment improves outcome. High clinical sensitivity (few false-negatives) and specificity (few false-positives) are essential prerequisites as the population screened is asymptomatic.	*Population screening*: Fecal occult blood testing for colorectal cancer. *Screening of high-risk groups:* Serum hCG testing for choriocarcinoma in women who have had a molar pregnancy.
Diagnosing cancer	High clinical sensitivity and specificity as for screening. Most serum tumor markers are not cancer specific and/or organ specific and are raised in different cancers and/or in nonmalignant disease, severely limiting their use in diagnosis.	*As a diagnostic aid in high-risk groups:* Serum AFP testing as an adjunct to ultrasound for hepatocellular carcinoma (HCC) in patients with cirrhosis who are at high risk of developing HCC. *Differential diagnosis:* CA 125 contributes (with menopausal status and ultrasound findings) to calculation of the risk of malignancy index, which is used to differentiate patients with benign and malignant pelvic masses.
Assessing prognosis	A test that can provide a probability estimate of outcome (risk of relapse or disease progression) and/or differentiate indolent from aggressive disease for a heterogeneous population of patients, thereby influencing treatment decisions.	Serum AFP, hCG, and lactate dehydrogenase measurements are mandatory for determining prognosis and selection of chemotherapy in patients with metastatic nonseminomatous germ cell tumors. Microsatellite instability (MSI) occurs when germline alleles of microsatellites (short stretches of DNA that are repeated at multiple locations throughout the genome) lose or gain a repeat unit due to failure of mismatch repair. Presence of MSI is associated with good prognosis in colorectal cancer patients, especially in stage II or III disease, and their presence, associated with other established prognostic factors, may obviate the need for adjuvant chemotherapy in patients with stage II disease. Gene expression profiles such as the Oncotype Dx test, which measures expression of 21 genes in breast tumor tissue, enable calculation of a recurrence score that predicts the risk of distant disease at 10 years in a specific subset of breast cancer patients.
Prediction of treatment response	A test that can identify whether or not a potential treatment is likely to be effective and of benefit to the patient.	Measurement of estrogen receptors is mandatory for all newly diagnosed invasive breast cancer to predict response to treatment with antiestrogen therapy (e.g., aromatase inhibitors and tamoxifen). The additional measurement of progesterone receptors can improve the accuracy of prediction. Measurement of human epidermal growth factor receptor 2 (HER2) is essential to identify patients with HER2-positive breast cancer who are likely to benefit from treatment with anti-HER2 monoclonal antibody therapy.
Monitoring response during and/or shortly after treatment	A test that assesses whether the tumor is responding to treatment, enabling withdrawal and/or change of ineffective treatment.	In patients receiving treatment for ovarian cancer a reduction of at least 50% in CA 125 concentrations from a pretreatment sample has been defined as a response provided it is confirmed and maintained for at least 28 days and the pretreatment CA 125 concentration was greater than twice the upper limit of the reference interval and measured within 2 weeks of the start of chemotherapy. Biologic response to systemic chemotherapy, as assessed by serial measurement of serum CEA and CA 19-9 in patients receiving chemotherapy for colorectal cancer with liver metastases, agrees with response as assessed by radiology in >94% and 91% of patients, respectively, suggesting that this could decrease the need for imaging studies, such as CT, which require exposure of patients to significant doses of radiation.

TABLE 20.1 **Current Clinical Applications of Tumor Markers—cont'd**

Clinical Application	Requirements	Examples
Monitoring disease post treatment to detect progression	A test that reliably indicates earlier than clinical symptoms whether disease is progressing in patients with no evidence of disease post therapy and/or in patients with detectable disease. Whether this information is helpful crucially depends on whether alternative therapies are available.	Inclusion of CEA measurements in intensive follow-up strategies for patients with nonmetastatic colorectal cancer following curative surgery have been shown to improve overall survival. Such follow-up is also associated with earlier detection of recurrence.

AFP, α-Fetoprotein; *CA 125,* cancer antigen 125; *CA 19-9,* cancer antigen 19-9; *CEA,* carcinoembryonic antigen; *CT,* computerized tomography; *hCG,* Human chorionic gonadotropin.

Genetic predisposition to some cancers is increasingly identifiable due to advances in molecular techniques. Simple genetic analysis of mutations within cancerous cells may be too simplistic to explain a complex process, but increased understanding of such interactions is already leading to more finely targeted therapies.

Cancer remains a leading cause of death in the United States, second only to heart disease, although death rates have decreased significantly over the past three decades. Early detection and more effective treatment combined with prevention (e.g., decreased smoking, improved diet) may further reduce the cancer mortality rate. However, identification of smaller more indolent cancers with newer techniques and screening programs may also contribute to the perceived increase in survival.

Reflecting advances in treatment over recent decades, many patients survive for much longer than previously. Although surgical removal of a primary tumor is the only curative therapy for almost all cancers, new chemotherapeutic and biologic therapies that prevent progression are increasingly available. Some relatively indolent cancers are now almost regarded as chronic diseases like diabetes, to be controlled rather than cured. Breast and prostate cancers are among the best examples. However, for these and other cancers, distinguishing cancers that are likely to progress rapidly from those that are slow growing or indolent is critical and remains a challenge.

Early detection of malignancy optimizes any opportunities for curative surgery for some in situ cancers. Unfortunately, most cancers do not produce symptoms until tumors are too large to be removed surgically or until cancerous cells have spread to other tissue either by invading local lymph nodes or by distant spread (metastasis) to other organs. Systemic treatments (chemotherapy, endocrine therapy, or immunotherapy) are then usually the only options but are not curative. Any residual viable cancerous cells remaining after treatment may proliferate, develop resistance to further therapy, and ultimately cause the death of the patient.

HISTORICAL BACKGROUND

The introduction of new tumor markers has mirrored developments in analytical techniques and more recently the availability of new therapies, many of which are effective only in subsets of cancer patients.

The first tumor marker to be used clinically was Bence Jones protein, identified in 1847, and it is still the basis of many diagnoses of multiple myeloma. In the first half of the 20th century the presence or absence of other hormones, enzymes and isoenzymes, and blood group antigens was recognized to be associated with malignancy, but it was not until development of the technique of radioimmunoassay (RIA) in the 1960s that these observations could be translated into routine clinical practice (refer to Chapter 15). Monoclonal antibody technology subsequently enabled the development of two-site immunometric assays of complex cancer-associated mucins identified primarily by their reactivity with given monoclonal antibody pairs (e.g., the cancer-associated antigens CA 125 and CA 15-3). As automated methods became available, many more laboratories added tumor markers to their test repertoires.

Advances in molecular genetics, including the study of oncogenes, suppressor genes, and genes involved in DNA repair, have led to rapid understanding and use of tumor markers at both the molecular and cellular levels, particularly in high-risk individuals. Estrogen, progesterone, and epidermal growth factor receptor (EGFR) measurements are now routinely performed for breast cancer patients to inform treatment decisions, and measurement of the breast cancer susceptibility genes *BRCA1* and *BRCA2* provides additional prognostic information.

Mass spectrometry is also increasingly used both as a discovery and a diagnostic tool, albeit with limited success thus far in translating new tumor marker tests into clinical practice. These developments require sophisticated bioinformatics support, which encourages multiparametric analysis for cancer diagnosis, prognosis, and therapy prediction.

GENERAL PRINCIPLES GUIDING THE USE OF ALL TUMOR MARKERS

Tumor markers can help make or confirm a cancer diagnosis, monitor treatment effectiveness and the course of disease, estimate prognosis, and/or predict whether a specific therapy is likely to be successful (Table 20.1). Their measurement should permit more efficient application of therapies, by ensuring that these are applied only to those patients most likely to benefit and reducing exposure to unnecessary toxicity for those patients unlikely to benefit. Tumor markers

TABLE 20.2 Properties and Applications of Commonly Requested Serum Tumor Markers

Tumor Marker	Biochemical Properties	Molecular Weight	Main Clinical Applications
Alkaline phosphatase	Phosphohydrolase	Variable	Raised activities associated with presence of liver and/or bone metastases
α-Fetoprotein	Glycoprotein, 4% carbohydrate; considerable homology with albumin	~70 kDa	Diagnosis and monitoring of primary hepatocellular carcinoma, hepatoblastoma, and GCTs. Prognosis of GCT
Cancer antigen 125	Mucin identified by monoclonal antibodies OC125 and M11. Developed from serous cystadenocarcinoma cell line	~200 kDa	Monitoring ovarian carcinoma. Measurement required for determination of the risk of malignancy index for ovarian carcinoma
CEA	Family of glycoproteins, 45%–60% carbohydrate	~180 kDa	Monitoring colorectal adenocarcinomas
ConfirmMDx	A multiplex epigenetic assay	Not applicable	Reducing unnecessary repeat prostatic biopsies in men who have had at least one negative biopsy
Human chorionic gonadotrophin	Glycoprotein hormone consisting of two noncovalently bound subunits (α and β). α-Subunit similar to LH, FSH and TSH. β-Subunit considerable homology with LH	~36 kDa	Diagnosis, prognosis, and monitoring of GCTs and gestational trophoblastic neoplasia
Lactate dehydrogenase	Enzyme of the glycolytic pathway	Variable	Diagnosis, prognosis, and monitoring of GCTs. Used to monitor a wide range of malignancies including hematologic malignancies
Paraproteins	Monoclonal immunoglobulins	Variable	*See Chapter 18*
Prolactin	Pituitary hormone	~22 kDa but high molecular forms also exist	*See Chapter 40*
PSA	Glycoprotein, member of the kallikrein family with serine protease activity. Circulates as free enzyme or complexed to α_1-antichymotrypsin (measurable) or α_2-macroglobulin (not detected by most immunoassays)	~30 kDa (free enzyme)	Diagnosis, risk assessment and monitoring of prostate carcinoma. Lower concentrations of free PSA relative to complexed PSA (i.e., free:total ratio) found in prostatic cancer as compared with benign prostatic hypertrophy

CEA; Carcinoembryonic antigen; *GCT,* germ cell tumor; *PSA,* prostate-specific antigen.

should be measured only after considering whether the result is likely to provide information that may improve the outcome for the individual patient or as part of a clinical trial.

An evidence-based approach to clinical decision making is therefore essential, and careful consideration must be given to whether an individual tumor marker can fulfill requirements in selected clinical circumstances. Guidelines published by the National Comprehensive Cancer Network (NCCN) in the United States include patient pathways that clearly indicate which tumor markers should be measured and when. Complementing these clinically oriented guidelines, those published by the National Academy of Clinical Biochemistry (NACB) focus on the use of tumor markers, considering in detail requirements for their appropriate use in all phases of both the patient pathway and laboratory provision, helpfully comparing recommendations made with those of other groups. These are summarized later in this chapter.

Well-established and commonly requested serum tumor markers whose measurements are generally available in most large clinical laboratories are described in Table 20.2. Some less commonly requested serum tumor markers with an established clinical role are shown in Table 20.3. Table 20.4 lists some tissue markers essential to informed selection of endocrine and immunotherapy in individual patients.

The broad principles guiding the validation of any new tumor marker and its subsequent introduction into routine clinical practice are essentially similar and are considered in detail in the following sections. Early critical aspects include rigorous validation of analytical performance and clinical utility and demonstration of clinical value. Objective assessment of the potential clinical value of a proposed new tumor marker together with an estimate of the magnitude of its benefit, including effect on patient outcome, is now a prerequisite for introducing a new test into routine practice. Three key issues in tumor marker evaluation are clinical utility, magnitude of effect, and reliability. Only when these have been established and regulatory requirements fulfilled should the tumor marker be adopted into clinical practice. It is then essential to ensure that the new test is requested appropriately, that those requesting the test are aware of any preanalytical requirements, that these are met in routine practice, that analytical performance is reliable as confirmed by rigorous

TABLE 20.3 Properties and Applications of Less Commonly Requested Serum Tumor Markers

Tumor Marker	Biochemical Properties	Molecular Weight	Main Clinical Applications
Calcitonin	32–Amino acid peptide	~3.5 kDa	Monitoring medullary carcinoma of the thyroid
Cancer antigen 15.3 (CA 15.3), BR 27.29	Mucin (MUC-1 glycoprotein peptide) identified by monoclonal antibodies	>250 kDa	Monitoring breast cancer
Cancer antigen 19.9 (CA 19.9)	Glycolipid carrying the Lewis[a] blood group determinant	~1000 kDa	Monitoring pancreatic carcinoma following curative resection
Catecholamines	Biogenic amines	~0.2 kDa	*See Chapter 26*
Chromogranin A	Member of the granin family of acid secretory glycoproteins	~49 kDa	Monitoring neuroendocrine tumors
CYFRA 21-1	Fragments of cytokeratin 19	~30 kDa	Monitoring lung carcinoma
Gut hormones	Small peptide hormones including vasoactive intestinal peptide, pancreatic polypeptide, somatostatin, gastrin		*See Chapter 38*
Human epididymis protein 4 (HE4)	Product of the WFDC2 (HE4) gene that is overexpressed in patients with ovarian carcinoma	~25 kDa	Monitoring ovarian cancer. Potentially an aid in diagnosis as part of the risk of malignancy algorithm. Under evaluation
Inhibin A (α-β_A) Inhibin B (α-β_B)	Heterodimeric glycoproteins composed of an α- and a β-subunit. There are several forms of the β-subunit, as indicated	32 kDa	Monitoring of ovarian granulosa cell tumors and testicular Sertoli and Leydig tumors
β_2-microglobulin	Component polypeptide chain of the HLA antigen complex	~12 kDa	Prognostic assessment of multiple myeloma
Neuron-specific enolase	Dimer of the enzyme enolase	~87 kDa	Monitoring small cell lung carcinoma, neuroblastoma, and neuroendocrine tumors
Placental alkaline phosphatase	Heat-stable isoenzyme of alkaline phosphatase	~86 kDa	Monitoring of germ cell tumors (seminomas)
Progastrin-releasing peptide	More stable precursor of gastrin-releasing peptide	~16 kDa	Monitoring small cell lung carcinoma
Prostate cancer gene 3 protein (PCA3)	Protein product of PCA3 gene in urine or prostate cancer patients	~86 kDa	Potentially an aid to diagnosis of prostate cancer particularly in biopsy-negative men. Under evaluation
Prostate marker algorithms	Phi	Not applicable	An algorithm combining serum measurements of total PSA, free PSA, and proPSA in serum
	4K	Not applicable	An algorithm combining serum measurements of free and total PSA, human kallikrein 2, and intact PSA
Squamous cell carcinoma antigen	Serine protease inhibitor isolated from glycoprotein subfraction of tumor antigen TA-4	48 kDa	Monitoring squamous cell carcinomas (e.g., cervix)
S-100 proteins	Family of proteins characterized by two calcium-binding sites that have helix-loop-helix confirmations	~20 kDa	Melanoma; clinical utility in melanoma to be established
Thyroglobulin	Glycoprotein dimer of two identical subunits	670 kDa	Monitoring differentiated thyroid cancer
Tissue polypeptide antigen	Fragments of cytokeratin 8, 18, and 19	~22 kDa	Monitoring bladder and lung carcinoma

PSA, Prostate-specific antigen.

TABLE 20.4 Tissue Markers Whose Measurement Informs Selection of Endocrine and Immunotherapy in Individual Patients

Tumor marker	Biochemical Properties	Molecular Weight	Main Clinical Applications
Anaplastic lymphoma receptor tyrosine kinase *(ALK)*	Enzyme encoded by the *ALK* gene	~176 kDa	Predicting response to the tyrosine kinase inhibitor) crizotinib in non–small cell lung cancer
BRAF	Proto-oncogene that encodes the serine/threonine protein kinase B-Raf	~87 kDa	Predicting response to inhibitors of *BRAF*-mutated protein including vemurafenib and dabrafenib in melanoma
Estrogen receptor	Nuclear transcription factor	~5 kDa	Predicting response to endocrine therapy in breast cancer
Human epidermal growth factor receptor-2 (HER2 or c-erb2)	Transmembrane glycoprotein encoded from HER2/neu oncogene	~185 kDa	Predicting response to trastuzumab (Herceptin) in breast cancer Predicting response to tyrosine kinase inhibitors in non–small cell lung cancer
KRAS mutation	Guanosine nucleotide–binding protein	~23 kDa	Predicting response to tyrosine kinase inhibitors in non–small cell lung cancer and reducing need for testing for *HER2* and *ALK* alterations
Progesterone receptor	Nuclear transcription factor	A form ~4 kDa B form: ~120 kDa	Predicting response to endocrine therapy in breast cancer

internal quality control (IQC) and proficiency testing (PT), and that postanalytical reporting of results to clinical users is both informed and informative.

Tumor Marker Validation

Assuming there is preliminary evidence that the marker has clinical potential, requirements for validation prior to introduction into routine use include:

1. *Protocols for sample collection and processing.* Availability of clinical specimens (e.g., serum, plasma, tissue) that have been collected, processed, and stored following validated standard operating procedures (SOP). The panel should include samples from patients with a variety of malignant and nonmalignant conditions. Specimens from biobanks may be appropriate provided their provenance is well documented.
2. *Analytical/technical validation.* Development and validation of each specific assay to be used in the clinic, having regard to the requirements described as follows, including biologic variation of the marker where feasible.
3. *Clinical validation.* Definition of the specific context in which the tumor marker is to be used so as to identify the type of specimens to be used in the clinical validation step, the informative statistics, and the acceptability criteria. Clinical validation should be reported in terms of diagnostic accuracy for tests intended for use in diagnosis. The risk of pitfalls including study bias, overfitting of data, and other statistical problems can be reduced by strict adherence to a written protocol and rigorous reporting of all work.
4. *Demonstration of clinical value.* Strong evidence that the test will benefit the patient (e.g., by increasing overall survival or leading to fewer invasive procedures or hospital visits). Ideally this should include high-level evidence from a prospective or retrospective randomized clinical trial or systematic review of the literature.
5. *Achieving regulatory approval.* Requirements vary but in the United States involve either clearance approval by the US Food and Drug Administration (FDA) and/or evaluation in a Clinical Laboratory Improvement Amendments (CLIA)-certified laboratory or by a laboratory-developed test pathway.
6. *Postmarket evaluation.* Analytical performance of the test in the routine clinical setting should be subject to continued evaluation, through rigorous implementation of IQC and PT procedures (refer to Chapter 7). Ongoing clinical audit to ensure that the marker retains its value in the clinical setting is also essential. When a newly introduced tumor marker replaces an earlier one, it is important to ensure that the first test is discontinued.

"Ideal" Tumor Marker

An "ideal" tumor marker should be the following:

1. Detectable only in a given malignancy and absent in the healthy population or in nonmalignant conditions (i.e., have clinical specificity and clinical sensitivity approaching 100%)
2. Present in a convenient biologic matrix (e.g., serum, urine, tumor tissue)
3. Present at concentrations proportional to tumor burden
4. Conveniently measured by a readily available, simple, reproducible, and inexpensive procedure
5. Beneficial to clinical care with a measurable effect on patient outcome

No such tumor marker has yet been identified. However, measurement of hCG approaches ideal in the screening and monitoring of women who have had a molar pregnancy and are at risk of developing gestational trophoblastic neoplasia (GTN) (choriocarcinoma). Other tumor markers can also contribute usefully to patient management provided they are used intelligently and with regard to requirements of the clinical application (Tables 20.2 to 20.4) as outlined next.

Clinical Applications of Tumor Markers

Screening for Cancer

Screening for disease differs fundamentally from diagnosis or "case finding" because the aim of screening is to detect unsuspected disease in asymptomatic subjects (Table 20.1). The World Health Organization (WHO) criteria for screening programs originally proposed by Wilson and Jungner in 1968 remain valid. High sensitivity and high specificity of the test for the target disease is essential, with the success of a screening program also dependent on the prevalence of the condition in the population being screened.

Diagnostic accuracy is established by studying test performance at different decision (cutoff) concentrations in appropriately selected diseased and disease-free populations. The relationship between these parameters is most conveniently depicted graphically as a receiver operating characteristic (ROC) plot. Changing the decision concentration selected alters both sensitivity and specificity but cannot improve both concurrently. These parameters define the characteristics of the test and, together with the prevalence of the disease in the population studied, yield its positive predictive value. The low prevalence of the majority of cancers, even in age groups most at risk, together with the low sensitivity and specificity of most tumor markers, generally precludes their use in screening. However, screening for early colorectal cancer using a fecal immunochemical test (FIT) for hemoglobin is routine in some countries, the efficacy of screening patients at high risk of hepatocellular carcinoma (HCC) and ovarian cancer is being actively investigated, and there is continuing debate about whether to screen for prostate cancer and some genetic conditions.

Diagnosis of Cancer

The lack of sensitivity and specificity that limits the application of tumor markers in screening also applies to diagnosis and case finding (Table 20.1). Tumor markers are not helpful for diagnosis in patients with nonspecific symptoms and cannot replace biopsy for establishing a cancer diagnosis. However, in carefully selected undiagnosed patients, raised concentrations of tumor markers may contribute to diagnosis (e.g., when the patient is unable or unwilling to undergo more invasive testing such as colonoscopy). In general, the higher the serum tumor marker value the greater the likelihood of malignancy, but results within the reference interval never necessarily exclude malignancy. Requests for panels of tumor marker measurements should be actively discouraged, as should requests for prostate-specific antigen (PSA) in women or CA 125 in men.

Assessing Prognosis

Prognostic markers predict the likely outcome of disease with respect to risk of relapse or disease progression (Table 20.1). They are generally most useful at the time of initial diagnosis, when a marker of good prognosis is suggestive of prolonged survival and/or the possibility of cure. A marker of poor prognosis indicates an increased probability of early recurrence. Prognostic markers can help identify patients with aggressive tumors that require further treatment (e.g., with systemic adjuvant chemotherapy following surgery) while minimizing the risk of overtreatment of patients with indolent disease that is unlikely to progress. Few prognostic markers are used in routine clinical practice as they provide a probability estimate of outcome for a heterogeneous population rather than for an individual patient. Most of those used are measured in tissue. Exceptionally, measurement of AFP, hCG, and lactate dehydrogenase (LDH) in serum or plasma following primary surgery in patients with nonseminomatous germ cell tumors (GCTs) is mandatory. These well-validated markers provide strong independent prognostic information complementing that of conventional prognostic factors.

Prediction of Treatment Response

Even cancers of the same histologic type may vary widely in response to a particular treatment, and for most cancers only a subset of patients will respond (Table 20.1). Predictive markers that can identify such patients are therefore critical in treatment planning, so that treatment can be offered to patients who will benefit and enabling alternative treatment to be offered to those who will not, also sparing them unnecessary side effects. Predictive markers, most of which are measured in tissue rather than serum, are increasingly both clinically and economically essential to identify patients who may benefit from new and costly molecularly targeted treatments. Several important predictive markers for breast cancer are already well established in clinical use, with further promising markers now available for breast cancer, colorectal cancer, glioma, melanoma, and non–small cell lung cancer (NSCLC).

Monitoring Response During and/or Shortly After Treatment

The main application of most serum tumor markers is in monitoring the course of disease. In general, it is helpful to measure the relevant tumor marker pretreatment to provide a baseline for subsequent interpretation, and for some cancers measurements made during and immediately after treatment are also desirable. In general, a decrease in marker concentrations to within "normal" limits following treatment is a favorable sign. In some cases, subnormal or absent posttreatment values should be achieved (e.g., for PSA post prostatectomy or thyroglobulin following thyroid ablation). Measurement of AFP and hCG is mandatory for GCTs prior to surgical excision, to enable calculation of the half-life of tumor marker decline following chemotherapy.

Monitoring Disease Post Treatment to Detect Progression or Recurrence

Serial monitoring of tumor marker concentrations post treatment provides early indication of disease progression, often months before there are clinical signs and symptoms. Whether this information is of benefit depends crucially on whether the increase will prompt clinical action (e.g., early ultrasound, computed tomography (CT), or magnetic resonance imaging [MRI] scanning), whether alternative treatment

is available and/or whether that treatment can be implemented without confirmatory scan evidence. If not, and if the patient is not enrolled in a clinical trial, the patient should be made aware of the potential implications of having the test done, and whether tumor marker monitoring is likely to be of benefit should be seriously considered and discussed.

QUALITY REQUIREMENTS FOR TUMOR MARKER USE IN THE CLINIC

The clinical laboratory should provide readily available information to those requesting tumor markers, to encourage selection of the correct test or tests (and, importantly, to discourage inappropriate requesting) and to ensure that specimen timing is appropriate and other preanalytical requirements are met. Ensuring that results are analytically correct and that accurate and informative reports are returned to the requesting clinician are also laboratory responsibilities. Laboratory-oriented guidelines for tumor markers that outline quality requirements for provision of a high-quality tumor marker service have been developed by the NACB and form the basis of the following discussion.

Pre-preanalytical Requirements

Although numerous serum tumor markers can be measured (Tables 20.2 to 20.4), relatively few are commonly requested from the clinical laboratory. Table 20.5 summarizes the current recommendations of the NACB for their appropriate clinical use, and typical clinical presentations that might prompt a request are described in Table 20.6.

Requests for tumor markers may be for diagnosed cancer patients who have already been referred to specialist centers and are being or have been treated, for patients who have been referred to hospital for investigation of suspected malignancy and further investigation, and for patients who have presented to their family doctor with symptoms suspicious for malignancy. Requests in the first category (with a baseline measurement prior to therapy always desirable) are most likely to be appropriate.

Nonspecialist users should always consider carefully whether a tumor marker result is likely to be helpful before making a request. Requestors should be aware of the lack of sensitivity of most tumor markers, particularly for early stage disease, their lack of specificity, and the numerous nonmalignant conditions in which they may be increased (Table 20.7). Nonspecialist users should also be aware that increased tumor marker concentrations do not necessarily indicate malignancy. Attempting to identify the reason for an increased tumor marker that should never have been requested and may not be associated with malignancy can be expensive and time consuming, and psychologically stressful for the patient. Conversely, requestors should be reminded that a result within the reference interval never necessarily excludes malignancy or progressive disease. Clinical biochemists should also think carefully before requesting additional tests that might lead to a diagnosis of malignancy and should seek the agreement of the clinician caring for the patient before doing so.

Unfocused requests such as "tumor marker screen" or "malignancy?" should be actively discouraged and met with offers for educational support. Advice about appropriate test selection should be made readily available at the time of requesting, ideally electronically.

Preanalytical Requirements

Specimen Timing

Timing of specimens within the day for tumor marker measurement is not usually critical because there is little evidence of diurnal variation for most markers. A pretreatment specimen is always helpful when interpreting subsequent results. Specimens should be taken before some procedures that may transiently release tumor markers into the circulation (e.g., CA 125 following abdominal surgery, CEA following colonoscopy, and PSA after prostatic biopsy or digital rectal examination [DRE]). Awareness of benign conditions that may cause transient increases in tumor marker concentrations (Table 20.7) and avoiding inappropriate timing reduces the risk of misinterpreting spuriously raised results that may cause undue distress to the patient and also decrease confidence in laboratory testing.

Specimen Handling

Serum and/or plasma specimens are appropriate for most tumor marker measurements, although gel tubes may not be suitable for some assays. Requirements should be checked in the product information supplied. Tumor markers are generally stable, but serum or plasma should be separated from the clot as soon as possible and then stored at 4°C (short term) or −30°C or below. For longer-term storage, specimens should be stored at −70°C or below. Standardized conditions of specimen collection and fixation are crucial for immunohistochemical analyses.

Analytical Requirements

Assay Validation

Prior to introduction into routine clinical practice, both immunoassays and immunohistochemical methods must be validated, as described previously, by defined and well-characterized protocols that meet regulatory guidelines (e.g., FDA approval in the United States and CE marking in Europe). Individual laboratories should verify analytical performance prior to introduction into routine use. Internationally recognized guidelines for the performance of immunohistochemical tests should be adopted if available and methods described in detail if appropriate high-quality reference materials are not available.

Internal Quality Control

Robust procedures for IQC should be established. Within-run variability less than 5% and between-run variability less than 10% should be readily achievable for automated tumor marker methods. Some newer techniques may perform significantly better than this, although manual and/or research assays may be less precise. Appropriate action should be taken immediately if an assay run fails to meet objective criteria for assay acceptance so that no potentially erroneous results are reported. Criteria for acceptance should be predefined and preferably based on logical criteria such as those of Westgard.

TABLE 20.5 Most Frequently Requested Serum Tumor Markers and National Academy of Clinical Biochemistry Recommendations for Their Appropriate Clinical Use

		CURRENTLY RECOMMENDED CLINICAL APPLICATIONS				
	Relevant Cancer	Screening or Early Detection	Diagnosis or Case Finding	Prognosis (With Other Factors)	Detecting Recurrence	Monitoring Therapy
AFP	Germ cell/testicular tumor	√[a]	√	√	√	√
	HCC		√[b]	√	√	√[c]
Calcitonin	Medullary thyroid carcinoma		√		√	√
CA 125	Ovarian cancer	√[d]	√[e]	√	√	√[f]
CA 15-3	Breast cancer				√[g]	√[h]
CA 19-9	Pancreatic cancer		√[i]	√	√	√[j]
CEA	Colorectal cancer			√	√[c]	√[c,h]
hCG	Germ cell and testicular cancers; gestational trophoblastic neoplasia[k]		√	√	√	√
Paraproteins (M protein/ Bence Jones protein) *(Also measured in urine)*	Multiple myeloma		√		√	√
Prostate-specific antigen	Prostate cancer		√	√	√	√
Thyroglobulin	Thyroid cancer (follicular or papillary)				√	√

[a]Only for subjects in high-risk groups (e.g., with chronic hepatitis B virus, hepatitis C virus, or cirrhosis) and only in conjunction with ultrasound (impact on mortality unclear).
[b]In conjunction with liver imaging, AFP levels >200 μg/L are regarded as virtually diagnostic of HCC in patients with hypervascular lesions.
[c]Especially for disease that cannot be evaluated by other modalities.
[d]Only for women at high risk of ovarian cancer and only in conjunction with transvaginal ultrasound.
[e]Only for differential diagnosis of pelvic masses, especially in postmenopausal women.
[f]Preliminary results of a randomized trial show no survival benefit from early treatment based on a raised serum CA 125 level alone so this recommendation may be modified to exclude asymptomatic patients.
[g]Post surgery when may provide lead time for early detection of metastasis but clinical value unclear.
[h]Especially in patients with nonevaluable disease (for which CEA is also recommended in carefully selected patients).
[i]In patients in whom pancreatic disease is strongly suspected, CA 19-9 may complement other diagnostic procedures.
[j]Especially after chemotherapy and in combination with imaging.
[k]Use of hCG in screening for gestational trophoblastic neoplasia, a rare malignancy that develops most frequently following a molar pregnancy provides an excellent example of "best practice" in screening. Further information is available from the Trophoblastic Tumour Screening and Treatment Centre, Department of Medical Oncology, Charing Cross Hospital, London W6 8RF, UK (Tel: +44 (0) 20 8846 1409) or at http://www.hmole-chorio.org.uk/.
AFP, α-Fetoprotein; *CA 125,* cancer antigen 125; *CA 15-3,* cancer antigen 15-3; *CA 19-9,* cancer antigen 19-9; *CEA,* carcinoembryonic antigen; *HCC,* hepatocellular carcinoma; *hCG,* human chorionic gonadotrophin.

TABLE 20.6 Typical Clinical Presentations That Might Prompt a Request for the Most Frequently Used Tumor Markers[a]

Tumor Marker	Relevant Cancer	Typical Clinical Presentation	Other Cancers in Which Marker May Be Raised[b]
α-Fetoprotein	Germ cell/testicular tumor	Diffuse testicular swelling; hardness	Gastric; colorectal; biliary; pancreatic; lung
	Hepatocellular carcinoma	Ascites, encephalopathy, jaundice; upper abdominal pain; weight loss; early satiety in high-risk subjects (i.e., hepatitis B– or C–related cirrhosis)	

Continued

TABLE 20.6 Typical Clinical Presentations That Might Prompt a Request for the Most Frequently Used Tumor Markers[a]—cont'd

Tumor Marker	Relevant Cancer	Typical Clinical Presentation	Other Cancers in Which Marker May Be Raised[b]
Cancer antigen 125	Ovarian cancer	Pelvic mass; persistent, continuous or worsening unexplained abdominal or urinary symptoms; bloating	Breast; endometrial; cervix; peritoneal; uterus; lung; pancreas; hepatocellular; non-Hodgkin lymphoma
Cancer antigen 19-9	Pancreatic cancer	Progressive obstructive jaundice with profound weight loss and/or pain in the abdomen or mid-back	Colorectal; gastric; hepatocellular; esophageal; ovary
Carcinoembryonic antigen	Colorectal cancer	Intermittent abdominal pain, nausea, vomiting or bleeding; palpable abdominal mass	Breast; gastric; lung; mesothelioma; esophageal; pancreas
Human chorionic gonadotrophin	Germ cell/testicular tumor Gestational trophoblastic neoplasia	Diffuse testicular swelling, hardness, and pain Symptoms leading to x-ray showing cannon ball secondaries; previous history of hydatidiform mole	Lung cancer
Paraproteins (M protein/ Bence Jones protein) *(Also measured in urine)*	Multiple myeloma	Combination of symptoms including some/ all of the following: anemia; back pain; weakness or fatigue; osteopenia; osteolytic lesions; raised ESR or globulins; spontaneous fractures; recurrent infections	
Prostate-specific antigen	Prostate cancer	Frequency, urgency, nocturia, dysuria; acute retention; back pain, weight loss, anemia	None

[a]Tumor markers are often not helpful in diagnosis (see text).
[b]This list is not comprehensive. The marker may be raised in other cancers.

TABLE 20.7 Nonmalignant Conditions That May Cause Increases (Some Transient) in Serum Tumor Marker Concentrations That May Lead to Incorrect Interpretation[a]

Acute cholangitis	CA 19-9
Acute hepatitis	CA 125, CA 15-3
Acute and/or chronic pancreatitis	CA 125, CA 19-9
Acute urinary retention	CA 125, PSA
Arthritis/osteoarthritis/rheumatoid arthritis	CA 125
Benign prostatic hyperplasia	PSA
Cholestasis	CA 19-9
Chronic liver diseases (e.g., cirrhosis, chronic active hepatitis)	CEA, CA 125, CA 15-3, CA 19-9
Chronic renal failure	CA 125, CA 15-3, CEA, hCG
Colitis	CA 125, CA 15-3, CEA
Congestive heart failure	CA 125
Cystic fibrosis	CA 125
Dermatologic conditions	CA 15-3
Diabetes	CA 125, CA 19-9
Diverticulitis	CA 125, CEA
Endometriosis	CA 125
Heart failure	CA 125
Irritable bowel syndrome	CA 125, CA 19-9, CEA
Jaundice	CEA, CA 19-9
Leiomyoma	CA 125
Liver regeneration	AFP
Menopause	hCG
Menstruation	CA 125
Mesothelioma	CEA
Nonmalignant ascites	CA 125
Ovarian hyperstimulation	CA 125
Pancreatitis	CA 125, CA 19-9
Pericarditis	CA 125

TABLE 20.7 Nonmalignant Conditions That May Cause Increases (Some Transient) in Serum Tumor Marker Concentrations That May Lead to Incorrect Interpretation[a]—cont'd

Peritoneal inflammation	CA 125
Pregnancy	AFP, CA 125, hCG
Prostatitis	PSA
Recurrent ischemic strokes in patients with metastatic cancer	CA 125
Respiratory diseases (e.g., pleural inflammation, pneumonia)	CA 125, CEA
Sarcoidosis	CA 125
Systemic lupus erythematosus	CA 125
Urinary tract infection	PSA

[a]This list is not comprehensive; these markers may be raised in other nonmalignant conditions.
AFP, α-Fetoprotein; *CA 125,* cancer antigen 125; *CA 19-9,* cancer antigen 19-9; *CEA,* carcinoembryonic antigen; *hCG,* human chorionic gonadotrophin; *PSA,* prostate-specific antigen.

Given the long-term monitoring involved in cancer care, assay stability should be ensured over prolonged periods.

Quality control (QC) material not provided by the method manufacturer is preferable because kit controls may provide an overly optimistic impression of performance. At least one authentic serum matrix control from an independent source should be included in addition to QC materials provided by the manufacturer. Negative and low positive controls should be included for all tumor markers and should include concentrations close to important decision points (e.g., 0.1 and 3 or 4 μg/L for PSA; 4 to 7 μg/L for AFP; 5 U/L for hCG). The broad concentration range should be covered, and ideally a high concentration control should occasionally be included to check the accuracy of dilution, whether manual or on-board.

Proficiency Testing/External Quality Assessment

PT specimens should ideally be prepared from authentic patient sera, which for tumor markers may require dilution of high-concentration patient sera into a normal serum base pool. PT specimens should be commutable with patient specimens to ensure valid between-method comparisons. Specimens prepared by spiking purified analyte into serum base pools are likely to provide an overly optimistic impression of between-method performance. Specimens should assess performance over the working range and include assessment of linearity on dilution, baseline security, and stability of results.

Standardization

Major international efforts continue to be directed toward encouraging manufacturers to calibrate their methods accurately in terms of commutable International Standards (IS) or reference reagents, where available. Improved understanding of what is being measured is also critical if improved comparability is to be achieved. Manufacturers should provide clear information about the specificity of the antibodies used in their methods and data on cross-reactivity that is readily comparable with that of other methods so that users are aware of the differences.

Despite major international efforts to improve comparability, the molecular heterogeneity of most tumor markers means that results obtained using different methods are not interchangeable and considerable care is required in the interpretation of serial results obtained using more than one method.

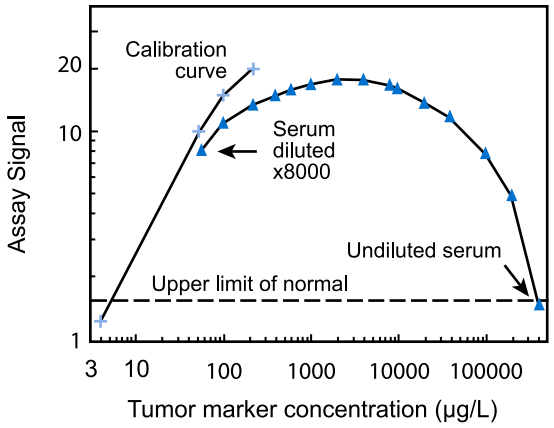


Fig. 20.1 Schematic illustration of high-dose hooking in tumor marker immunoassays. The extremely high tumor marker concentration blocks antibody-binding sites in the immunoassay such that the apparent result "hooks" back onto the calibration curve and an erroneously low result may be reported. [See text and Chapter 15 for further details.]

Clinically Relevant Interferences

Tumor marker measurements are subject to the same interferences as all immunoassays, but the following are particularly relevant:

High-dose "hooking." As tumor marker concentrations can range over several orders of magnitude, protocols enabling identification of high-dose "hooking" are essential to minimize the risk of reporting erroneously low results, particularly in patients for whom markers are being measured for the first time. An example is shown in Fig. 20.1. The risk of "hooking" is reduced by using solid phase antibodies of higher binding capacity, by using sequential assays that include a wash step, and by assaying specimens at two dilutions. Methods vary in their vulnerability to this interference (for additional information, refer to Chapter 15).

Specimen carry-over. Specimen carry-over is possible whenever high concentration specimens are assayed, so it is desirable to check periodically for this possibility. Serum concentrations of hCG can be greater than 1,000,000 U/L, so carry-over as low as 1/10,000 can still lead to a falsely positive result in the following sample.

Interference from heterophilic or human anti-mouse antibodies. Some patient sera contain anti-immunoglobulin (Ig)

antibodies (most often IgG) that may react with immunoassay reagent antibodies. High concentrations of human anti-mouse antibodies (HAMAs) may also be present in serum from cancer patients who have received treatment with mouse monoclonal antibodies for imaging or therapeutic purposes. For either interference, results may be falsely high or low. Identifying such interference requires a high degree of clinical suspicion, which is facilitated by availability of relevant clinical details. Possible interference can be investigated by assaying the specimen at several dilutions, by reassaying after treatment with a commercially available blocking agent, by adding further nonimmune mouse serum to the reaction mixture and reassaying, and/or by reassaying the specimen using a different method and preferably a different methodology. Caution should be applied in interpretation. For example, linearity on dilution does not always exclude the presence of HAMAs (for additional information, refer to Chapter 15).

Postanalytical Requirements

Brief information on the request form about the source of the suspected or diagnosed malignancy and the treatment stage (e.g., preoperative, postoperative, prechemotherapy) is highly desirable. This information should be recorded both in the laboratory computer and on the laboratory report. Reports should include cumulative and, if possible, graphical presentation of results because trends in tumor marker results are generally most informative. Such reporting facilitates interpretation of results and can also help to identify occasional errors in requesting or in the laboratory (e.g., incorrect sample identification or missampling on an analyzer), as well as highlighting unexpected results (e.g., sudden changes that are out of accord with the clinical picture) that require confirmation and further investigation. Brief comments relating to interpretation of the analytical results (e.g., whether an increase is likely to be clinically significant or not) and helpful advice about the frequency of monitoring and the need for confirmatory specimens is also desirable.

Urgent results that may be required for immediate patient management should be identified by the reporting biochemist so as to ensure that these reach the relevant clinician promptly (e.g., by telephone). These include tumor marker results that can be used to diagnose advanced disease in critically ill but treatable patients (e.g., AFP in hepatoblastoma, hCG in choriocarcinoma, AFP and hCG in nonseminomatous GCT, and PSA in men with advanced prostate cancer). The consequences of failure to telephone such results can be severe.

In view of the method-related differences discussed previously and as recommended by the NACB, the method used should be stated on the clinical report so that changes of method are readily identifiable. If there has been a method change, it is highly desirable that the laboratory indicates whether this is likely to influence interpretation of the trend in results. There should be a defined protocol for method changes and the likely effect communicated to clinical users prior to the change. Managing the change may necessitate analyzing the previous specimen by the new method or requesting a further specimen to reestablish the baseline and/or confirm the trend in marker concentrations.

Reference intervals should be derived using an appropriate healthy population and should ideally be specific to the method used. They are usually most relevant for cancer patients pretreatment because subsequently the patient's individual "baseline" results provide the most important reference point for most marker results. Provided "baseline" marker concentrations are well established in diagnosed patients post treatment, sustained increases even within the reference interval may be significant and should be treated as possible relapse.

The percentage increase or decrease that constitutes a significant change should be defined and should take account of analytical and biologic variation, as well as the time between samples. For tumor markers a confirmed increase or decrease of ±25% is frequently considered to be clinically significant.

Laboratories should be able to provide calculated tumor marker half-lives or doubling times for markers for which these are relevant (e.g., AFP and hCG). Half-lives are defined as the time to 50% reduction of circulating tumor marker concentration following complete removal of tumor tissue. Their calculation may be irrelevant if a 50% reduction does not represent a significant change. (See Germ Cell Tumor section for calculation details.)

Proactive provision of a high-quality tumor marker service helps to encourage good communication between laboratory and clinical staff and is likely to encourage appropriate use of tumor marker tests, as well as early identification of any results that are not in accord with the clinical picture and require investigation. An example of a laboratory report that fulfills many of these requirements is shown in Fig. 20.2.

POINTS TO REMEMBER

- Cancer is a heterogeneous disease that is a leading cause of death, although death rates for individual cancers vary markedly and there have been considerable advances in treatment over recent decades.
- Tumor markers are surrogate indicators that can help to make or confirm a cancer diagnosis, monitor treatment effectiveness and the course of disease, estimate prognosis, and/or predict whether a specific therapy is likely to be effective.
- Optimal use of tumor markers requires an evidence-based approach to clinical decision making and knowledge of the limitations of these tests particularly in relation to clinical sensitivity and specificity.
- Tumor marker results are rarely diagnostic and cannot replace biopsy for the primary diagnosis of cancer.
- Tumor marker measurements are not recommended for patients with vague symptoms when the population likelihood of cancer is low.
- A raised tumor marker result never necessarily indicates malignancy, and conversely a result within the reference interval never necessarily excludes malignancy.
- In diagnosed patients a pretreatment result is essential and provides the baseline against which subsequent results can be assessed.
- Tumor marker results should always be confirmed on a repeat specimen if decisions about therapy depend on the result.
- Tumor marker results should always be interpreted in the context of all available information, and the possible influence of other factors (e.g., medication, analytical effects) should be carefully considered.

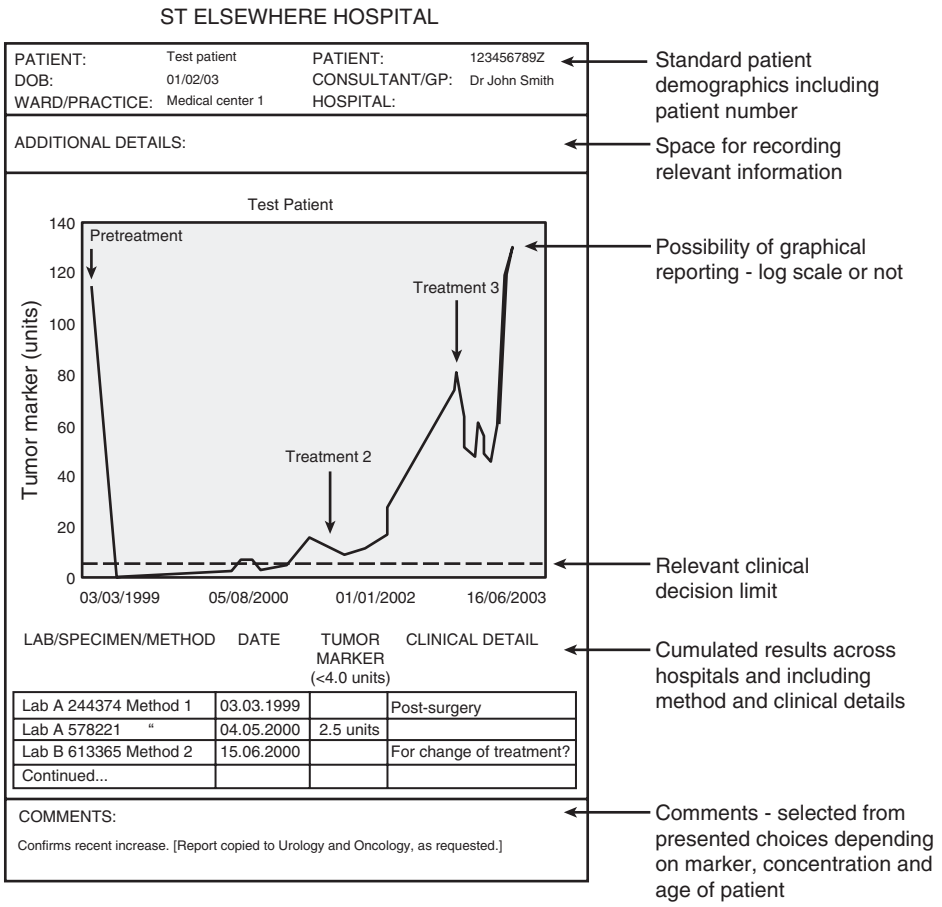


Fig. 20.2 Schematic diagram of a desirable clinical laboratory report for tumor markers that fulfills current reporting recommendations.

Use of Tumor Markers in Specific Malignancies

The extent to which tumor markers currently contribute to the management of a number of important malignancies is briefly reviewed in the following sections.

Bladder Cancer

Approximately 600,000 Americans are currently affected by bladder cancer, with 74,000 new cases diagnosed each year. The most common symptom is intermittent hematuria, which is present in 80% to 85% of patients, with voiding problems or dysuria also common. Transitional cell carcinomas (TCCs) account for the majority of bladder cancers, but adenocarcinomas, squamous cell carcinomas, and sarcomas occur. Diagnosis is usually established by cystoscopic evaluation. Primary treatment is complete surgical resection, usually by transurethral resection with or without additional immunotherapy, chemotherapy, or radiotherapy. There is a high risk of recurrence, and 50% to 70% of patients will develop tumor recurrence within 5 years, depending on the stage of disease. Lifelong surveillance is therefore required.

Tumor Markers Relevant to Bladder Cancer

Markers evaluated for bladder cancer include nuclear matrix proteins (NMPs), human complement factor H–related protein, fibronectin, telomerase, cytokeratins, and survivin. NMPs make up the internal structure of the nucleus and contribute to the regulation of some of the key functions that occur in the nucleus.

The FDA has approved an enzyme-linked immunosorbent assay (ELISA) for the measurement of nuclear mitotic apparatus protein, a component of the nuclear matrix that is overexpressed in bladder cancer. Approved uses of the NMP-22 test are as an aid in the diagnosis of symptomatic patients or those with risk factors suggesting TCC and in the management and monitoring of patients with TCC. A qualitative point-of-care version of the test is available as an aid in monitoring patients with a history of bladder cancer.

The bladder tumor–associated antigens (BTAs), human complement factor H–related protein, and related proteins are involved in the regulation of the alternative pathway of complement activation that prevents complement-mediated damage to healthy cells. The BTA-Trak and BTA-Stat tests have been approved by the FDA for use in conjunction with cystoscopy in the management of bladder cancer patients.

Breast Cancer

Breast cancer is the most common cancer in women worldwide, affecting 10% to 12% of women. In symptomatic women the main presenting features include a lump in the breast, nipple change or discharge, and skin contour changes. Worldwide the incidence appears to be increasing, but in some Western countries mortality rates are declining. This decrease has been attributed to earlier detection by systematic screening with mammography, greater awareness among women of early signs of breast cancer, and availability of adjuvant treatment for newly diagnosed cases. In the United States, breast cancer represents approximately 14% of all new cancer cases and nearly 3 million women are living with the disease. For the approximately 60% of breast cancer patients in the United States who have localized disease at diagnosis (i.e., confined to the primary site), the 5-year survival rate is 98.6% as compared with 25.9% for the 6% of patients who have distant metastases at diagnosis.

Primary treatment for localized breast cancer is either breast-conserving surgery and radiation or mastectomy. Following primary treatment, most women with invasive breast cancer receive systemic adjuvant such as chemotherapy, hormone therapy, or immunotherapy or a combination of these. Depending on estrogen receptor (ER) status and other factors, not all patients with breast cancer require adjuvant treatment.

Tumor Markers Used in the Management of Breast Cancer

Tissue tumor marker measurements essential for the management of breast cancer include those for receptors for estrogen, progesterone, and HER2, and CA 15-3 and the closely related BR 27.29 are the serum markers of choice. Measurement of CEA may also be helpful in patients with metastatic breast cancer. In early-stage disease, concentrations of CA 15-3 may be similar to those found in healthy women or women with benign breast disease.

Screening and Diagnosis. Early detection undoubtedly improves 5-year survival, but the low specificity and sensitivity of currently available serum tumor markers, especially in early stage disease, preclude their use in screening for breast cancer, and x-ray imaging with mammography remains the only screening modality.

Women at increased risk of breast cancer because of a strong family history of breast cancer (e.g., relatives diagnosed at a young age) or because they are carriers of the *BRCA1*, *BRCA2,* or *TP53* genetic mutations are likely to benefit from additional screening with MRI, which is current policy in the United Kingdom.

Definitive diagnosis of breast cancer requires biopsy and histology. Serum tumor markers do not contribute to this, but preoperative measurements of CA 15-3 and/or CEA are desirable if either marker is going to be used for posttreatment monitoring. A high CA 15-3 concentration (e.g., >40 to 50 kU/L) in a patient with apparently localized breast cancer should prompt further investigation to exclude the possibility of metastatic disease.

Prognosis. Accurate assessment of **prognosis** is essential for optimal management of breast cancer patients so as to avoid undertreatment of patients with advanced disease and overtreatment of patients with indolent disease. The extracellular serine protease urokinase plasminogen activator (uPA) and its endogenous inhibitor plasminogen activator inhibitor 1 (PAI-1) are the best validated tissue prognostic markers for breast cancer but are not currently widely used in clinical practice.

Multigene Profiling in Prognosis. Several gene expression profiles or multigene panels have been proposed for determining prognosis in breast cancer patients. The Oncotype DX test measures the expression of 16 cancer-associated and 5 control genes in breast tumor tissue at the RNA level and stratifies newly diagnosed patients with invasive breast cancer as being at low, medium, or high risk of recurrence.

Other gene profiles include the MammaPrint test, which measures the expression of 70 genes involved in signal transduction pathways responsible for breast cancer metastasis. MammaPrint is approved by the FDA to assist in assignment of women with ER-positive or ER-negative breast cancer into high- or low-risk groups for recurrence.

Preoperative serum concentrations of CA 15-3 predict shortened disease-free or overall survival in breast cancer patients and can provide additional prognostic information, perhaps because the marker is detecting micrometastases or occult disease that is not clinically evident. Because CA 15-3 measurement is relatively convenient and inexpensive, it may be appropriate to use it when planning optimal treatment of newly diagnosed patients with breast cancer, but further evaluation is required.

Therapy Prediction. Measurement of ER, the longest-established molecular marker for breast cancer, is mandatory in all newly diagnosed invasive breast cancer patients to determine whether anti-estrogen endocrine therapy is likely to be effective. Patients with ER-positive tumors tend to respond to hormonal therapy, whereas those with ER-negative tumors will be treated using chemotherapy or other therapies. The negative predictive value of ER is high (i.e., patients who are ER negative are very unlikely to benefit from endocrine therapy). However, in both early and advanced breast cancer, its positive predictive value is less accurate.

Patients with breast cancers that overexpress the *HER2* gene are candidates for treatment with anti-HER2 immunotherapies, which have been cleared by the FDA. Because amplification of the *HER2* gene occurs in only 20% of patients with invasive breast cancers, it is important to identify these patients so as to avoid treating patients unlikely to respond.

Detailed recommendations and good practice guidelines for measurement of ERs and progesterone receptors (PRs) have been published by the American Society for Clinical Oncology (ASCO) and the College of American Pathologists (CAP) and endorsed by the NCCN.

Monitoring. Most women who have been treated for breast cancer, with or without adjuvant therapy and/or curative surgery, are subsequently evaluated at regular intervals with history, physical examination, and annual mammography. The aim of follow-up is to identify early recurrence or metastatic disease on the assumption that early intervention will improve survival. If serum tumor markers are measured, CA 15-3, CEA, and/or BR 27.29 are the markers of choice.

However, the clinical value of serial measurements of CA 15-3 and other tumor markers is unclear. Their regular measurement can provide a lead time of 5 to 6 months, but whether early intervention based solely on tumor marker increase alone improves outcome is not well established and guideline recommendations vary. The NACB states that tumor markers can be measured where other methods of evaluation are not possible. The NCCN includes rising CA 15-3, CEA, and/or BR 27.29 in its definition of disease progression and states that marker increases raise the suspicion of tumor progression, also cautioning that such increases may be seen in responding disease. ASCO states that these markers may be used as adjunctive assessments, but not alone, to contribute to decisions regarding therapy in women with metastatic breast cancer. The Southwest Oncology Group is undertaking a prospective clinical trial to evaluate the impact of serial testing.

Cervical Cancer

Cervical cancer is the fourth most common cancer in women worldwide and a leading cause of cancer death in women in the developing world, where 85% of cases occur. In the United States, 12,360 new cases were diagnosed in 2014 and 4020 patients died of the disease. The most important factor in the development of cervical cancer is persistent infection with human papillomaviruses (HPVs). Primary prevention is therefore feasible with prophylactic vaccination.

Most women are asymptomatic and are frequently diagnosed only after abnormal cervical cytology is identified in a screening program. Introduction of widespread cytology screening in the United States decreased the incidence of cervical cancer from 14.8 per 100,000 women in 1975 to 6.6 per 100,000 in 2008, with a similar reduction in mortality. HPV testing is gradually being added to screening programs in a number of countries. Women with higher-grade abnormalities identified by cytology are referred immediately for colposcopy, whereas for low-grade or borderline cell abnormalities, the sample is automatically tested for HPV.

Treatment planning of patients with cervical cancer is primarily determined by the clinical stage of the disease and histologic type. Early-stage disease can be treated by radical hysterectomy and pelvic lymphadenectomy or radiotherapy. For more advanced disease, chemoradiation or neoadjuvant chemotherapy may also be required.

Tumor Markers Used in the Management of Cervical Cancer

Squamous cell carcinoma antigen (SCCA) is the serum tumor marker of choice for squamous cell cervical carcinoma, whereas CEA and CA 125 may have possible utility in patients with cervical adenocarcinoma. Serum SCCA concentrations correlate with tumor stage, tumor size, residual tumor post treatment, recurrent or progressive disease, and survival in patients with squamous cell cervical carcinoma. However, whether serum SCCA measurements are useful in clinical practice remains uncertain. There is no evidence that earlier detection of recurrent disease through tumor marker monitoring influences outcome or prognosis. Consequently, no currently available serum tumor markers are recommended for routine clinical use in patients with cervical cancer.

Colorectal Cancer

Colorectal cancer is the second leading cause of cancer-related death both worldwide and in the United States, where it is also the fourth most frequently diagnosed cancer. Typical symptoms include altered bowel habit, weight loss, anemia, and fecal occult blood. The natural history of the disease involves slow multistage progression from mucosal cell hyperplasia through adenoma formation, growth, and dysplasia, followed by malignant transformation and invasive cancer. In the United States, 5-year survival is approximately 90% for locally confined tumors, approximately 70% for those that have spread regionally, but only approximately 10% if distant metastases are present, highlighting the importance of detecting early-stage disease.

Tumor Markers Used in the Management of Colorectal Cancer

CEA, an oncofetal antigen first identified as being associated with colorectal cancer in the 1960s, remains the most relevant serum tumor marker for colorectal cancer.

Screening for Colorectal Cancer Using Tumor Markers. Several randomized controlled trials have shown that population-based screening using fecal occult blood testing (FOBT) can reduce mortality from colorectal neoplasia, with an average reduction in mortality of at least 16%. Based on this evidence, a number of countries have introduced screening for colorectal cancer, replacing the early guaiac-based tests with newer FIT that detect the globin component of hemoglobin. Because both types of test lack specificity for colorectal cancer, follow-up colonoscopy is required for positive tests.

Typically, patients collect fecal samples at home on a card that is mailed to a central testing laboratory for either test. FITs are easier to use than guaiac-based tests for a number of reasons. They can be automated, provide quantitative rather than qualitative results with adjustable cutoff points, are less vulnerable to interference (e.g., due to diet or drugs), have greater analytical specificity and better clinical sensitivity for cancers and advanced adenomas, and are cost effective. Ease of use is particularly important as the success of any screening program depends on its acceptability to the population being screened and consequently its uptake.

Analysis of genetic and/or epigenetic markers in fecal material forms the basis of an automated DNA screening test for colorectal cancer that has been approved by the FDA.

Individuals with a strong family history of colorectal cancer, hereditary nonpolyposis colon cancer (Lynch syndrome), or familial adenomatous polyposis (FAP) are at high risk of developing colorectal cancer and should be referred for risk assessment and endoscopic screening according to agreed protocols. Genetic counseling and germline gene testing for Lynch syndrome should be offered to patients or microsatellite instability (MSI) or mismatch repair (MMR) gene mutations.

Diagnosis. The low clinical sensitivity of CEA, particularly for early-stage disease, precludes its use in diagnosis. It will be increased in only 30% to 50% of patients at the time of diagnosis and cannot be used in isolation to diagnose even advanced disease. Its specificity is also low because it can be increased in nonmalignant liver and renal disease, as well as in other malignancies. However, clearly increased concentrations can assist in diagnosis in certain clinical circumstances and may be suggestive of metastatic disease, although they do not necessarily indicate a colorectal primary.

Prognosis, Staging, and Therapy Prediction. Increased preoperative CEA concentrations at the time of initial presentation of patients with colorectal cancer are associated with adverse outcome and can provide both prognostic information and a baseline value for interpreting subsequent results.

Patients with advanced colorectal cancer may be treated with anti-EGFR monoclonal antibodies, which bind to the extracellular domain of EGFR. Patients with a specific mutation of the *KRAS* gene appear to benefit most from such treatment, and modeling studies suggest that administering anti-EGFR antibody treatment only to patients with wild-type *KRAS* would result in significant net cost savings.

Monitoring. Meta-analyses of a number of randomized controlled trials designed to compare intensive follow-up with minimal follow-up of colorectal cancer patients following curative resection have demonstrated that intensive follow-up, which includes CEA, improves overall survival. It is now recommended that CEA testing should be performed every 3 to 6 months for 5 years following curative surgery for colorectal cancer, although results within the reference interval do not necessarily exclude progression.

An increase in CEA should prompt referral for early investigation of possible recurrence but should be confirmed by a second sample taken within approximately 1 month. What constitutes a significant increase in CEA is not yet well established. An elevation at least 30% over that of the previous concentration is regarded by some groups as significant, whereas smaller consecutive increases (e.g., 15% to 20%) observed on at least three successive occasions with intervals of at least 2 weeks between them may also prompt intervention.

Resection of metastatic deposits in the liver is an option that can prolong survival for a small but increasing number of patients with Dukes stage B or C disease, provided they are detected early, before symptoms develop. It is therefore recommended that CEA is measured every 3 months for at least 3 years following surgery in patients with Dukes stage B or C disease who would be candidates for liver resection. Measurement of CEA following liver resection provides a helpful indication of its success.

Its measurement is also recommended during chemotherapy for metastatic disease, but care in interpretation is required because CEA concentrations may be affected by factors other than tumor progression (e.g., liver damage). CEA-defined responses agree well with radiologic responses (the "gold standard") in greater than 90% of cases, enabling the conclusion that use of CEA is as accurate as CT imaging for assessing the response of colorectal cancer liver metastases to chemotherapy.

Monitoring of colorectal cancer outside a clinical trial should be undertaken only if it is likely to benefit the patient. CEA concentrations in individual patients should be monitored using the same method because results obtained in two methods may not be the same. If a change of method is unavoidable, re-baselining may be necessary, as described previously. Coordination of follow-up monitoring should ideally be supported by an automated computer system.

Gastric Cancer

Gastric cancer is relatively rare, although it is the second most common gastrointestinal cancer worldwide. Approximately 25,000 new cases are diagnosed in the United States each year, more than 60% of which will have already spread to regional lymph nodes or metastasized. Survival is closely related to stage at diagnosis, with 5-year survival approximately 65% for localized disease and less than 5% for patients with distant metastases.

Tumor Markers Used in the Management of Gastric Cancer

Although CEA and CA 19-9 are increased in 20% to 50% of patients with advanced disease and AFP in 20% to 25%, these markers are raised in less than 20% of patients with early-stage disease and consequently cannot be used for screening or diagnosis of gastric cancer. Measurement of serum CEA or CA 19-9 may provide early indication of recurrence in a proportion of patients, but their use is not recommended because there is currently no evidence that this improves clinical outcome.

This may change with availability of more effective systemic treatment because 10% to 25% of patients overexpress HER2. Because patients with advanced HER2-positive gastric tumors benefit from treatment with the anti-HER2 antibody, if such patients are being considered for systemic therapy, HER2 should be measured.

Gastrointestinal Stromal Tumors

Gastrointestinal stromal tumors (GISTs) present predominantly in middle-aged and older individuals and, although rare, are the most common mesenchymal tumors of the gastrointestinal tract, occurring in the stomach, small bowel,

large bowel, esophagus, or omentum. Increased awareness of this tumor type since the 1990s has led to appropriate pathologic diagnoses, with incidence estimated between 0.68 and 1.5 per 100,000 population in several Western countries.

Tumor Markers Used in the Management of Gastrointestinal Stromal Tumor

Molecularly, greater than 95% of GISTs are characterized by the presence of KIT protein (also known as CD117 antigen), whose measurement in serum is recommended by a number of expert groups. Mutations in the *KIT* gene occur in approximately 80% to 85% of all cases, and mutations in the homologous *PDGFRA* gene are found in 5% to 10% of GIST patients.

The outcome for patients with GISTs has been dramatically improved in recent years with the advent of tyrosine kinase inhibitors, which block *KIT* and *PDGFRA,* as well as BCR-ABL. Response to these drugs depends on the mutational status of the *KIT* gene with patients, so *KIT* and *PDGFRA* mutational analysis is also recommended prior to prescribing imatinib to patients with GISTs.

Germ Cell Tumors

Although representing only 1% of all male tumors, GCTs are the solid tumor most frequently diagnosed in men between the age of 20 and 34 years and account for approximately 95% of malignant tumors arising in the testes. They also occur elsewhere, particularly along the "midline" (i.e., the mediastinum, retroperitoneum, or perineal gland). The incidence of GCTs has increased over the past few decades, with 8430 new cases predicted in the United States in 2015. Risk factors include positive family history, cryptorchidism, testicular dysgenesis, Klinefelter syndrome, and a prior history of GCT.

Although GCTs are usually aggressive, they respond well to treatment with surgery and, where appropriate, adjuvant chemotherapy and/or radiotherapy. Anticipated 5-year survival rates are now 98% in the United States, although 5-year survival rates in patients presenting with advanced disease due to delay in diagnosis are only 50% to 60%. There are well-established national and international guidelines relating to the management of patients with GCTs. Treatment in specialist centers is advantageous, and all patients should be discussed at multidisciplinary team meetings at which a clinical biochemist is present.

Tumor Markers Used in the Management of Germ Cell Tumors

Measurements of hCG, AFP, and LDH are integral to the management of patients with GCTs. Tumor marker results should be reviewed together with histologic immunostaining and radiologic results and any inconsistencies noted when treatment decisions are considered.

Screening. The low prevalence of GCTs in the general population, together with the relatively low sensitivity and specificity of AFP, hCG, and LDH, preclude the use of tumor markers in screening.

Diagnosis. Although rare, the possibility of a GCT should be considered in any patient with a poorly defined epithelial malignancy, especially young individuals with midline masses. Plasma or serum AFP and hCG should be measured in any male with a suspicious lump in the testes and in any patient with malignancy of unknown origin. Hyperthyroidism may be a presenting feature if the hCG concentration is very high because hCG shares structural similarities with thyroid-stimulating hormone (TSH). Patients with clinical findings consistent with GCTs (e.g., testicular lump, lung metastases, and/or abdominal mass) and markedly increased serum concentrations of AFP and/or hCG should be urgently discussed with physicians at the regional cancer center. Referral for immediate chemotherapy may be appropriate before later surgery to remove residual tumor.

Measurement of serum tumor markers is essential for the differential diagnosis of GCTs because treatment differs according to whether the GCT is seminomatous, nonseminomatous (NSGCT) (with yolk sac elements), or a combined tumor with both seminomatous and nonseminomatous elements. Serum or plasma concentrations of AFP and/or hCG are increased in 80% to 85% of men with NSGCT, whereas less than 25% of those with seminomas have raised hCG and none have raised AFP. Nonseminomas are more clinically aggressive, so tumors with both elements are managed as for nonseminomas. Increased AFP precludes a diagnosis of seminoma. Marker measurements can sometimes modify histopathologic diagnoses. Increased AFP in a patient diagnosed with seminoma suggests that the AFP may not be related to the tumor (e.g., may be of liver origin) or that yolk sac elements were overlooked.

hCG and AFP should be measured before surgery for a suspected GCT to allow the rate of fall of the markers to be monitored post treatment.

As for all tumor markers, it is important to remember that concentrations within the reference interval do not necessarily exclude malignancy. Only a small proportion of seminomas or dysgerminomas (the female equivalent of seminomas) produce hCG and up to 25% of NSGCTs do not produce AFP or hCG.

As for other oncology applications, the hCG method used should recognize both intact hCG and its free β-subunit (hCGβ). hCG increases have to be interpreted with caution because hypogonadism and marijuana use may cause benign serum increases. Some immunoassay methods for hCG may be vulnerable to interference from **heterophilic antibodies** and **high-dose "hooking"** is always a possibility (see Interferences discussion earlier). Both AFP and hCG can be increased in nonmalignant diseases and other malignancies, and both are significantly raised in pregnancy.

Prognosis. The lowest tumor marker concentration reached post surgery, the primary tumor site, and the sites of metastatic disease all contribute to the prognostic classification of GCTs using criteria established by the International Germ Cell Cancer Collaborative Group. Treatment depends on whether the tumor is classified as having good, intermediate, or poor prognosis.

Monitoring. Provided disease is confined to the testis or ovary, serum AFP and/or hCG should decline to normal with respective apparent half-lives of 5 to 6 days for AFP and 1 to 2 days for hCG. Fig. 20.3 shows the decrease in AFP observed in a patient with widespread metastatic disease who had a complete radiologic and marker response to chemotherapy. The equation used to calculate the apparent half-life ($t_{½}$) of the tumor marker is

$$t_{1/2} = \frac{-0.3t}{\log_{10}\frac{[M]_T}{[M]_{T_0}}}$$

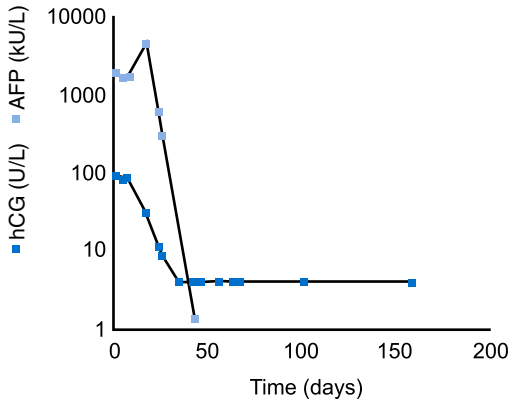


Fig. 20.3 Tumor marker responses in a patient with widespread metastatic disease who had a complete radiologic and marker response to chemotherapy. *AFP*, α-Fetoprotein; *hCG*, human chorionic gonadotropin.

where $[M]_T$ and $[M]_{T0}$ are tumor marker concentrations at time T and T_0, respectively, and t is the difference in days between T and T_0.

Further treatment with chemotherapy or radiation is required if AFP and/or hCG remains increased following surgery or if imaging identifies residual or metastatic disease.

Long-Term Surveillance. Following treatment, tumor markers should be measured regularly following defined clinical protocols, which depend on prognostic category and treatment. Clearly any significant increases in tumor markers must be reported immediately to the relevant clinical team. The importance of careful monitoring of tumor markers is illustrated in Fig. 20.4. After an initially very good response to a number of different chemotherapy agents, the patient required surgery, further chemotherapy, and ultimately a successful bone marrow transplant.

Gestational Trophoblastic Disease

Gestational trophoblastic disease (GTD) is a rare and previously fatal disease of pregnancy that develops in the trophoblast cells that surround the embryo immediately after conception. They may occur following a molar pregnancy, a nonmolar pregnancy, or a live birth. The possibility of GTD should be considered in any woman who develops hyperemesis, excessive uterine enlargement, or persistent irregular vaginal bleeding after a pregnancy or a failed pregnancy. A urine hCG pregnancy test should be performed in any woman presenting with such symptoms. Diagnosis

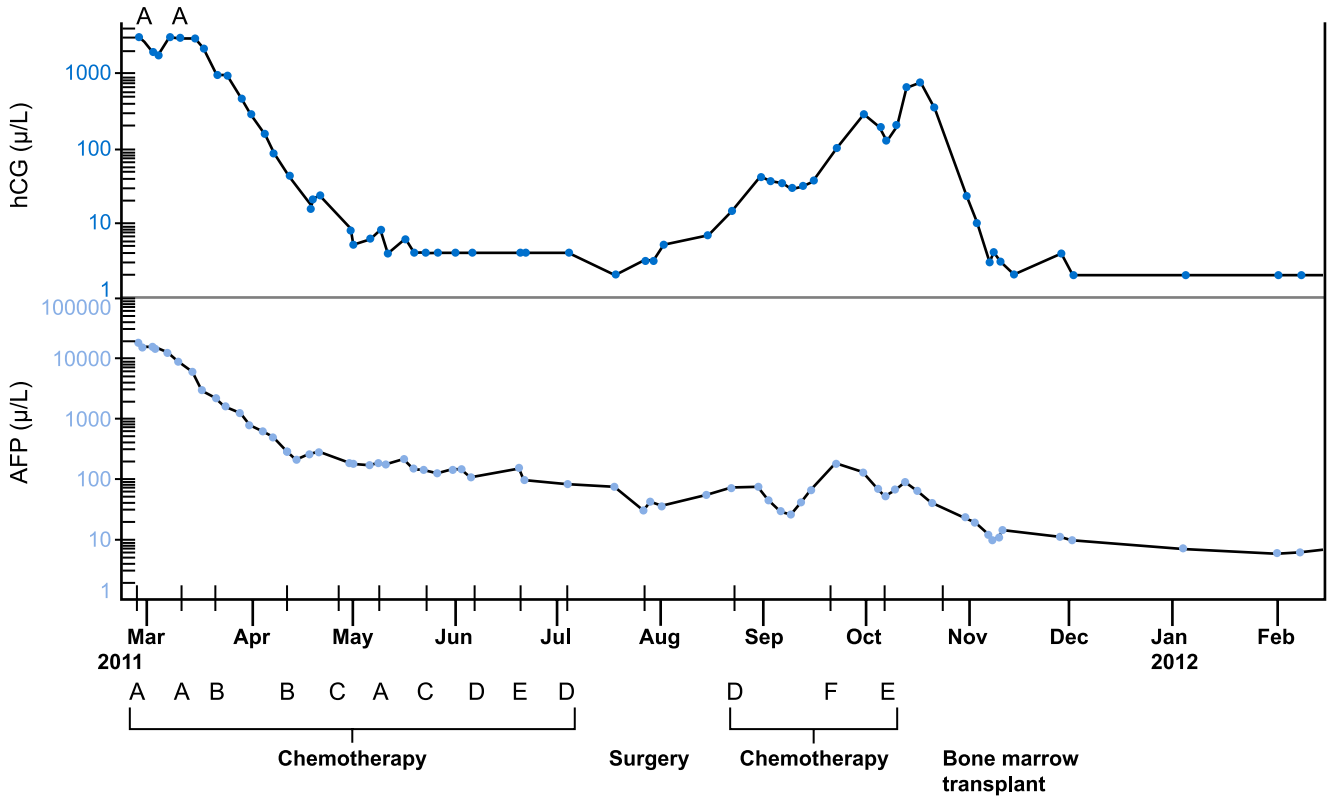


Fig. 20.4 Tumor marker monitoring of a patient who required a number of chemotherapy treatments prior to surgery, further chemotherapy and a bone marrow transplant. *AFP*, α-Fetoprotein; *hCG*, human chorionic gonadotropin.

requires histologic examination of the products of conception following surgical evacuation. If there is any evidence of persistence of GTD, which is most commonly defined as a persistent increase of serum hCG, the condition is referred to as GTN.

Characteristics of the main forms of GTD are briefly described next.

Hydatidiform Moles

Hydatidiform moles ("molar pregnancies") are the most common form of GTD and are nonmalignant. They may develop when a sperm fertilizes an "empty" egg containing no nucleus (a "complete" mole) or when two sperm fertilize a normal egg ("partial" mole). Fetal tissue is not present in complete moles. Following surgical removal, few patients with a partial mole require further treatment, and these moles rarely become malignant.

Invasive Moles

Following surgical removal of a complete mole, approximately 20% of women develop invasive moles. The tumor metastasizes to other sites (most often lung) in approximately 15% of patients. The risk of metastasis is increased if more than 4 months elapse between cessation of periods and treatment, if the uterus has become very large, if the woman is older than 39 years, or if she has had a previous GTD.

Choriocarcinoma

Choriocarcinoma most frequently develops from a complete hydatidiform mole but can occur after normal pregnancy or after early fetal loss in pregnancy. Metastasis to distant organs is more likely than for invasive moles but can be treated highly effectively with chemotherapy.

Placental Site Trophoblastic Tumors

These rare forms of GTD develop if the placenta remains attached to the uterus after a normal pregnancy or abortion. The tumors must be completely removed surgically and are not sensitive to chemotherapy, but they do not usually metastasize.

Tumor Markers in Gestational Trophoblastic Disease

Successful clinical management of GTD depends on accurate measurement of hCG, which has approximately 99% specificity and sensitivity for this disease in the clinical setting described previously, achieving almost ideal tumor marker performance in this application. Measurement of hCG using a method that recognizes both hCG and hCGβ is essential.

Human Chorionic Gonadotropin in Screening for Gestational Trophoblastic Disease

Screening of women at high risk of developing GTD (i.e., women who have had a partial or complete hydatidiform mole in a previous pregnancy) is conveniently performed by measuring hCG in urine collected by the patient in her own home and sent by mail to a central screening center. In the United Kingdom, there is a well-established registration system that enables lifetime follow-up, which is recommended because the disease can occur years after the affected pregnancy. Three specialist centers provide hCG testing in urine, with all patients referred for treatment to a specialist unit in London. Costs of the program are relatively low compared with the social cost avoided per life saved.

Human Chorionic Gonadotropin as an Aid to Diagnosis of Gestational Trophoblastic Disease. Definitive diagnosis requires histologic examination of the products of conception, but hCG serum measurements may be helpful in diagnosing molar pregnancies because values greater than two multiples of the median for the gestational date are suggestive of GTD.

Human Chorionic Gonadotropin in Prognosis. Prognosis is assessed according to the International Federation of Gynaecology and Oncology staging system for GTN, which includes pretreatment hCG concentration as one of eight prognostic factors used for scoring. Women at high risk of recurrence are treated with multiagent chemotherapy until the hCG concentration returns to normal and then for a further 6 weeks.

Monitoring

Posttreatment monitoring is essential in the follow-up of molar pregnancy and involves serial measurements of hCG in either blood or urine. A urinary hCG pregnancy test should be performed 3 weeks after medical management of failed pregnancy if products of conception are not sent for histologic examination. Following a diagnosis of GTD and surgical evacuation, if the hCG concentration has reverted to normal within 56 days, then follow-up will be for 6 months from the date of uterine evacuation. If hCG has not reverted to normal within 56 days, then follow-up will be for 6 months from normalization of the hCG concentration. In the United Kingdom, women who have had GTD in a previous pregnancy are requested to notify the screening center following any future pregnancies, whatever the outcome of the new pregnancy. hCG concentrations are then checked 6 to 8 weeks after the end of that pregnancy to exclude disease recurrence.

Hepatocellular Carcinoma (Primary Liver Cancer)

The 11th most commonly diagnosed cancer in the United States, HCC is the sixth most common cause of cancer death. Worldwide, however, HCC is the second leading cause of cancer death, with an incidence that has increased exponentially in recent decades. Asian countries account for approximately 78% of the 600,000 new cases of HCC reported annually. Risk factors are generally the same as those for liver cirrhosis. In China, Southeast Asia, and sub-Saharan Africa the major causative factors are chronic infection with hepatitis B virus (HBV) and ingestion of food contaminated with aflatoxin B, a fungal toxin that is genotoxic and carcinogenic. Nonviral causes associated with increased risk for HCC are more common in Western countries and include alcoholic cirrhosis, Wilson disease, and stage IV primary biliary cirrhosis. The increasing incidence of HCC in Europe probably reflects the increased frequency of chronic hepatitis C infection and liver cirrhosis, both of which are strong predisposing factors for HCC.

Signs and symptoms of HCC at presentation may include jaundice, hepatomegaly, ascites, peripheral edema, pruritus,

bleeding esophageal varices, and right upper quadrant pain. CT, MRI, and/or biopsy are required for diagnosis. Depending on imaging results, biopsy confirmation of a diagnosis may not be required for liver nodules greater than 1 cm in size.

Potentially curative treatment options include liver transplantation (for which only approximately 5% of patients are suitable), surgical resection (which requires very good liver function and has a recurrence rate of 50% to 60% after 5 years), ablative therapy (with alcohol, radiofrequency, and/or microwaves), chemoembolization, and chemotherapy. Eligibility for these treatments depends on diagnosing early stage disease, which is often asymptomatic. In developed countries, only 30% to 40% of HCC patients are diagnosed early enough for curative treatment.

Tumor Markers for Hepatocellular Carcinoma

AFP is the most clinically useful tumor marker for HCC. AFP concentrations may range from within the reference interval to as high as 10×10^6 μg/L (8.3×10^6 kU/L) but may not be increased in up to 20% of untreated patients with HCC, especially in early-stage disease. AFP concentrations may be transiently raised in regenerating liver nodules during recovery from viral hepatitis, as well as in some benign conditions.

Screening for Hepatocellular Carcinoma in High-Risk Groups Using α-Fetoprotein

There is increasing evidence that early detection of HCC through well-organized screening of high-risk populations with abdominal ultrasound and 3- or 6-monthly AFP measurements can improve clinical outcome. Studies suggest that serial AFP testing more accurately identifies those patients with hepatitis C virus (HCV) and/or advanced fibrosis or cirrhosis who are most likely to develop HCC. If serum AFP is rising or if a liver mass nodule is identified on ultrasound, the NCCN recommends additional CT or MRI imaging. According to the NACB, AFP concentrations greater than or equal to 20 μg/L (i.e., ≥17 kU/L) and rising should prompt further investigation even if ultrasound is negative.

α-Fetoprotein as an Aid to Diagnosis. The same caveats that apply to use of AFP for screening apply to its use as an aid to diagnosis. In general, ultrasound scan (USS)-detected nodules of less than 1 cm should be monitored at 3-monthly intervals but those of 1 to 2 cm in cirrhotic livers should be investigated by two imaging modalities (e.g., CT and MRI) and treated as HCC if imaging results are consistent with this. For liver lesions greater than 2 cm in size where USS is typical of HCC and AFP is greater than 200 μg/L (i.e., ~170 kU/L) and USS, a diagnosis of HCC can be made without biopsy. If no mass is detected, AFP testing and liver imaging should be continued every 3 months.

α-Fetoprotein in Prognosis. AFP concentrations may provide prognostic information in untreated HCC patients and in those being considered for liver resection or transplantation, with high concentrations generally associated with poor prognosis. Preoperative AFP decision limits ranging from 200 to 1000 μg/L (i.e., ~170 to 830 kU/L) have been suggested to predict outcome after liver transplantation, but there is no international agreement about the decision concentration. In the United Kingdom, current National Health Service (NHS) Blood and Transplant guidelines state that patients with serum AFP greater than 1200 μg/L (i.e., >1000 kU/L) should be deselected from eligibility for transplant.

α-Fetoprotein in Posttreatment Monitoring. Measurement of AFP at follow-up visits is recommended to monitor disease status after liver resection or liver transplantation, for detection of recurrence or after ablative therapies and/or during palliative treatment. This reflects the general consensus that earlier identification of disease may enable alternative treatments to be instituted or patients entered in a clinical trial. Current practice suggests monitoring patients with AFP every 3 months for 2 years and then every 6 to 12 months. Rising AFP concentrations should be confirmed with a repeat specimen within 3 months. Reevaluation, including AFP measurement, should be instituted earlier if there is clinical suspicion of disease recurrence.

Lung Cancer

Lung cancer has been the most common cancer and cause of cancer death worldwide for several decades, with an estimated 1.8 million new cases in 2012, 58% of which occurred in less developed regions. It is also the most common cause of death from cancer worldwide, accounting for nearly 20% of all cancer deaths. Survival rates depend strongly on tumor stage at diagnosis. Trends in incidence and mortality generally reflect smoking habits and/or exposure to other environmental or occupational carcinogens.

Early symptoms are nonspecific and include dyspnea, cough, and thoracic pain, with hemoptysis often indicating advanced disease. Investigations for suspected lung cancer include CT and/or MRI, bronchoscopy, and biopsy. Histologic differentiation and staging of lung cancer is essential because the two major histologic types have different clinical behaviors and sensitivity to chemotherapy and radiotherapy.

Approximately 75% to 85% of lung cancers are primarily NSCLC and require surgery as first-line treatment, with subsequent adjuvant radiotherapy and/or chemotherapy if appropriate. Several new biologic therapies are also available, targeting EGFR or vascular endothelial growth factor (VEGF). Small cell lung cancers (SCLCs) account for the other 15% to 25%. These often have neuroendocrine components and respond to chemotherapy and/or radiotherapy. They are aggressive tumors that double rapidly and metastasize early.

Tumor Markers in the Management of Lung Cancer

The serum tumor markers most frequently associated with SCLC are neuron-specific enolase (NSE) and progastrin-releasing peptide (ProGRP), whereas Cyfra 21-1 and CEA are primarily associated with NSCLC. SCCA may be raised in the subset of NSCLC patients with squamous cell carcinoma.

Screening and as Aids to Diagnosis. No serum tumor markers are sufficiently sensitive or specific, singly or in combination, for use in screening for lung cancer, even in high-risk groups such as smokers.

In patients in whom inoperable lung cancer is suspected but for whom histology is not available, clearly increased serum NSE and/or ProGRP support a diagnosis of SCLC and clearly raised serum SCCA concentrations are consistent with squamous cell cancer.

Prognosis. Measurements of Cyfra 21-1, CEA, and/or LDH may provide prognostic information for NSCLC, and NSE and LDH can be used similarly in SCLC.

Therapy Prediction. Serum tumor markers cannot be used to predict therapy response in lung cancer patients. However, as for colorectal cancer, anti-EGFR therapy is available for treatment of patients with NSCLC.

In agreement with other groups, ASCO recommends EGFR and ALK testing for all patients with lung adenocarcinoma or mixed lung cancer (both primary and metastatic) with a component of adenocarcinoma.

Monitoring. Monitoring of lung cancer patients with serum tumor markers is controversial in view of the often aggressive nature of the disease. Serial determinations following surgery may provide an indication of the completeness of tumor removal and an early indication of subsequent recurrence. The latter is likely to benefit the patient only if alternative treatment is available, and there is as yet no evidence to suggest that monitoring improves outcomes.

Melanoma

Melanoma is a skin cancer that can occur anywhere on the body but is most common in skin that is frequently exposed to sunlight. The incidence of melanoma has steadily increased over the past 30 years. Signs of melanoma include changes in the appearance of moles or pigmented areas of the skin. Surgery is curative in early-stage disease, but the prognosis is poor in metastatic disease.

Tumor Markers Used in the Management of Melanoma

Diagnosing malignant melanoma in pathologic specimens involves staining with antibodies to S100, a family of multifunctional proteins with regulatory roles in a number of cellular processes. S100 is increased in serum in patients with malignant melanoma, but its clinical utility has yet to be established and it lacks sensitivity in early disease. Rising concentrations are specific and sensitive for progression in patients with advanced disease, but its measurement is appropriate only if further treatment options are available. Although very nonspecific, LDH can be used to monitor patients with melanoma and may have prognostic value in advanced disease.

Therapy Prediction. The *BRAF*-mutated protein can be selectively targeted by a number of inhibitors that are very effective in patients with *BRAF* mutation–positive advanced melanomas. Commercially available predictive biomarker assays have now been cleared by the FDA for the identification of patients with metastatic or unresectable melanoma who are likely to benefit from these drugs.

Neonatal and Pediatric Tumors

In children younger than 15 years, cancer is the second leading cause of death, but advances in treatment mean that more than 70% of children diagnosed with cancer are now cured. Tumor markers can contribute significantly to the management of childhood neuroblastomas, malignant hepatic tumors, and GCTs. The general principles are the same as for adults with some additional caveats.

Tumor Markers Used in the Management of Childhood Malignancies

AFP and hCG are the tumor markers most frequently used in childhood malignancies. AFP is markedly increased at birth and then declines steadily to adult concentrations by 6 to 12 months. Age-related reference intervals are therefore essential. AFP is higher in preterm infants and may remain increased for longer in children with delayed development. Serial measurements are often more useful than isolated results, particularly in neonates, in whom acute hepatocellular damage may result in marked increases in AFP concentrations.

Additional care must be taken to avoid reporting results that are inappropriately low due to high-dose "hooking" because tumor marker concentrations may be very high in some childhood cancers. Considerable care is also required to ensure that dilutions of serum are made and recorded correctly. Because AFP and hCG requests for young children are relatively infrequent, it is highly desirable that all such samples are routinely assayed at more than one dilution.

Childhood Germ Cell Tumors

Measurement of AFP and hCG is mandatory because either or both are frequently increased at the time of diagnosis. Yolk sac tumors are the most common pure malignant GCTs in childhood. Dysgerminomas are the most common pure malignant GCT occurring in the ovary and central nervous system in girls, who may present with precocious puberty.

Hepatoblastoma

Hepatoblastoma and HCC are the most frequent malignant hepatic tumors in childhood and differential diagnosis is essential. More than 80% of children with hepatoblastoma and approximately 50% of those with HCC have tumors that produce AFP, often at extremely high concentrations (e.g., 1.2×10^6 μg/L [1.0×10^6 kU/L]). Children with hepatoblastomas that secrete hCG may develop precocious puberty. Complete surgical resection is the treatment of choice for hepatoblastomas and HCC, with adjuvant chemotherapy also important for hepatoblastomas.

Neuroblastoma

Neuroblastomas are malignant embryonal tumors that may exhibit extremely malignant behavior or may regress spontaneously. They account for 8% to 10% of childhood cancers, with approximately 80% of cases occurring in children younger than 4 years. Treatment includes surgery, chemotherapy, and radiotherapy. The diagnosis can be confirmed and the success of treatment monitored by measuring urinary catecholamines, which are increased in greater than 90% of cases.

Ovarian Cancer

Ovarian cancer is relatively rare (approximately 12.1 new cases per 100,000 women per year in the United States) and most frequently diagnosed in women aged between 55 and 64 years old. Five-year survival rates are greater than 90% for women diagnosed with disease confined to the ovary but approximately 28% for women with advanced disease at diagnosis. Early detection is difficult due in part to the relative rarity and heterogeneity of the disease, and as many as 30% of women with ovarian cancer may present for the first time at the emergency department.

Most malignant ovarian tumors (80% to 85%) are surface epithelial carcinomas that occur in five distinct histologic subtypes, predominantly serous. When evaluating the clinical utility of tumor markers in ovarian cancer, the different clinical behavior of the different subtypes must be considered.

GCTs account for approximately 15% of malignant ovarian tumors (see Germ Cell section) or sex cord stromal tumors, two-thirds of which are granulosa cell tumors.

Tumor Markers Used in the Management of Ovarian Cancer

CA 125 is the most widely used serum tumor marker for all epithelial ovarian cancers but is most sensitive for serous adenocarcinomas. CA 125 is raised in many nonmalignant conditions (Table 20.7), so careful interpretation of results in the clinical context is essential. Occasional and sometimes transiently raised concentrations of CA 125 greater than 5000 kU/L may be seen in patients with nonmalignant ascites, and CA 125 is increased in other malignancies, particularly adenocarcinomas. Some methods are vulnerable to interference, particularly from heterophilic antibodies.

Human epididymal protein 4 (HE4), a member of the "four disulfide core" family of proteinase inhibitors, has recently emerged as a promising serum tumor marker that could potentially complement measurement of CA 125 in ovarian cancer patients. Serum HE4 is less sensitive than CA 125 for detecting early-stage ovarian cancers among asymptomatic women but has better sensitivity and specificity for differentiating malignant from nonmalignant pelvic masses, especially in premenopausal patients. HE4 is not increased during menstruation but significantly increases in renal failure and with advancing age.

The serum markers of choice for ovarian GCT are AFP, hCG, and LDH, whereas for granulosa cell tumors, inhibin is the appropriate tumor marker. An inhibin method that detects all forms of inhibin, including inhibin A, inhibin B, and pro-αC, is required.

Screening for Ovarian Cancer. A reliable screening test for early ovarian cancer could make a major difference to outcome because most patients with advanced disease die of it. Assuming a disease prevalence of 40 per 100,000 in women older than 50 years, very high specificity (99.7%) is required to achieve an acceptable positive predictive value of 10% for ovarian cancer. The feasibility of screening women older than 50 years using CA 125 and ultrasound has been investigated extensively in several large major randomized controlled trials, with conflicting conclusions.

In the United Kingdom Collaborative Trial of Ovarian Cancer Screening (UKCTOCS), CA 125 was measured annually for 202,638 postmenopausal women. The number of screen-detected primary invasive epithelial ovarian cancers was double the number that would have been detected using a fixed cutoff, supporting the use of velocity-based algorithms in cancer screening strategies that use blood biomarkers. Survival data from this trial are awaited.

Women who are *BRCA1* or *BRCA2* mutation carriers are at high risk because 1.0% to 2.5% or 0.4% to 0.8%, respectively, will develop ovarian cancer each year. The American College of Obstetricians and Gynecologists recommends risk-reducing salpingo-oophorectomy (i.e., total removal of the ovaries and fallopian tubes) by age 40 years for these women. For women declining this operation, screening with transvaginal ultrasound and/or CA 125 can be offered at the clinician's discretion. Outcome data from the United Kingdom Familial Ovarian Cancer Screen Study (UK FOCSS) and from a prospective longitudinal CA 125 screening study of high-risk women by the Gynecologic Oncology Group are awaited.

Diagnosis. CA 125 measurement in isolation is not recommended for diagnosis due to its low sensitivity and specificity for ovarian cancer. A level within the reference interval does not necessarily exclude ovarian cancer, and an increased concentration does not necessarily indicate it, particularly in premenopausal women. Nevertheless, extremely high CA 125 values may be helpful in evaluating premenopausal women provided other potential causes are excluded.

In postmenopausal women, CA 125 measurement can help in differentiating malignant from nonmalignant pelvic masses using well-established RMI algorithms, which incorporate ultrasound, menopausal status, and CA 125. At an RMI cutoff value of 200, the positive predictive value for malignancy is approximately 80%.

Women with ovarian cancer may develop nonspecific symptoms several months before diagnosis of early-stage disease. Reflecting this, UK National Institute for Health and Clinical Excellence (NICE) guidelines recommend that serum CA 125 be measured in women presenting with persistent, frequent (>12 times a month), and continuous symptoms including abdominal bloating or distension, early satiety and/or loss of appetite, pelvic or abdominal pain, and/or increased urinary urgency and/or frequency. If CA 125 is greater than or equal to 35 kU/L, an USS of the abdomen is arranged and the RMI calculated. Women with RMI greater than 250 are referred urgently to a specialist team. Audit will be required to assess the cost effectiveness of this approach.

Prognosis. Higher preoperative CA 125 concentrations are generally associated with worse prognosis. Five-year survival rates in patients with a preoperative CA 125 concentration greater than 65 kU/L have been found to be significantly lower than those for patients with CA 125 less than 65 kU/L. CA125 values greater than 500 kU/L may predict the occurrence of difficult or complex surgery, potentially aiding treatment planning.

Following primary treatment with surgery and/or chemotherapy, the rate and extent of fall of CA 125 during the initial two cycles of platinum-based chemotherapy generally reflect prognosis. Factors influencing CA 125 concentrations post surgery must be taken into account to minimize the risk of misinterpreting results.

Monitoring. Criteria used to evaluate changes in CA 125 during treatment and follow-up are included in the Response Evaluation Criteria in Solid Tumors (RECIST). A CA 125 response is defined as a 50% decrease from the pretreatment concentration provided the decrease is maintained for at least 28 days. The RECIST criteria enable objective assessment of response, which can influence decisions to continue effective treatment or change ineffective treatment.

Following completion of treatment, standard practice has been to monitor complete remission and then potential increases in CA 125 prior to clinical signs of relapse (median lead time approximately 4 to 5 months). The UK-based Medical Research Council/European Organisation for Research and Treatment of Cancer (MRC-EORTC) conducted a randomized trial to evaluate the utility for such monitoring of CA 125. Results indicated that early intervention based on CA 125 conferred no survival advantage but resulted in earlier deterioration in quality of life. It was concluded that the value of routine measurement of CA 125 in the follow-up of patients with ovarian cancer who attain a complete response after first-line treatment is not proven and that these patients should be told that routine CA 125 monitoring is not necessary until or unless they are worried or develop any signs suspicious of tumor relapse.

Results of the MRC-EORTC trial challenge existing assumptions and have generated considerable debate, with a number of aspects of the study questioned. The European Society of Gynaecological Oncology (ESGO) issued a consensus statement advising that the use of CA 125 should not be universally abandoned in the routine follow-up of all patients with ovarian cancer. Their recommendations are broadly in line with those of the MRC-EORTC trial coordinator and of other groups, most of which recommend that the pros and cons of monitoring with CA 125 should be objectively discussed with the patient.

Pancreatic Cancer

Pancreatic cancer is the 12th most common cancer in the United States, representing 3% of all new cancer cases and is the fourth leading cause of cancer death. Most patients with pancreatic cancer present with symptoms that may include jaundice, weight loss, abdominal pain, nausea or vomiting, and/or pruritus. Only 9% are diagnosed when the tumor is still localized and resectable.

Tumor Markers Used in the Management of Pancreatic Cancer

The most clinically useful serum tumor marker for pancreatic cancer is CA 19-9, a sialylated Lewis a antigen (Le^a) commonly produced in pancreatic and hepatobiliary disease but also increased in many other conditions (Table 20.7). CA 19-9 is undetectable and therefore uninformative in the 5% to 10% of the general population who are Le^a negative.

Clinical Applications of CA 19-9 in Pancreatic Cancer

Screening and Diagnosis. The low positive predictive value of CA 19-9 means it cannot be used in screening, but in symptomatic patients it can contribute to diagnosis. It should be measured only in patients in whom pancreatic cancer is suspected and with the usual caveats that a result within the reference interval does not exclude pancreatic malignancy and vice versa. CA 19-9 may be increased in biliary infection (cholangitis), inflammation, or biliary obstruction, so it is best measured after complete biliary decompression and when bilirubin is normal. It is frequently increased in patients with biliary obstruction due to benign pancreatobiliary and other conditions.

Prognosis and Therapy Prediction. Preoperative CA 19-9 concentrations correlate with staging and disease burden, so they can complement information from imaging, laparoscopy, and biopsy when assessing whether a pancreatic tumor is resectable. In patients without biliary obstruction, a high CA 19-9 concentration can be used as an indication for staging laparoscopy, with concentrations greater than 130 kU/L strongly associated with subradiographic unresectable disease. Low postoperative CA 19-9 concentrations at 3 months and before adjuvant chemotherapy are independently prognostic, with normalization of CA 19-9 following neoadjuvant therapy associated with longer median survival. Postoperative CA 19-9 concentrations may also help to predict the benefit of adjuvant therapy.

Monitoring. Rising CA 19-9 concentrations suggest progression and can be used, together with radiographic and clinical data, to influence decisions to initiate palliative treatment in a patient whose disease progresses after surgery or to change (or discontinue) chemotherapy in patients progressing during treatment. Obtaining a CA 19-9 measurement immediately before each therapeutic intervention is essential to have a baseline with which to compare subsequent results. ASCO does not recommend the routine use of serum CA 19-9 alone to monitor response to treatment but states that in advanced disease CA 19-9 can be measured at the start of treatment for locally advanced metastatic disease and every 1 to 3 months during active treatment. An increase may indicate progression and should prompt confirmation with other modalities.

It is particularly important that the same CA 19-9 method be used for serial monitoring because results are not interchangeable between methods.

Prostate Cancer

Prostate cancer is most frequently diagnosed in men aged 65 to 74 years old. It accounts for approximately 13.3% of all new cancer cases in the United States in 2015 but for only 4.7% of all cancer deaths. Although many more men with prostate cancer die with, rather than of, their cancer (by the age of 80 years, autopsy studies show that approximately 80% of men have some cancer cells in their prostate), some prostate

cancers are lethal. Developing a reliable and noninvasive means of differentiating aggressive cancers from indolent cancers remains a major challenge.

Common symptoms at presentation include frequency, urgency, nocturia, dysuria, acute retention, prostatic enlargement, back pain, weight loss, and anemia. The first six of these are also associated with benign prostatic hyperplasia (BPH), a nonmalignant condition reflecting increased prostate size with advancing age. The last three symptoms tend to be symptomatic of more advanced prostate cancer.

Treatment options for early-stage disease include curative surgery (prostatectomy), brachytherapy, and radiotherapy, and "active surveillance" without intervention is also a realistic option. In advanced metastatic disease, endocrine (antiandrogen) therapy is effective because prostate cancers are dependent on testosterone. However, many advanced cancers eventually no longer respond to traditional androgen-reducing treatments and other treatment options, including chemotherapy, are required.

Widespread measurement of serum PSA, which can identify potential prostatic disease in both asymptomatic and symptomatic patients, informs decisions to proceed to prostatic biopsy, and probably accounts for an apparent increased incidence of prostate cancer. However, many of the cancers identified are indolent and such overdiagnosis is a major issue for health systems.

Tumor Markers Used in the Management of Prostate Cancer

Prostate-Specific Antigen. Optimal treatment of patients with prostate cancer relies on PSA measurements. Unlike many other tumor markers, PSA is essentially organ-specific. However, PSA is not cancer-specific because increased serum concentrations occur in men with benign prostatic disease and/or urinary tract infections and following intervention involving the prostate. Prostatic biopsy is therefore required for definitive diagnosis.

A member of the kallikrein serine protease family, 70% to 90% of immunoreactive PSA protein in blood is bound to protease inhibitors, primarily α_1-antichymotrypsin (ACT). The remaining 10% to 30% of the immunoreactive protein circulates in a free or unbound form, thought to be biologically inactive and known as "free PSA." Multiple forms of free PSA and its precursor forms exist in serum and have been proposed as new markers for prostate cancer. In clinical practice, methods measuring total PSA (i.e., free PSA and the PSA-ACT complex) are most often used.

Several strategies to improve the diagnostic accuracy of PSA with the primary aim of decreasing the number of unnecessary biopsies are described as follows.

Age-Related Reference Intervals. Prostate size can increase with age, particularly in men with BPH, and may be associated with increased serum PSA values. Historically, a single PSA reference interval of less than 4.0 μg/L has been used. Higher PSA reference intervals in older men could decrease unnecessary biopsies (i.e., increase specificity), whereas lower reference intervals in younger men could increase cancer detection (i.e., improve sensitivity). Race-specific reference ranges have similarly been suggested because PSA tends to be higher in some populations (e.g., African-Americans). The value of age- and/or race-related reference intervals is still unclear. The NHS Prostate Cancer Screening Management Programme recommends use of age-related reference intervals, but other groups make no recommendations regarding their routine use.

Percent-Free Prostate-Specific Antigen ("Free to Total" Ratio). The ratio of free to total PSA is lower in prostate cancer patients than in men with BPH and those without prostatic disease. Determination of the ratio may therefore enable better discrimination of prostatic cancer and BPH. Measurement of percent-free PSA is most informative when total PSA is between 3 or 4 and 10 μg/L. Although measurement of percent-free PSA has been shown to reduce the number of unnecessary biopsies by 25% to 40%, some cancers may be missed.

Preanalytical requirements for measurement of free PSA are more rigorous than for total PSA because its half-life in blood is much shorter than that of total PSA (approximately 2.5 hours compared with 3 to 4 days). Later addition of the test to a stored sample may therefore be inappropriate. Percent-free PSA must be determined using paired kits from the same supplier.

Prostate-Specific Antigen Density. PSA density potentially differentiates men with high serum PSA due to a large prostate from those with prostate cancer. PSA density is estimated by dividing serum PSA (μg/L) by prostate volume (cc) as determined by transrectal ultrasound (TRUS).

Some studies suggest that this improves PSA specificity in men with larger prostates and improves PSA sensitivity in men with smaller glands, but other studies have not confirmed this. PSA density is rarely used in clinical practice.

Prostate-Specific Antigen Velocity. PSA velocity (PSAV), the rate of change of PSA over time, has been suggested to reflect aggressiveness of prostate cancer, with higher rates of change associated with greater risk of prostate cancer death. There are a number of caveats to the use of PSAV, which is not useful in patients with PSA values greater than 10 μg/L.

Other Tumor Markers for Prostate Cancer

Tests that are not at present widely used in clinical practice include PCA3 (a noncoding, prostate tissue–specific RNA measured in urine), phi (a combination of total PSA, free PSA, and proPSA measured in serum), 4K score (a combination of free and total PSA, human kallikrein 2, and intact PSA measured in serum), and ConfirmMDx (a multiplex epigenetic assay).

Prostate-Specific Antigen in Screening for Prostate Cancer. Screening asymptomatic men for prostate cancer with PSA is highly controversial, primarily because its inability to differentiate aggressive and indolent prostate cancers can inevitably lead to overdiagnosis and overtreatment of some men.

The feasibility of screening with annual PSA measurements has been investigated in five trials that were the subject of a Cochrane systematic review. Three trials used methodology that was assessed as posing a high risk of bias. The two larger studies, the European Randomized Study of Screening for Prostate Cancer (ERSPC) and the US Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial, were both assessed as posing a low risk of bias but reported contradictory results.

The ERSPC study was the only one of the trials to report significant (21%) reduction of prostate cancer–specific mortality in a prespecified subgroup of men aged 55 to 69 years after 13 years of follow-up. After the same length of follow-up, the PLCO study found no evidence of a mortality benefit for organized annual screening as compared with opportunistic screening, which forms part of standard care. Coordinators of the ERSPC study concluded that the balance of benefit and harm in prostate cancer screening has not yet been adequately characterized. Authors of both studies agreed that population screening should not be implemented at present and that asymptomatic men requesting PSA measurement should be objectively informed of the risks of overdiagnosis and overtreatment and potential adverse side effects before undergoing the test.

Prostate-Specific Antigen as an Aid to Diagnosis of Prostate Cancer. Men with symptoms suggestive of prostate cancer should be offered a PSA test and a DRE before referral for specialist care. A PSA result within the reference interval does not necessarily exclude prostatic disease, and an increased PSA does not necessarily indicate malignancy. However, in general the higher the PSA result the greater the likelihood of prostate cancer, providing prostatitis, urinary tract infection, catheterization, and other possible confounding factors that may increase PSA have been excluded. Results above the upper limit of the reference interval but less than 10 μg/L are considered to be within a "gray zone" where additional testing (e.g., percent free PSA) may be undertaken in some centers before proceeding to biopsy. A second specimen to confirm a raised PSA result is always desirable.

In general, diagnosis of prostate cancer requires biopsy and pathologic confirmation unless clinical suspicion of malignancy is high, there is evidence of bone metastases, and the PSA concentration is clearly increased (e.g., >100 μg/L). Prior to prostatic biopsy, the benefits and potential side effects should be discussed with the patient, and individual risk factors (e.g., age and ethnicity) and comorbidities carefully considered.

Prostate-Specific Antigen in Prognosis and Therapy Prediction of Prostate Cancer. In patients with immunohistochemically confirmed prostate cancer, treatment depends on whether the disease is confined to the prostate or has spread to other organs, so accurate pretreatment staging is crucial. The pretreatment serum PSA value correlates with the risk of extraprostatic extension and lymph node involvement and is an independent predictor of response to all forms of treatment. Predictive tables or method-specific nomograms are available that combine pretreatment PSA concentrations with information on clinical stage and the immunohistochemical Gleason score, which assesses the degree of cellular differentiation in the prostate tissue examined and ranges from 2 in well-differentiated tumors to 10 in completely anaplastic tumors.

Such nomograms can provide a reasonable indication of whether disease is localized to the prostate and aid in predicting treatment outcome. For example, a patient with a PSA of less than 10 μg/L and a Gleason score of less than or equal to 6 is unlikely to have metastatic disease so pelvic lymph node dissection may not be necessary, and treatment options will include radical prostatectomy, localized radiotherapy, or active surveillance as described previously.

Pretreatment PSA can contribute to assessment of the need for bone, CT, and/or MRI scans for staging asymptomatic men with localized prostate cancer. Bone scans are generally not necessary in patients with newly diagnosed prostate cancer who have a PSA less than 20.0 μg/L unless history or clinical examination suggests bony involvement and/or if the patient has a Gleason score greater than 8. Similarly, CT is rarely positive when pretreatment PSA is less than 20.0 μg/L unless disease is locally advanced or the Gleason score is greater than or equal to 8.

Following radical prostatectomy, PSA should decrease to and remain at undetectable concentrations, with measurable concentrations indicative of residual disease or possibly the presence of benign glands. An initial post-prostatectomy PSA value of greater than 0.2 μg/L followed by a subsequent confirmatory value is likely to indicate residual disease or recurrence but may not indicate need for initiation of treatment because the median interval from PSA recurrence to cancer death is between 5 and 12 years.

PSA decreases more slowly after radiotherapy, rarely falling to less than 0.2 μg/L. However, the level should remain stable. Any rise in PSA value greater than or equal to 2.0 μg/L predicts treatment failure with great sensitivity and specificity after radiotherapy with or without androgen deprivation.

Following hormonal therapy, a low PSA nadir can be linked to survival. Failure to achieve a PSA nadir of less than 4.0 μg/L 7 months after initiation of therapy is associated with very poor prognosis (median survival approximately 1 year), whereas patients achieving a nadir of less than 0.2 μg/L have relatively good prognosis (median survival >6 years).

Prostate-Specific Antigen in Monitoring of Prostate Cancer. Sustained increases in PSA after treatment provide objective evidence of disease progression, often months before other diagnostic procedures. Consistently rising PSA usually, although not always, indicates recurrence. Conversely, a stable PSA does not necessarily exclude progression if clinically suspected. An increase in PSA should always be confirmed, taking into account intraindividual biologic variation in PSA of up to 20% to 30% before concluding an increase is clinically significant. For patients in whom it can be calculated, the PSA doubling time can provide valuable information about the likely time to progression. Patients with a PSA doubling time greater than 15 months have a low likelihood of prostate cancer–specific mortality over a 10-year period. In some patients, knowledge of increasing PSA concentrations may have adverse psychological

consequences, and monitoring with PSA may be undesirable, particularly if effective alternative treatments are not available.

Testicular Cancer

GCTs account for more than 90% of testicular tumors in adults (see earlier). Leydig and Sertoli cell tumors develop in the stroma, the hormone-producing tissue of the testicles, and represent 4% of adult testicular tumors and 20% of childhood tumors. Leydig cell tumors normally produce androgens. They usually do not spread and can be cured by surgical removal. Those that do spread are usually resistant to both chemotherapy and radiation. Sertoli cell tumors are very similar and are also difficult to treat if they spread beyond the testes.

As for granulosa cell tumors of the ovary, the marker of choice for both Sertoli and Leydig cell tumors is inhibin and the same analytical requirements apply.

Thyroid Cancer

Thyroid cancers represent approximately 2% to 4% of cancers in all age groups, with women more than 3 times likely to be affected. There are four histologic types: papillary, follicular, and anaplastic thyroid cancer and medullary thyroid carcinoma (MTC). The diagnosis is often made when a patient presents to the thyroid clinic with a painless neck lump. Because the major differential diagnosis required is between benign and malignant thyroid nodules, the reader is referred to Chapter 42, but use of the relevant tumor markers is described briefly here.

Tumor Markers Used in the Management of Thyroid Cancers

Papillary and follicular thyroid cancers are differentiated thyroid cancers and can be monitored post treatment by measuring serum thyroglobulin. A large complex glycoprotein synthesized by the follicular cells of the thyroid and stored in the colloid space, thyroglobulin is present in benign and malignant thyroid tissue. Its production is stimulated by TSH, either from the pituitary or by injection of recombinant human TSH. In patients with an intact thyroid, thyroglobulin reflects thyroid volume and injury and also TSH receptor stimulation. It should be undetectable in serum in the absence of thyroid cells.

MTC is a rare and challenging malignancy that often presents as a lump in the neck, with dysphagia or systemic effects such as diarrhea or flushing. The 10-year survival is approximately 75%, but patients may survive for many more years even with a significant tumor burden. Most appropriately managed in specialist centers, MTC has a hereditary component in approximately 25% of cases and can be associated with multiple endocrine neoplasia (MEN) 2B. Germline testing for the RET proto-oncogene distinguishes hereditary from sporadic MTC and should be offered to all patients with a family history consistent with MEN 2 or familial MTC. Prophylactic thyroidectomy may be indicated in children, and screening for other endocrine tumors (e.g., pheochromocytoma, pituitary, parathyroid) is essential. Calcitonin, a 32–amino acid polypeptide produced by the C cells of the thyroid is an excellent tumor marker for MTC. CEA is also a very good marker of tumor burden in MTC.

Screening, Diagnosis, and Prognosis. Serum thyroglobulin measurement is not helpful in screening for or diagnosis of thyroid cancer and has no value as a prognostic marker. Basal calcitonin measurements can contribute to the diagnosis of C-cell hyperplasia, and MTC and CEA should be measured preoperatively.

Monitoring. Thyroglobulin has an important role in monitoring patients with papillary or follicular thyroid cancer after treatment with surgery and/or radioiodine. Following either of these treatments, TSH is usually kept suppressed by adjusting the replacement dose of thyroxine to minimize the risk of stimulating the growth of any remaining thyroid cancer cells. If TSH is suppressed, serum thyroglobulin should be undetectable following complete ablation. A confirmed rise in thyroglobulin is consistent with recurrence.

Calcitonin and CEA should be measured before surgery and 6 weeks post surgery. Any increase should be confirmed with an early repeat determination using the same analytical method. Provided the concentrations fall to within reference limits following treatment, annual monitoring may be adequate after the first year.

Analytical and Reporting Requirements for Thyroglobulin. There are several potential pitfalls in thyroglobulin measurement. Analytically, RIA and immunometric methods for thyroglobulin are vulnerable to interference from anti-thyroglobulin autoantibodies that may be present in some patient sera. Results obtained may be falsely low or falsely high, with immunometric methods being particularly prone to giving falsely low results. Appropriate cautions should be added to reports.

Cancers of Unknown Primary

When cancer is found in one or more metastatic sites but the primary site of origin cannot be determined, the diagnosis made is of cancer of unknown primary or occult primary cancer. Such cancers account for 2% to 3% of all new cancer diagnoses, with approximately 31,000 cases diagnosed per year in the United States. Symptoms vary according to the sites of metastases and may include swollen lymph nodes, abdominal mass, shortness of breath, pain in the chest or abdomen, bone pain, skin tumors, anemia, weakness, fatigue, poor appetite, and/or weight loss. The average survival time is approximately 9 to 12 months after diagnosis.

Routine clinical investigations to demonstrate the organ of origin are generally unproductive, with an approximate 30% rate of identification, which has not been associated with significant improvement in survival. However, for a small subset of patients with cancer of unknown primary (possibly <1%), tumor marker measurements can contribute to diagnosis of an unsuspected but potentially treatable malignancy that has presented in an atypical manner. These include gestational choriocarcinoma, GCT of the testis or ovary, prostate cancer, ovarian cancer, and breast cancer.

Tumor Markers Used in the Management of Cancers of Unknown Primary

Measurement of AFP and hCG is recommended in patients with midline metastatic disease (mediastinal nodes or presence

POINTS TO REMEMBER

- Carefully used, tumor markers can contribute usefully to patient management but awareness of their limitations is essential.
- Measurement of tissue markers including *ALK, BRAF, HER2, KRAS,* and the ER and PR informs selection of endocrine and immunotherapy in individual patients in NSCLC, melanoma, and/or breast cancer.
- In the management of GCT, AFP, and hCG measurement in serum is mandatory.
- Measurement of CEA in serum is recommended for postoperative follow-up of stage II and III colorectal cancer patients if further treatment is an option.
- PSA measurement in serum contributes to diagnosis of prostate cancer and may be used for detecting recurrence and monitoring therapy.
- Measurement of serum AFP, CA 125, or CA 19-9 may aid early detection of HCC, ovarian cancer, or pancreatic cancer, respectively, in appropriately selected high-risk patients.
- Opportunistic screening with panels of tumor markers is not helpful.
- Serial results are nearly always more useful than single isolated results because the main application of tumor markers is in monitoring.
- Excellent communication between clinical and laboratory colleagues is essential for optimal use of tumor markers.

of a retroperitoneal mass) to assess the possibility of a GCT. CA 125 measurement is recommended in women with symptoms that could be consistent with an occult non–germ cell ovarian malignancy, together with a gynecologic-oncologic consultation, if indicated. Men presenting with bone metastases or multiple sites of involvement should have PSA levels assessed. Chromogranin A measurement is recommended by the European Society for Medical Oncology (ESMO) in patients with features of neuroendocrine tumors. Measurement of estrogen and PR may identify hormone-sensitive tumors amenable to specific therapy. In patients in whom a pancreatic tissue origin is suspected, measurement of CA 19-9 is recommended by the NCCN. The usual caveats apply to the use of these tumor markers in diagnosis. If their measurement identifies a likely primary source of malignancy (e.g., a GCT), monitoring protocols, if required, should be according to protocols previously described for these cancers and tumor markers.

CONCLUSION

Tumor marker measurements contribute significantly to the management of cancer patients, but considerable care is required to ensure that their use is appropriate. The clinical laboratory should proactively encourage correct test selection, sample handling, analysis, reporting, and interpretation of tumor marker results, taking particular care to remind users of the caveats associated with these tests. This is essential not only for the well-established tumor markers but also for the new generation of molecular and genetic tumor markers on which clinicians increasingly will rely on to provide optimal care. Involvement of laboratorians in developing clinical pathways that will enable optimal use of tumor markers—both the well established and the newly developing—in routine practice provides new and exciting challenges that clinical laboratories are well positioned to meet.

REVIEW QUESTIONS

1. Which of the following best describes appropriate use for most serum tumor marker measurements?
   a. Diagnosing or excluding a diagnosis of malignancy
   b. Providing definitive information about likely prognosis in individual patients
   c. Monitoring treatment effectiveness and the course of disease
   d. Identifying the primary cancer by testing panels of serum tumor markers
   e. Testing that is essential in the follow-up of all cancer patients
2. For which of the following has measurement of a serum tumor marker been unequivocally demonstrated to be essential for early detection of disease in the appropriate screening population?
   a. Breast cancer
   b. Prostate cancer
   c. Ovarian cancer
   d. Malignant melanoma
   e. Gestational trophoblastic disease
3. Which of the following are mandatory for the clinical management of the majority of breast cancer patients?
   a. ERs
   b. CA 15-3
   c. CEA
   d. uPA
   e. Gene profiling
4. Age-related reference intervals for most tumor markers are either not relevant and/or are not universally applied for most clinical applications, but are mandatory for which of the following age groups?
   a. AFP in men older than 30 years
   b. AFP in neonates and infants younger than 12 months
   c. hCG in men older than 30 years
   d. hCG in neonates and infants younger than 12 months
   e. CEA in men older than 30 years
5. In women with suspected ovarian cancer the RMI is used primarily for which of the following?
   a. To exclude the need for pelvic ultrasound
   b. To identify whether a colorectal surgeon should be present at the time of operation

c. To identify women with stage I disease where clear cell and endometrioid histologies predominate
d. To assist in triaging women with ovarian cysts
e. To determine menopausal status

6. Which of the following statements best describes the contribution of IS and reference reagents to tumor marker measurements?
   a. They ensure that results are reported worldwide in the same units of measurement
   b. They can be used as kit standards by manufacturers
   c. Their introduction guarantees improved between-method agreement
   d. They enable accurate calibration of all the most widely used serum tumor markers
   e. They provide a benchmark against which the accuracy of calibration can be assessed

SUGGESTED READINGS

American Cancer Society. (2017). *Cancer Facts and Figures*. https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2017.html.

Dalton-Fitzgerald, E., Tiro, J., Kandunoori, P., Halm, E. A., Yopp, A., & Singal, A. G. (2015). Practice patterns and attitudes of primary care providers and barriers to surveillance of hepatocellular carcinoma in patients with cirrhosis. *Clinical Gastroenterology and Hepatology*, *13*, 791–798 e1.

Duffy, M. J. (2014). PSA in screening for prostate cancer: more good than harm or more harm than good? *Advances in Clinical Chemistry*, *66*, 1–23.

Duffy, M. J., & Crown, J. (2014). Precision treatment for cancer: role of prognostic and predictive markers. *Critical Reviews in Clinical Laboratory Sciences*, *51*, 30–45.

Duffy, M. J., Lamerz, R., Haglund, C., Nicolini, A., Kalousova, M., Holubec, L., et al. (2014). Tumor markers in colorectal cancer, gastric cancer and gastrointestinal stromal cancers: European Group on Tumor Markers 2014 guidelines update. *International Journal of Cancer*, *134*, 2513–2522.

Duffy, M. J., Sturgeon, C. M., Sölétormos, G., Barak, V., Molina, R., Hayes, D. F., et al. (2015). Validation of new cancer biomarkers: a position statement from the European Group on Tumor Markers. *Clinical Chemistry*, *61*, 809–820.

International Germ Cell Consensus Classification. (1997). A prognostic factor-based staging system for metastatic germ cell cancers. International Germ Cell Cancer Collaborative Group. *Japan Society of Clinical Oncology*, *15*, 594–603.

Jassam, N., Jones, C. M., Briscoe, T., & Horner, J. H. (2006). The hook effect: a need for constant vigilance. *Annals of Clinical Biochemistry*, *43*, 314–317.

NACB. (2008). National Academy of Clinical Biochemistry Laboratory Medicine Practice Guidelines for use of tumor markers. *Clinical Chemistry*, *54*, e1–e10.

NCCN. National Comprehensive Cancer Network Clinical Practice Guidelines in Oncology. https://www.nccn.org/professionals/physician_gls/f_guidelines.asp.

Sturgeon, C. M., Lai, L. C., & Duffy, M. J. (2009). Serum tumour markers: how to order and interpret them. *British Medical Journal*, *339*, b3527.

Sturgeon, C. M., & Viljoen, A. (2011). Analytical error and interference in immunoassay: minimizing risk. *Annals of Clinical Biochemistry*, *48*, 418–432.

Kidney Function Tests

*Edmund J. Lamb, Graham R.D. Jones**

OBJECTIVES

1. Define the following:
   - Creatinine
   - Urea
   - Uric acid
   - Gout
2. Diagram the pathway that results in the formation of creatinine.
3. State the clinical usefulness of measuring serum and urine creatinine.
4. State the principle of the Jaffe reaction and list five interferences that affect this method.
5. List and describe three enzymatic assays used to measure creatinine in serum.
6. Describe "compensated assays" in relation to the measurement of creatinine.
7. Diagram the catabolic pathway that results in the formation of urea.
8. State the clinical usefulness of measuring serum urea; list the causes of increased and decreased serum urea.
9. Describe the urease approach for measuring urea in serum; list one interference that affects this method.
10. Convert serum urea nitrogen given in milligrams per deciliter to urea nitrogen in millimoles per liter.
11. Diagram the pathway that results in the formation of uric acid.
12. State the clinical usefulness of measuring serum uric acid.
13. List the causes of hyperuricemia; explain the pathogenesis of gout and urinary tract uric acid stones.
14. Compare and contrast primary and secondary gout.
15. List the causes of hypouricemia.
16. Describe the uricase method of measuring uric acid in serum; list two interferences that affect this method.
17. List and describe the proteins found in urine, including their characteristics, how they are affected in disease, and the laboratory methods of protein measurement.

KEYWORDS AND DEFINITIONS

BUN (blood urea nitrogen) Obsolete term used to report results of a urea assay, particularly in the United States.

Creatinine A nonprotein nitrogen compound derived from the spontaneous hydrolysis of creatine or the cyclization of phosphocreatine; creatinine production is relatively constant, is related to muscle mass, and is used as a marker of the glomerular filtration rate of the kidneys.

Gout A group of disorders of purine metabolism; can be due to primary (inherited) or secondary causes such as chronic kidney disease.

Hyperuricemia An excess of uric acid or urates in the blood with many causes; it is a prerequisite for the development of gout and may lead to renal disease.

Hypouricemia Decreased uric acid concentration in the blood, secondary to a number of underlying conditions such as severe hepatocellular disease and defective renal tubular reabsorption.

Jaffe reaction The reaction of creatinine with alkaline picrate to form a colored compound; this creatinine assay is subject to numerous interferences.

Urea The major nitrogen-containing metabolic product of protein catabolism in humans.

Urease methods Enzymatic assays that initially involve the hydrolysis of urea by urease to generate ammonia, which is quantified by a variety of methods.

Uric acid A nitrogenous compound derived from the catabolism of purine nucleosides.

*Data on current creatinine method usage in the United Kingdom were provided by Finlay Mackenzie from the United Kingdom National External Quality Assessment Scheme (UKNEQAS), Birmingham, United Kingdom. Further information about the UKNEQAS estimated glomerular filtration rate (GFR) scheme may be found at <https://birminghamquality.org.uk/>. We are grateful to the College of American Pathologists and the Royal College of Pathologists of Australasia Quality Assurance Program for data on current creatinine method usage and performance in the United States and Australasia, respectively. We also acknowledge the input of Professor Christopher Price to previous editions of this chapter.

Uricase methods A group of enzymatic assays that initially involve oxidation of uric acid by uricase to eventually produce a chromogen that is spectrophotometrically measured to determine uric acid concentration.

Creatinine, **urea**, and **uric acid** are nonprotein nitrogenous metabolites that are cleared from the body by the kidney following glomerular filtration. Measurements of plasma or serum concentrations of these metabolites are commonly used as indicators of kidney function and other conditions. In particular, creatinine is commonly used in equations to estimate the glomerular filtration rate (GFR). Small molecular weight proteins such as cystatin C are also increasingly being used to estimate the GFR. The measurement of total protein and/or albumin in urine can give useful information regarding the integrity of the glomerular filter and tubular function.

PROTEINURIA: TOTAL PROTEIN AND ALBUMIN

Higher molecular weight proteins are retained within the circulation by the glomerular filter, and lower molecular weight proteins are freely filtered and reabsorbed and catabolized within the tubular cells. Consequently, the appearance of notable amounts of protein in the urine suggests renal disease of either the glomerulus, the tubules or both. The association between kidney disease and proteinuria dates as far back to at least the early 19th century, when Bright first described albuminous nephritis. Proteinuria can be classified as either tubular or glomerular, depending on the pattern of proteinuria observed. A third category, overflow proteinuria, is also recognized in which filtration of excessive amounts of low-molecular-weight protein exceeds the tubular capacity for reabsorption. Examples of the latter include Bence Jones proteinuria and myoglobinuria. Proteinuria is a potent risk marker for progressive kidney disease, and reduction of protein excretion is a therapeutic target. Proteinuria may be detected and measured using reagent strip devices or laboratory measurements of either total protein or albumin.

Sample Collection for Urinary Total Protein and Albumin Measurement

It is generally recognized that a 24-hour sample is the definitive means of demonstrating and quantifying the presence of proteinuria but this is a difficult procedure to control effectively, and inaccuracies in urinary collection may contribute to errors in estimation of protein losses. Overnight, first void in the morning (early morning urine [EMU]), second void in the morning, or random sample collections have also been used. Because creatinine excretion in the urine is fairly constant throughout the 24-hour period, measurement of the albumin-to-creatinine ratio (ACR) or protein-to-creatinine ratio (PCR) allows the use of a spot sample, with correction for variations in urinary concentration in an individual. Spot samples for ACR or PCR measurement generally have good diagnostic performance and correlation with the 24-hour collection, and are widely recommended for routine clinical use. An EMU sample is often preferred because it is unaffected by orthostatic (postural) proteinuria.

The diagnosis of albuminuria requires the demonstration of increased albumin loss (either increased ACR or increased albumin in a timed collection) in at least two out of three urine samples collected in the absence of infection or an acute medical illness (Fig. 21.1). Establishing the diagnosis of albuminuria or proteinuria has both prognostic and management implications. In the setting of diabetes, the best possible metabolic control should be achieved before patients are examined for albuminuria, and patients should not be screened during other transitory illnesses. Screening should commence 5 years after diagnosis in patients with type 1 diabetes mellitus and at diagnosis in patients with type 2 diabetes, and should continue on an annual basis up to the age of 75 years. Patients demonstrating an ACR of 3.0 mg/mmol or greater should have urine samples sent to the laboratory on two other occasions (ideally within 2 months) for albumin estimation. Patients demonstrating increased ACRs in one or both of these additional samples are said to have persistent albuminuria. Diabetic nephropathy is uncommon in patients who have had type 1 diabetes for less than 5 years, and other causes of kidney disease should be considered.

Samples for urinary albumin (or total protein) measurement may be analyzed fresh, stored at 4°C for up to 1 week, or stored at −70°C for longer periods. Freezing at −20°C appears to result in loss of measurable albumin and is not recommended. For analysis, stored samples should be allowed to reach room temperature and be thoroughly mixed prior to testing.

Analytical Methods and Traceability: Urinary Total Protein

Numerous methods can be used for the measurement of protein in urine, including the original Lowry method, turbidimetry after mixing with trichloroacetic or sulfosalicylic acid, turbidimetry with benzethonium chloride (benzyl dimethyl {2-[2-(*p*-1,1,3,3-tetramethyl butylphenoxy)ethoxy] ethyl} ammonium chloride), and dye binding with Coomassie Brilliant Blue, pyrogallol red molybdate, and pyrocatecholviolet-molybdate, which is used in dry-slide applications.

Total protein measurement is more difficult in urine than in serum. The concentration of urinary protein is normally low (100 to 200 mg/L) and large sample-to-sample variation in the amount and composition of proteins is common. The concentration of nonprotein potentially interfering substances is high relative to the protein concentration and is highly variable, and the inorganic ion content is high. All these factors affect the precision and accuracy of the various methods.

Because of the need for automation, the benzethonium chloride and dye-binding methods have become the most popular in current clinical use but they do not provide equal analytical specificity and sensitivity for all proteins. Concerns about different

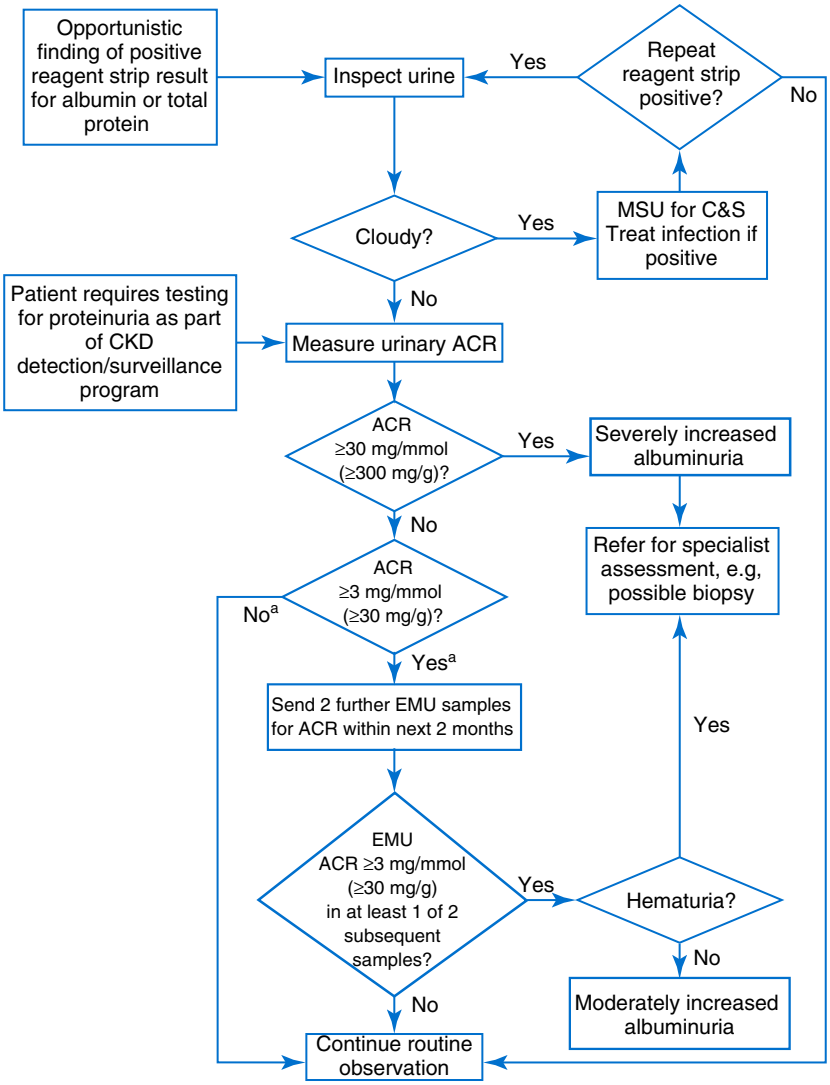


Fig. 21.1 Suggested protocol for the further investigation of an individual demonstrating a positive reagent strip test for albuminuria/proteinuria or quantitative albuminuria/proteinuria test. Reagent strip device results should be confirmed using laboratory testing of the albumin-to-creatinine ratio *(ACR)* on at least two further occasions. Patients with two or more positive (≥3 mg albumin/mmol creatinine) tests on early morning samples 1 to 2 weeks apart should be diagnosed as having persistent albuminuria. (The possibility of postural proteinuria should be excluded by the examination of an early morning urine.) Protein-to-creatinine ratio (PCR) measurement can be substituted for the ACR but is insensitive in the detection of mild to moderately increased albuminuria/proteinuria: Approximate PCR equivalent to an ACR of 30 mg/mmol is 50 mg/mmol. [a]Consider other causes of increased ACR (e.g., menstrual contamination, uncontrolled hypertension, symptomatic urinary tract infection, heart failure, other transitory illnesses, and strenuous exercise), especially in the case of type 1 diabetes present for less than 5 years. The presence of hematuria may indicate nondiabetic renal disease. *Note:* US guidelines express albuminuria or proteinuria as mg/g creatinine, whereas other guidelines use mg/mmol creatinine. An approximate conversion factor of 0.1136 can be used to convert results in mg/g to mg/mmol. However, for clarity and pragmatism, recent guidelines have accepted decision points that are approximately equivalent: hence, when using this protocol in the United States, 300 mg/g should be substituted for 30 mg/mmol and 30 mg/g for 3 mg/mmol. *C&S,* Culture and sensitivity; *CKD,* chronic kidney disease; *EMU,* early morning urine; *MSU,* midstream urine.

responses to different proteins have led to many variants of the published methods. Most methods tend to react more strongly with albumin than with globulin and other nonalbumin proteins. As different methods are responding to different properties of proteins and the abundance of these properties varies between the proteins found in urine, the "quantity intended to be measured" is not the same for different assays. For those reasons, there is no reference measurement procedure or standardized reference material for urinary total protein. The variety of methods in use means that significant between-laboratory variation is inevitable. This variation tends to diminish at higher concentrations of urinary total protein, presumably in part as albumin becomes the predominant protein and thus reduces a component of the between-sample variation. These methodological problems are a major reason why urine albumin is a preferred analyte for the assessment of proteinuria.

TABLE 21.1 Characteristics of Some Clinically Important Urinary Proteins

Protein	Mr (kDa)	Free Plasma Concentration (g/L)	Diameter (nm)	pI	Glomerular Sieving Coefficient	Filtered Load >(mg/L)[a]	Urinary Concentration (mg/L)[b]	% Reabsorbed
IgG	150	10	5.5	7.3	0.0001	1	0.1	99
Albumin	66	40	3.5	4.7	0.0002	8	5	99
α_1-Microglobulin	31	0.025	2.9	4.5	~0.3	7.5	5	99
Retinol-binding protein	22	0.025	2.1	4.5	~0.7	17.5	0.1	99
Cystatin C	12.8	0.001	3.0	9.2	~0.7	0.7	0.1	99
β_2-Microglobulin	11.8	0.002	1.6	5.6	0.7	1.1	0.1	99

IgG, Immunoglobulin G.
[a]Concentration in the glomerular filtrate.
[b]Typical concentrations observed in health.

Analytical Methods and Traceability: Urinary Albumin

Urinary albumin has been measured using immunoassay since the 1960s when the first such assays became available. Urinary albumin is predominantly measured using quantitative immunoturbidimetric or immunonephelometric approaches capable of detecting albumin at low concentrations. No Joint Committee for Traceability in Laboratory Medicine (JCTLM) listed reference measurement procedure or higher order reference material is currently available for urinary albumin. To date, most urinary albumin assays have been standardized against a serum-based calibrant (ERM-DA-470k/IFCC) distributed by the Institute for Reference Materials and Measurements of the European Commission.

Reference Intervals, Definitions of *Proteinuria* and *Albuminuria*

There is no consistent definition of *proteinuria*. The upper limit of the reference interval[†] for urinary total protein loss varies between 150 and 300 mg/day, depending on the laboratory. Given average daily creatinine excretion of about 10 mmol (0.11 g), an upper limit of normal protein loss of 150 mg/day is equivalent to a urinary PCR of approximately 15 mg/mmol (130 mg/g). The protein in the urine of healthy individuals is made up of albumin (<30 mg/day) and some smaller proteins, together with proteins secreted by the tubules, of which Tamm-Horsfall glycoprotein (THG) predominates. Typical concentrations of proteins found in urine are listed in Table 21.1. Readers should note that the units of expression used for ACR and PCR differ depending on geographical location, typically whether it is inside or outside the United States.[‡]

Proteinuria is often detected at the point of care (POC) using urine reagent strip devices, and clinical proteinuria has sometimes been defined as equivalent to a color change of "+" or greater on the relevant pad on the strip. This equates to approximately 300 mg/L of total protein or a PCR of 50 mg/mmol, or protein loss of approximately 500 mg/day (assuming an average urine volume of 1.5 L/day). Indeed, the limits for proteinuria and albuminuria are best described as clinical decision points because they are generally described or defined by expert groups rather than as the product of formal reference interval studies.

In health, relatively small amounts of albumin (<30 mg/day) are lost in the urine. In the international classification of kidney disease (see Chapter 35), proteinuria is categorized based on levels of albumin loss: A1 normal to mildly increased, less than 3.0 mg/mmol (approximately equivalent to <30 mg/g); A2 moderately increased, 3 to 30 mg/mmol (approximately equivalent to 30 to 300 mg/g); and A3 severely increased, greater than 30 mg/mmol (approximately equivalent to ≥300 mg/g). These categories are approximately equivalent to what would have formerly been considered normoalbuminuria, microalbuminuria, and macroalbuminuria (sometimes referred to as "clinical" or "significant" proteinuria), respectively. *Microalbuminuria* is a term that has been widely used to describe an increase in urinary albumin loss above the reference interval for healthy nondiabetic subjects, but at a level that is not generally detectable by less sensitive clinical tests such as reagent strips designed to measure total protein. The term *microalbuminuria* is somewhat misleading in that the albumin being measured is identical in form to that circulating in plasma, and the so-called microalbuminuric range refers to increased, not "micro-," albumin losses. Current guidelines do not support the continuing use of this term, and the term *albuminuria* is used in this chapter.

Clinical Significance

The physiology and pathophysiology of renal protein handling and the clinical significance of proteinuria in specific clinical situations are discussed in more detail in Chapter 35; more general considerations follow. While reagent strip tests are commonly used in clinical practice, false negative and positive results are common. Most authors agree that positive tests require confirmation by laboratory measurement of the PCR or ACR on an early morning or random urine sample (see Fig. 21.1).

[†]Laboratories should verify that these ranges are appropriate for use in their own settings.
[‡]US guidelines express albuminuria or proteinuria as mg/g creatinine, whereas other guidelines use mg/mmol creatinine. A conversion factor of 0.1136 can be used to convert ACR or PCR results in mg/g to mg/mmol (e.g., 200 mg/g = 23 mg/mmol, 30 mg/g = 3.4 mg/mmol).

Most commonly, proteinuria reflects albuminuria and it is suggested that urinary total protein measurement can be replaced by urine albumin measurement. Strong evidence has linked increased urinary albumin loss to mortality and morbidity, including kidney disease progression, in people both with and without diabetes: albuminuria is generally held to be a more sensitive indicator of kidney disease than increased total protein loss. However, increases in albumin loss may also reflect overall changes in vascular permeability, and therefore may not indicate explicit deterioration in renal function. Increased urinary albumin loss is common among the general population and is not solely attributable to the presence of diabetes.

POINTS TO REMEMBER

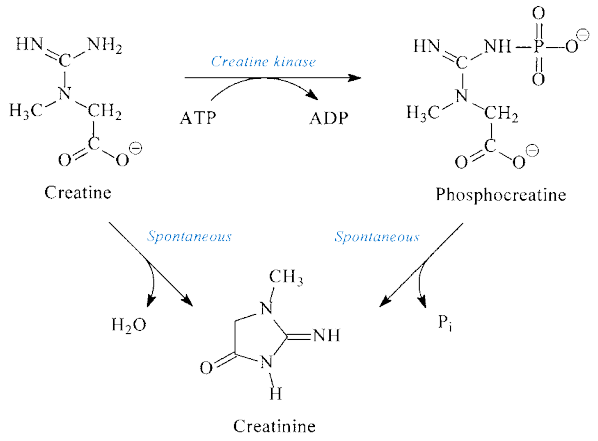
Urinary Albumin

- Albumin (Mr 66 kDa) is predominantly retained in the circulation by the glomerular basement membrane due to its size and negative charge.
- In health, only small amounts of albumin are filtered. This is largely reabsorbed in the proximal tubule.
- Albuminuria may arise due to both glomerular (increased filtration) and tubular (decreased resorption) disease.
- Albumin is the predominant urinary protein in the setting of proteinuria.
- Due to high biological variability and nonrenal influences, albuminuria should be confirmed on at least two occasions, preferably with first morning samples.
- Measurement of the ACR allows the use of a spot sample with correction for variations in urinary concentration in an individual.

CREATININE

Biochemistry and Physiology

Creatine, the immediate precursor of creatinine, is synthesized in the kidneys, liver, and pancreas and is then transported in blood to other organs, such as muscle and brain, where it is phosphorylated to phosphocreatine, a high-energy compound.



Interconversion of phosphocreatine and creatine is a particular feature of the metabolic processes of muscle contraction. A proportion of free creatine in muscle (thought to be between 1% and 2%/day) spontaneously and irreversibly converts to its anhydride waste product, creatinine. Thus the amount of creatinine produced each day in an individual is fairly constant and is related to the muscle mass. In health, the blood concentration of creatinine is also fairly constant although it may be influenced by diet (see later). Creatinine (Mr approximately 113 Da) is present in all body fluids and secretions and is freely filtered by the glomerulus. Although it is not reabsorbed to any great extent by the renal tubules, a small but notable tubular secretion is present, as well as concentration-related losses in the gut.

Sample Collection

Creatinine in serum, plasma, or urine is stable for at least 7 days at 4°C, and serum creatinine is stable during long-term frozen storage (at −20°C and below) and after repeated thawing and refreezing. However, it should be noted that delayed separation (beyond 14 hours) of serum from erythrocytes leads to a significant increase in apparent serum creatinine concentration using some kinetic Jaffe (but not enzymatic) assays, possibly as the result of release of noncreatinine chromogens from the red cells (see later). Creatinine concentration increases in blood after meals containing cooked meat, due to the conversion of creatine to creatinine, and ideally blood for serum creatinine measurement should be obtained in the fasting state.

Analytical Methods

Serum creatinine is measured in virtually all clinical laboratories as a test of kidney function. Most laboratories use adaptations of the same assay for measurements in both serum and urine. Both chemical and enzymatic methods are used to measure creatinine in body fluids.

Most chemical methods for measuring creatinine are based on its reaction with alkaline picrate. As first described by Jaffe in 1886, creatinine reacts with picrate ion in an alkaline medium to yield an orange-red complex.

A serious analytical problem with the **Jaffe reaction** is its lack of specificity for creatinine. For example, many compounds have been reported to produce a Jaffe-like chromogen, including (1) ascorbic acid, (2) blood-substitute products, (3) cephalosporins, (4) glucose, (5) guanidine, (6) ketone bodies, (7) protein, and (8) pyruvate. The degree of interference from these compounds is dependent on the specific reaction conditions chosen. The effect of ketones and ketoacids are probably of the greatest significance clinically, although the effect is very method dependent. Thus reports on acetoacetate interference vary from a negligible increase to an increase of 3.5 mg/dL (310 μmol/L) in the apparent creatinine concentration at an acetoacetate concentration of 8 mmol/L. Bilirubin is a negative interferant with the Jaffe reaction. The addition of buffering ions, such as borate and phosphate, together with surfactant, has been used to minimize the effects of this interference. In addition, ferricyanide—O'Leary method—has been added that oxidizes bilirubin to biliverdin, hence reducing its interference. For rate assays a rate-blanking approach has been successful. Noncreatinine chromogens do not generally contribute to measured urinary creatinine concentration.

The greatest success in terms of common usage and specificity has come from the use of a kinetic measurement approach in combination with the careful choice of reactant concentrations. In general, manual methods have traditionally been equilibrium methods, with 10 to 15 minutes allowed for color development at room temperature. Kinetic assays have been developed to provide more specific, faster, and automated analyses. Early studies of interferences in the kinetic methods identified two kinds of noncreatinine chromogens. In one group, the rate of adduct formation is very rapid and occurs in the first 20 s after mixing reagent and sample. Acetoacetate is an example of this type of interferant. In the second group, the rate of adduct formation does not become significant until 80 to 100 s after mixing (e.g., protein). The "window" between 20 and 80 s therefore was a period in which the rate signal being observed could be attributed predominantly to the creatinine-picrate reaction. Thus improvement of specificity in the kinetic assays was achieved by selecting times for rate measurements 20 to 80 s after initiation of the reaction (mixing). This approach has been implemented on various automated instruments, and kinetic assays are now widely used to measure creatinine concentrations in body fluids.

Extensive literature exists on the choice of reactant concentrations and reading interval, and on the choice of wavelength and reaction temperature. Brief comments follow.

Picrate Concentration

The Jaffe reaction is pseudo first order with respect to picrate up to 30 mmol/L, with the majority of methods employing a concentration between 3 and 16 mmol/L. At concentrations above 6 mmol/L, the rate of color development becomes nonlinear, so a two-point fixed interval rather than a multiple data point approach is required.

Hydroxide Concentration

The initial rate of reaction is pseudo first order with respect to hydroxide concentrations above 0.5 mmol/L. However, at 500 mmol/L there is an increased degradation of the Jaffe complex. Furthermore, at hydroxide concentrations above 200 mmol/L, the blank absorbance increases significantly.

Wavelength

Although the absorbance maximum of the Jaffe reaction is between 490 and 500 nm, improved method linearity and reduced blank values have been reported at other wavelengths, the choice varying with hydroxide concentration.

Temperature

The rate of Jaffe complex formation and the absorptivity of the complex are temperature dependent, measurable differences being observed even between 25°C and 37°C. Consequently, temperature control is an important component of assay reproducibility.

Compensation

As a result of reaction with noncreatinine chromogens, Jaffe methods often have historically overestimated true serum creatinine concentrations by up to 20% compared with high-performance liquid chromatography (HPLC) or isotope dilution mass spectrometry (IDMS) methods, at physiologic concentrations. In an attempt to adjust for this, some manufacturers have introduced so-called "compensated" Jaffe assays, in which a fixed concentration is automatically subtracted from each result. For example, Roche Diagnostics Ltd. (Lewes, Sussex, United Kingdom) has realigned its assays on the Cobas Integra and Hitachi systems by −0.20 mg/dL and −0.32 mg/dL (−18 and −28 μmol/L), respectively. Such assays produce lower results more closely aligned with IDMS reference measurement procedures at concentrations within the reference interval. However, they make an assumption that the noncreatinine chromogen interference is a constant between samples; this is clearly an oversimplification, especially when adult and pediatric samples are compared.

Enzymatic Methods

Enzymes from a number of metabolic pathways have been investigated for the enzymatic measurement of creatinine. All of the methods involve a multistep approach leading to a photometric equilibrium (Fig. 21.2). There are primarily three approaches, described below.

Creatininase

Creatininase (EC 3.5.2.10; creatinine amidohydrolase) catalyzes the conversion of creatinine to creatine. The creatine is then detected with a series of enzyme-mediated reactions involving creatine kinase, pyruvate kinase, and lactate dehydrogenase, with monitoring of the decrease in absorbance at 340 nm (see Fig. 21.2A). Initiating the reaction with creatininase allows for the removal of endogenous creatine and pyruvate in a preincubation reaction. The kinetics of the reaction are analytically problematic and a 30-minute incubation is required to allow the reaction to reach equilibrium. This shortcoming has been overcome by a kinetic approach but with a further reduction in the method's ability to detect creatinine. Consequently, this approach is not widely used.

Creatininase and Creatinase

An alternative approach has been the use of creatinase (EC 3.5.3.3; creatine amidinohydrolase) that yields sarcosine and urea, the former being measured with further enzyme-mediated steps using sarcosine oxidase (EC 1.5.3.1). This produces (1) glycine, (2) formaldehyde, and (3) hydrogen peroxide (see Fig. 21.2B) with the latter being detected and measured with a variety of methods. Care must be taken, however, because of interference (e.g., by bilirubin) in the final reaction sequence. This problem has been minimized by adding potassium ferricyanide (with limited success) or bilirubin oxidase. The potential interference caused by ascorbic acid has been overcome by the inclusion of ascorbate oxidase (L-ascorbate:oxygen oxidoreductase; EC 1.10.3.3). The influence of endogenous intermediate creatine and urea has been minimized by adding a

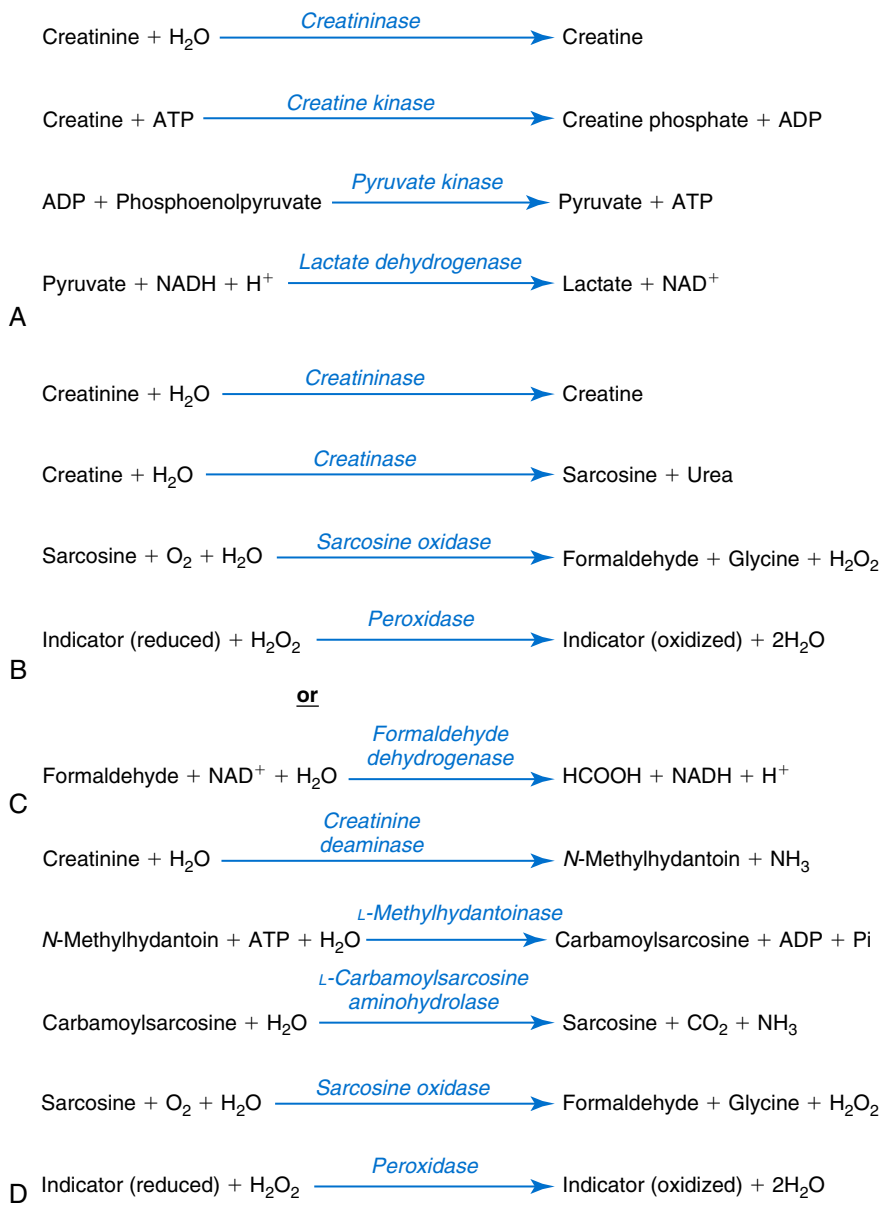


Fig. 21.2 (A–D) Determination of creatinine using a variety of enzymatic methods. For additional details, see text. *ADP*, Adenosine diphosphate; *ATP*, adenosine triphosphate; *NAD*, nicotinamide adenine dinucleotide.

preincubation step and then initiating the reaction with creatininase. This system has been incorporated in a POC testing device using polarographic detection. An alternative detection system involves measurement of the reduction of nicotinamide adenine dinucleotide (NAD) by formaldehyde in the presence of formaldehyde dehydrogenase (see Fig. 21.2C).

Creatinine Deaminase

Creatinine deaminase (EC 3.5.4.21; creatinine imino-hydrolase) catalyzes the conversion of creatinine to *N*-methylhydantoin and ammonia. Early methods concentrated on the detection of ammonia using either glutamate dehydrogenase or the Berthelot reaction. An alternative approach involves the enzyme *N*-methylhydantoin amidohydrolase (see Fig. 21.2D).

Dry Chemistry Systems

A number of multilayer dry reagent methods have been described for the measurement of creatinine using enzyme-mediated reactions. An early "two-slide" approach employed creatinine deaminase, with the ammonia diffusing through a semipermeable and optically opaque layer to react with bromophenol blue to give an increase in absorbance at 600 nm. A second multilayer film lacking the enzyme was used to quantitate endogenous ammonia, enabling blank correction. A later single-slide method used the creatininase-creatinase reaction sequence. Lidocaine metabolites have been reported to interfere with this method. The creatinine deaminase system described above has also been used and adapted for use as a POC testing device (see Fig. 21.2D). In all cases, the color produced in the film is quantified by reflectance spectrophotometry.

Other Methods

A definitive method employing IDMS was described by Welch in 1986. Gas or liquid chromatography-IDMS (GC-IDMS, LC-IDMS) are now accepted as the methods of choice for establishing the true concentration of creatinine in serum because of its excellent specificity and low imprecision. GC- and LC-IDMS methods have been approved by the JCTLM as reference measurement procedures for serum creatinine.

Quality Issues and Preanalytical Considerations With Creatinine Methods

Assessment of the methods used for the measurement of creatinine is complex by virtue of the number of variants of the Jaffe reaction and the innovations attempted using enzymatic procedures to overcome the limitations of the former. Although enzymatic methods are more expensive, they are used in dry chemistry systems (with their lower reagent requirement), including some POC testing devices. Kinetic Jaffe approaches generally predominate in wet chemistry analytical systems. Any laboratorian assessing a new creatinine method (e.g., as part of an analyzer purchase) should review the data for that method on common interferences. Despite criticism of the Jaffe methods, good correlation has been noted invariably between them and enzymatic procedures, with differences likely to be due as much to calibration as to interference.

Different methods for assaying serum creatinine have varying degrees of accuracy and imprecision. Mean within-individual biological variation for serum creatinine has been reported as approximately 6.0%, indicating a desirable analytical performance goal of less than 3.0%. Intralaboratory imprecision at a concentration of 88 μmol/L varies between approximately 2.0 and 8.4%. Clearly, many laboratories are therefore outside desirable and even minimum performance standards. Proficiency studies demonstrate that while between-laboratory CVs of approximately 3% are achievable within method groups, overall between-laboratory agreement across methods is much poorer. Further, within- and between-laboratory agreement deteriorates as serum creatinine concentration falls to within and below the reference interval; the exponential relationship between serum creatinine and GFR means that imprecision at lower creatinine concentrations contributes to greater error in GFR estimation than at higher creatinine concentrations.

Over the past 20 years, appreciation of chronic kidney disease (CKD) as a major public health issue and of its identification, staging and monitoring using GFR-estimating equations has led to increased focus on the measurement of creatinine. Creatinine-based estimates of GFR (see later) will clearly vary, depending on how accurate the creatinine measurement used in the calculation is. The more a method overestimates "true" creatinine, the greater will be the underestimation of GFR, and vice versa (see later).

Traceability of Serum Creatinine Measurement

Standardized serum matrix reference materials (SRM 967) with known creatinine concentrations (0.80 mg/dL [71 μmol/L] and 4.00 mg/dL [354 μmol/L]) have been prepared by the National Institute of Standards of Technology (NIST) and included in a list of higher order reference materials by the JCTLM. The material was value-assigned using mass spectrometry and issued in 2007. This material, in combination with GC-IDMS reference methodology, was used by reagent manufacturers to restandardize their methods and since 2009 most clinical laboratory methods have had calibration traceable to the reference measurement procedure and standard.

Although undoubtedly desirable, it must be recognized that standardization is only one arm of the problem. Standardization will not solve the problems of noncommutable calibrator materials and differential reactivity with noncreatinine chromogens across different patient samples, which can be resolved only by the use of highly specific creatinine methods. Wider adoption of enzymatic methods could further improve between-laboratory agreement in creatinine measurement, but these methods are used by a minority of laboratories at present.

Reference Intervals

Reference intervals for serum creatinine are method dependent. A systematic review of creatinine reference intervals in which studies were included only when, in particular, their calibration was traceable to the reference IDMS procedure; proposed adult reference intervals of 0.72 to 1.18 mg/dL (64 to 104 μmol/L) in men and 0.55 to 1.02 mg/dL (49 to 90 μmol/L) in women. These data were derived using an enzymatic (Roche Diagnostics Ltd) assay. Serum creatinine concentration in patients with untreated end-stage renal disease (ESRD) may exceed 11 mg/dL (1000 μmol/L).

Urinary creatinine excretion is higher in men (14 to 26 mg/kg per day, 124 to 230 μmol/kg per day) than in women (11 to 20 mg/kg per day, 97 to 177 μmol/kg per day). Creatinine excretion decreases with age. Typically, for a 70 kg man, creatinine excretion will decline from approximately 1640 to 1030 mg/day (14.5 to 9.1 mmol/day) with advancing age from 30 to 80 years. The measurement of urinary creatinine excretion has been found to be a useful indication of the completeness of a timed urine collection. Creatinine excretion is often used as a method of normalizing the urinary excretion of analytes, that is, the excretion of the test analyte (in millimoles or grams) is divided by the total amount of creatinine (in millimoles or grams) excreted in the same urine specimen. This method is a rough correction for volume differences between patient specimens. Similarly, expressing the concentration of a substance as a ratio to the creatinine concentration is a useful method of adjusting for urinary concentration differences in random ("spot") urine samples.

Clinical Significance

Serum creatinine concentration is maintained within narrow limits predominantly by glomerular filtration. Consequently, both serum creatinine concentration and its renal clearance ("creatinine clearance") have been used as markers of the *GFR*. The application and limitations of these tests are discussed later in this chapter.

POINTS TO REMEMBER

Creatinine

- Creatinine is produced at a fairly constant rate within an individual as a result of a breakdown of creatine within muscle tissue.
- Creatinine is freely filtered at the glomerulus and also secreted in the tubules to a lesser extent.
- Because it is predominantly excreted by the kidneys, plasma creatinine concentration is inversely proportional to GFR (halving the GFR approximately doubles the serum creatinine).
- In addition to GFR, other factors related to creatinine production affect plasma creatinine concentration, including age, gender, race, muscularity, certain drugs, diet, and nutritional status.
- Estimated GFR (eGFR) using the CKD-EPI$_{creat}$ equation is currently recommended for clinical use (see Table 21.2).

CYSTATIN C

Biochemistry and Physiology

Cystatin C is a low-molecular-weight (12.8 kDa) protein synthesized by all nucleated cells whose physiologic role is that of a cysteine protease inhibitor. The gene has been sequenced, and the promoter region has been identified as that of the housekeeping type, with no known regulatory elements, and, consequently, the rate of cystatin C release into the circulation was initially considered to be constant. Over time a number of factors have been identified that can affect cystatin C production and thus the circulating concentration (e.g., thyroid hormone concentration). Unlike creatinine, muscle mass does not have a major effect on cystatin C concentration.

Cystatin C is removed from the circulation by the kidneys with no known extrarenal routes of elimination. Due to its small size and high isoelectric point (pI 9.2), it is more freely filtered than some other putative protein markers of GFR. Following filtration, cystatin C is essentially fully resorbed and broken down in the tubules usually leading to very low concentrations in the urine.

In addition to renal function and pathophysiological factors, serum cystatin C concentrations are also affected by age, gender, pregnancy, weight, and height. The influences of age and gender are taken into account with some cystatin C-based GFR estimating equations (see later).

Analytical Methods

Cystatin C is measured by immunoassay, the most practical approaches being particle-enhanced turbidimetric or nephelometric immunoassay, which can run on automated chemistry analyzers or specific nephelometers. Using these assays, a between-day imprecision of CV 3% to 6% can be expected at the upper limit of the reference interval (≅1.00 mg/L) and less than 3% at higher values. The use of immunoassays makes cystatin C relatively free of the interferences that affect Jaffe creatinine assays although the reagents are markedly more expensive than those used for creatinine measurement.

Traceability of Cystatin C Measurement

An international standard, ERM-DA471/IFCC Cystatin C in human serum, has been developed by an International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) working party. The material has verified commutability and is listed on the JCTLM database. Most manufacturers have restandardized their assays against this material, and the use of these traceable assays in research and in the routine laboratory is recommended to allow direct comparability of results.

Reference Intervals

While single reference intervals for adult males and females of all ages are commonly used, typically of the order 0.6 to 1.1 mg/L, it is known that cystatin C concentrations are higher in males than females and climb throughout adult life, reflecting the fall in GFR with advancing age seen in most populations. In a similar manner to serum creatinine, clinical decisions based on cystatin C are more likely to be made using eGFR values derived from cystatin C than on the serum cystatin C concentration itself, making the reference interval less relevant as a decision support tool.

Clinical Significance

Because cystatin C is produced at a relatively constant rate, freely filtered at the glomerulus and neither secreted nor reabsorbed intact in the proximal tubule, its concentration in serum can serve as a marker of GFR, with GFR being related to the reciprocal of serum cystatin C. In a similar manner to serum creatinine, equations have been developed to estimate GFR based on serum cystatin C with or without added demographic variables and with or without serum creatinine; some of these equations are given in Table 21.2. The Chronic Kidney Disease-Epidemiology Collaboration (CKD-EPI) cystatin C equations are well validated in adults.

Generally the use of cystatin C equations provides some, albeit modest, improvement over equations based on serum creatinine alone. An increasing body of evidence suggests that the use of cystatin C GFR-estimating equations gives improved risk prediction of death and kidney failure unrelated to the accuracy of the GFR estimation; this improved predictive ability may relate to non-GFR determinants of serum cystatin C concentration. To delineate risk, KDIGO suggest that, if confirmation of CKD is required, cystatin C should be measured in adults who have creatinine-eGFR between 45 and 59 mL/min per 1.73 m^2 and who do not have markers of kidney damage.

POINTS TO REMEMBER

Cystatin C

- Cystatin C is a low-molecular-weight protein (12.8 kDa) produced at a fairly constant rate from most body cells.
- A modest increase in production is seen in obesity, hyperthyroidism, and inflammation.
- Cystatin C is removed from the circulation by the kidneys. It is freely filtered at the glomerulus and largely destroyed in the tubules.
- The clinical utility of cystatin C can be improved with a GFR estimating equation—for example, the CKD-EPI$_{cystatin\ C}$ equation.

TABLE 21.2 Representative Glomerular Filtration Rate Estimating Equations for Use in Adults

Abbreviation	Glomerular Filtration Rate Equation
Cockcroft and Gault	([(140 – age) × weight)] × 1.23)/(Scr) × 0.85 (if female)
MDRD (ID-MS traceable)	GFR (mL/min per 1.73 m^2) = 175 × $(Scr \times 0.01131)^{-1.154}$ × $(age)^{-0.203}$ × (1.210 if patient is black) × (0.742 if patient is female)
$CKD\text{-}EPI_{creat}$	141 × min(Scr × 0.01131/κ, 1)α × $max(Scr \times 0.01131/\kappa, 1)^{-1.209}$ × 0.993^{age} × 1.018 (if female) × 1.159 (if black), where Scr is serum creatinine, κ is 0.7 for females and 0.9 for males, α is –0.329 for females and –0.411 for males, min indicates the minimum of Scr/κ or 1, and max indicates the maximum of Scr/κ or 1.
$CKD\text{-}EPI_{cys}$	133 × $min(Scys/0.8, 1)^{-0.499}$ × $max(Scys/0.8, 1)^{-1.328}$ × 0.996^{Age} × 0.932 (if female), where min indicates the minimum of Scys/κ or 1, and max indicates the maximum of Scys/κ or 1.
$CKD\text{-}EPI_{creat\text{-}cys}$	135 × min(Scr × 0.01131/κ, 1)α × $max(Scr \times 0.01131/\kappa, 1)^{-0.601}$ × $min(Scys/0.8, 1)^{-0.375}$ × $max(Scys/0.8, 1)^{-0.711}$ × 0.995^{Age} × 0.969 (if female) × 1.08 (if black), where Scr is serum creatinine, Scys is serum cystatin C, κ is 0.7 for females and 0.9 for males, α is –0.248 for females and –0.207 for males, min indicates the minimum of Scr/κ or 1, and max indicates the maximum of Scr/κ or 1.
BIS1	3736 × $(Scr \times 88.4)^{-0.87}$ × $age^{-0.95}$ × 0.82 (if female)
BIS2	767 × $Scys^{-0.61}$ × $Scr^{-0.40}$ × $age^{-0.57}$ × 0.87 (if female)
CAPA cystatin C equation	130 × cystatin $C^{-1.069}$ × $age^{-0.117}$ – 7

Age is given in years, serum creatinine in micromoles per liter, serum cystatin C (Scys) in milligram per liter, and weight in kilograms.
BIS, Berlin Initiative Study; *CAPA,* Caucasian, Asian, pediatric, and adult; *CKD-EPI,* Chronic Kidney Disease–Epidemiology Consortium; *GFR,* glomerular filtration rate; *ID-MS,* isotope dilution-mass spectrometry; *MDRD,* modification of diet in renal disease; *Scr,* serum creatinine; *Scys,* serum cystatin C.

UREA

Biochemistry and Physiology

Urea ($CO[NH_2]_2$, Mr 60 Da) is the major nitrogen-containing metabolic product of protein catabolism in humans. During the process of protein catabolism, nitrogen derived from amino acids enters the urea cycle via intermediates, which include aspartate and ammonia. Urea is synthesized exclusively in the liver by the enzymes of the urea cycle (Fig. 21.3). The rate of urea production is dependent on the rate of protein catabolism from both dietary and endogenous sources, the latter being largely from muscle. Urea is distributed evenly throughout the total body water due to its ability to diffuse through most cell membranes facilitated by urea transporter proteins, and is therefore described as having a volume of distribution equal to the total body water.

Renal excretion accounts for more than 90% of urea removal from the body, with some losses through the gastrointestinal (GI) tract and skin. Urea is freely filtered at the glomerulus and, while it is neither actively reabsorbed nor secreted by the tubules, over 50% of urea moves passively out of the renal tubule, into the renal interstitium and back into the plasma. The back-diffusion of urea is dependent on urine flow rate, with less entering the interstitium in high-flow states (e.g., pregnancy) and more with low-flow situations (e.g., prerenal reduction in GFR due to fluid losses). Because water is resorbed to a far greater extent than the urea back-diffusion, urea is markedly concentrated during the process of urine formation, being present in urine at about 20 to 100 times the plasma concentration and is therefore the major contributor to urine

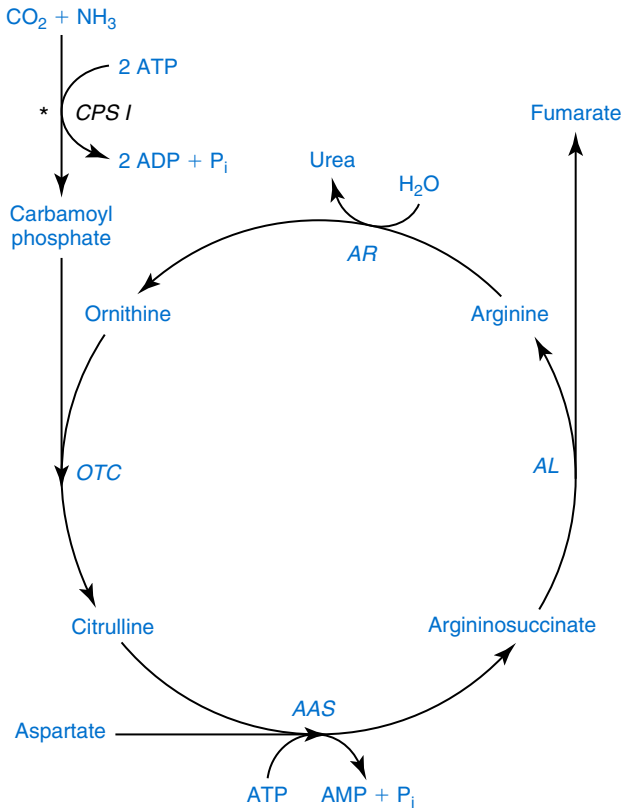


Fig. 21.3 The urea cycle pathway. *AAS,* Argininosuccinate synthetase; *AL,* argininosuccinate lyase; *AR,* arginase; *ADP,* adenosine diphosphate, *ATP,* adenosine triphosphate; *CPS I,* carbamyl phosphate synthetase I (**N*-acetylglutamate as positive allosteric effector); *OTC,* ornithine transcarbamylase; P_i, inorganic phosphate.

osmolality. The role of urea in the functioning of the kidney is complex, involving specific urea transporters under the action of vasopressin.

Analytical Methods

The vast majority of urea measurements in clinical laboratories are undertaken using enzymatic methods based on preliminary hydrolysis of urea with urease to generate ammonium ion, which then is quantitated. This approach has been used in end-point, kinetic, conductimetric, and dry chemistry systems.

The analytical specificity of all routine methods is good and typical assay precision is usually small compared with the biological variation of serum urea indicating good analytical performance. Reference materials, both pure and matrix matched, as well as reference methods based on IDMS are available for urea and are listed on the JCTLM database, and assays from different providers generally provide acceptably consistent results.

Reference Intervals

Serum urea concentrations are higher in men than in women, and there is an increase in the upper reference limit throughout adult life by a factor of about 50% from the 20s to the 70s. Pregnancy is associated with lower urea concentrations compared with age-matched nonpregnant women due to the associated glomerular hyperfiltration. In the pediatric age range, from after the first 2 weeks, the first year of life is associated with low serum urea, followed by a rise to slightly above adult concentrations to the age of about 10 years, then falling to young adult values by the late teens.

A reference interval for serum urea in healthy adults is 2.1 to 7.1 mmol/L (6 to 20 mg/dL expressed as urea nitrogen). Specific reference intervals will depend on the age, diet, and gender balance of the reference population (e.g., in adults older than 60 years of age), a higher reference interval (e.g., 2.9 to 8.2 mmol/L [8 to 23 mg/dL]) may be used. On an average protein diet, urinary urea excretion expressed as urea nitrogen is 12 to 20 g/day (430 to 710 mmol/day).

The term **blood urea nitrogen (BUN)** is used in some countries to refer to serum urea tests, with results reported as mg/dL of BUN. The SI system recommends the use of the term urea, expressed in mmol/L. To convert mg/dL BUN to mmol/L urea multiply the value by 2.8 (or by 0.357 for the reverse).

Clinical Significance

Measurement of blood and serum urea has been used for many years as an indicator of kidney function. However, it is generally accepted that creatinine measurement provides better information in this respect. Serum and urinary urea measurement may still provide useful clinical information in particular circumstances. The measurement of urea in dialysis fluids is widely used in assessing the adequacy of renal replacement therapy (see Chapter 35).

Urea concentration in serum is significantly affected by the rate of production as well as the rate of removal, which not only limits its value as a test of kidney function, but also allows its use for a range of other factors. For example, urea production, and therefore plasma concentration, is increased by a high-protein diet, increased endogenous protein catabolism, such as occurs in many hospitalized patients, reabsorption of blood proteins after GI hemorrhage, and treatment with cortisol or its synthetic analogues.

Urea removal from the circulation can be reduced due to any cause of reduced glomerular filtration, including prerenal, renal, or postrenal factors. The differential response of urea relative to creatinine can be useful diagnostically. In obstructive postrenal conditions (e.g., malignancy, nephrolithiasis, and prostatism), both serum creatinine and urea concentrations will be increased, although in these situations the increase in serum urea is greater than that of creatinine because of increased back-diffusion. These considerations give rise to the main proposed clinical use of serum urea—namely, its measurement in conjunction with that of serum creatinine and subsequent calculation of the urea to creatinine ratio. This can be used as a crude discriminator between prerenal and intrinsic causes of reduced GFR. For a normal individual on a normal diet, the reference interval for the ratio is between about 49 and 81 mmol urea/mmol creatinine (12 and 20 mg urea/mg creatinine). In the setting of known acute kidney injury (AKI), a raised urea to creatinine ratio (>81 mmol/mmol) is thought to be more indicative of a prerenal cause than an intrinsic renal cause, such as acute tubular necrosis.

A low serum urea concentration occurs in the setting of low protein intake—for example, starvation or anorexia nervosa. This is less commonly seen among hospital inpatients who, although often do not consume adequate nutrients, do not exhibit a low serum urea due to the presence of significant catabolism of endogenous protein. Low serum urea concentration can also be seen with end-stage liver disease due to decreased urea synthesis.

Measurement of urinary urea output provides a crude index of overall nitrogen balance and may be used as a guide to replacement in patients receiving parenteral nutrition.

URIC ACID (URATE)

Biochemistry and Physiology

Uric acid ($C_5H_4N_4O_3$, Mr 168) is the major product of catabolism of the purine nucleosides, adenosine, and guanosine in humans (Fig. 21.4). These precursors are metabolized into uric acid in a series of metabolic steps finishing with the conversion of xanthine to uric acid by the action of xanthine oxidase. The bulk of excreted uric acid arises from the degradation of endogenous nucleic acids (approximately 400 mg/day) with a lesser contribution from dietary sources (300 mg/day). Overproduction of uric acid may result from increased synthesis of purine precursors. In humans, approximately 75% of

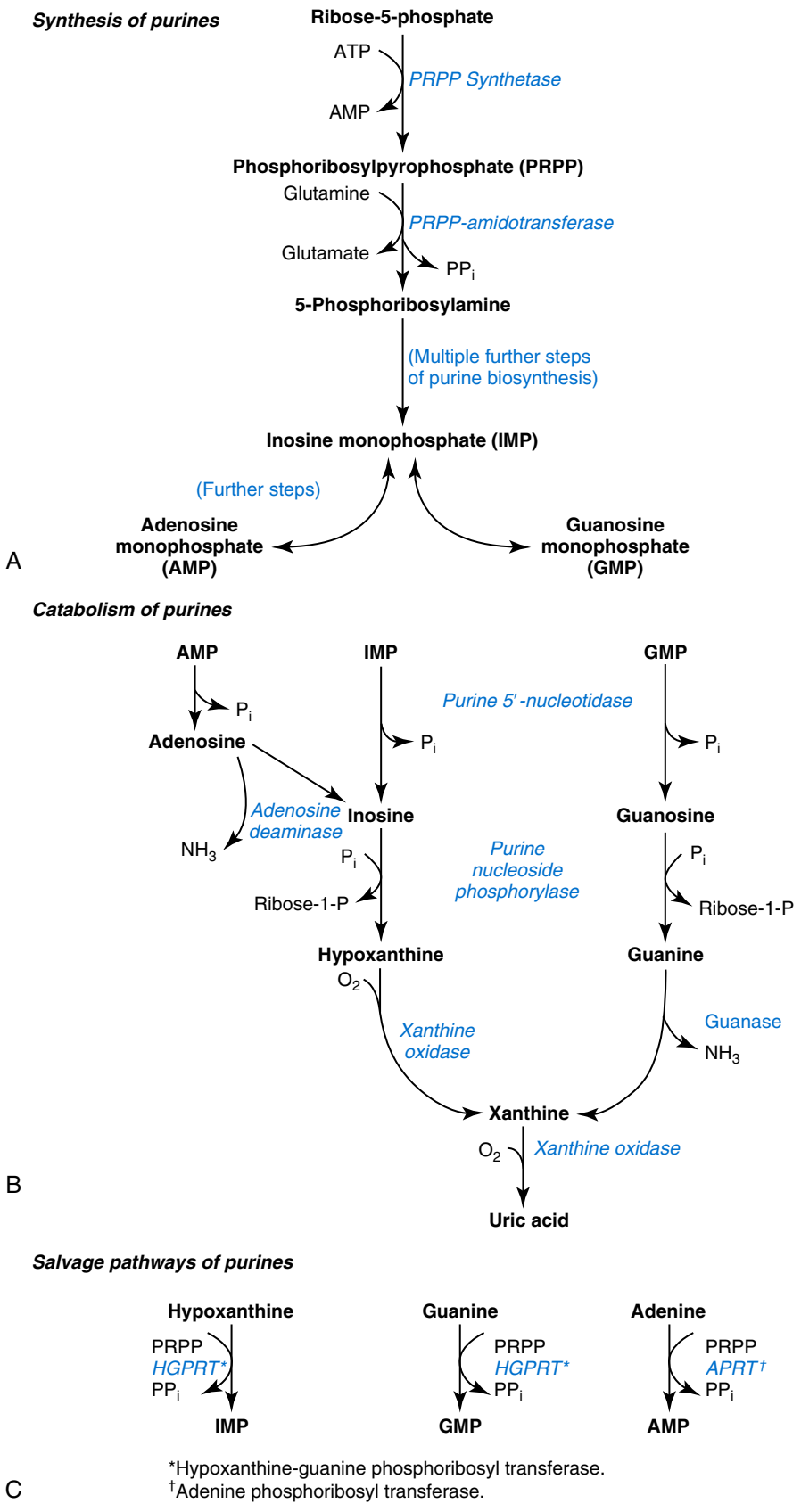


Fig. 21.4 Metabolism of purines: (A) Synthesis. (B) Catabolism. (C) Salvage pathways. *ATP,* Adenosine triphosphate.

uric acid excreted is lost in the urine; most of the remainder is secreted into the GI, where it is degraded to allantoin and other compounds by bacterial enzymes.

Renal handling of uric acid is complex and involves four sequential steps: (1) glomerular filtration of virtually all the uric acid in capillary plasma entering the glomerulus; (2) near complete reabsorption in the proximal convoluted tubule; (3) secretion into the distal portion of the proximal tubule; and (4) further reabsorption in the distal tubule. The net urinary excretion of uric acid is 6% to 12% of the amount filtered.

The physicochemical properties of uric acid are important in considering uric acid concentrations in the circulation, in tissue, and in the kidneys. The first pKa of uric acid is 5.75; above this pH, uric acid exists chiefly as urate ion, which is more soluble than uric acid. At a urine pH below 5.75, uric acid is the preponderant form. On the basis of the ionization form in the circulation, it has been proposed that urate should be the preferred term rather than uric acid when describing serum measurements.

Sample Collection

Most blood collection tubes are suitable for sample collection and urate is sufficiently stable that no special collection or sample handling conditions are required. Traditionally, urinary uric acid excretion is determined in a 24-hour urine sample. If analysis is not undertaken promptly, alkalinization of the sample is recommended to maintain uric acid in solution.

Analytical Methods

Methods based on the enzyme uricase ([urate:oxygen] oxidoreductase; EC 1.7.3.3) are the most commonly used routine methods for measuring uric acid. Uricase acts on uric acid to produce allantoin, hydrogen peroxide, and carbon dioxide. The reaction can be observed in either the kinetic or the equilibrium mode. Most current enzymatic assays for serum urate involve a peroxidase system coupled with one of a number of oxygen acceptors to produce a chromogen. The step for quantitation of the hydrogen peroxide produced is sometimes referred to as a Trinder reaction. One popular method measures hydrogen peroxide with the aid of horseradish peroxidase and an oxygen acceptor to yield a chromogen in the visible spectrum. The most common oxygen acceptor used is 4-aminophenazone, together with phenol or a substituted phenol. The JCTLM lists pure substance and matrix matched certified reference materials for uric acid as well as IDMS reference methods and reference measurement services for analysis in serum and urine. Evidence suggests that most routine methods are meeting clinical needs. Interference in Trinder assays can be caused by antioxidants such as ascorbate (e.g., intravenous vitamin C), bilirubin, and unspecified interferants in serum from patients with kidney failure. Rasburicase is a urate-consuming enzyme (urate oxidase) used to treat patients with tumor lysis syndrome. Rasburicase can lower serum urate concentrations in blood collection tubes ex vivo unless both whole blood and serum are cooled before and after separation or treated by acidification.

Reference Intervals

The serum urate concentration increases gradually with age, rising about 10% between the ages of 20 and 60 years. A significant rise is seen in women after menopause, reaching concentrations similar to those found in men. Additionally, higher concentrations of serum urate are found with increases in waist circumference, body mass index, and other components of the metabolic syndrome. A population reference interval may require partitioning on the basis of gender and age, and will be affected by the metabolic status of the population. It can be argued that a clinical decision point for the relevant clinical question may be more useful than a population reference interval given the interaction with common comorbidities. A reference interval for serum urate has been reported to be 3.5 to 7.2 mg/dL (0.21 to 0.43 mmol/L) for males and 2.6 to 6.0 mg/dL (0.16 to 0.36 mmol/L) for females.

During pregnancy, serum urate concentrations fall during the first trimester and until about 24 weeks' gestation, when values begin to rise and eventually exceed nonpregnant concentrations. Urinary uric acid excretion in individuals on a diet containing purines is 250 to 750 mg/day (1.5 to 4.5 mmol/day). Excretion may decrease by 20% to 25% on a purine-free diet to less than 400 mg/day (2.4 mmol/day).

Clinical Significance

The most common clinical use for urate assessment is risk assessment for gout and determination of therapy adequacy. Other clinical conditions where serum urate measurements have potential utility include cardiovascular risk assessment and a diagnosis of preeclampsia. Urine uric acid measurements may play a role in assessing the cause of hyperuricemia and in assessing the risk of renal stone formation.

Hyperuricemia

The major causes of **hyperuricemia** are summarized in Box 21.1. Asymptomatic hyperuricemia is frequently detected through biochemical screening, although there is no evidence for clinical benefit resulting from this practice.

Gout

Gout occurs when monosodium urate precipitates in joints and tissues from supersaturated body fluids. These deposits of urate are responsible for the clinical signs and symptoms. Gouty arthritis is caused by urate crystal formation in joint fluid and may be associated with deposits of crystals (tophi) in tissues surrounding the joint. The deposits may occur in other soft tissues as well, and wherever they occur they elicit an intense inflammatory response. The first metatarsophalangeal joint (big toe) is the classic site for gout. Gout is a condition characterized by occasional attacks and long periods of remission. It is important to appreciate that the serum uric acid concentration is often normal during an acute attack and is not a component of the diagnostic criteria. The demonstration of uric acid crystals in joint aspirate fluid is the pathognomonic sign for gout. For details on the diagnostic criteria and associated evidence base, readers are referred to the European League Against Rheumatism (EULAR)

BOX 21.1 Causes of Hyperuricemia

Increased Formation

Primary

Idiopathic
Inherited metabolic disorders

Secondary

Excess dietary purine intake
Increased nucleic acid turnover (e.g., leukemia, myeloma, radiotherapy, chemotherapy, trauma)
Psoriasis
Altered ATP metabolism
Tissue hypoxia
Preeclampsia
Alcohol

Decreased Excretion

Primary

Idiopathic

Secondary

Acute or chronic kidney disease
Increased renal reabsorption
Reduced secretion
Lead poisoning
Preeclampsia
Organic acids (e.g., lactate and acetoacetate)
Salicylate (low doses)
Thiazide diuretics
Trisomy 21 (Down syndrome)

ATP, Adenosine triphosphate.

guidelines. This report also provides information on the risk factors for the development of gout, which represent those for increasing concentrations of serum uric acid; of these, the major ones are male gender, CKD, hypertension, and obesity.

Gout may be classified as primary or secondary. Primary gout is associated with essential hyperuricemia, which has a polygenic basis. In more than 99% of cases, the cause is uncertain, but the condition is probably due to a combination of metabolic overproduction of purines, decreased renal excretion, and increased dietary intake. Very rarely, primary gout is attributable to inherited defects of enzymes in the pathways of purine metabolism.

Secondary gout is a result of hyperuricemia attributable to several identifiable causes such as CKD; as a consequence of administration of drugs (in particular diuretics); and ketoacidosis or to lactic acidosis (by reducing tubular secretion of urate). Increased nucleic acid turnover and a consequent increase in catabolism of purines may be encountered in rapid proliferation of tumor cells and in massive destruction of tumor cells on therapy with certain chemotherapeutic agents.

Management of an acute attack of gout can be divided into broad strategies: management of the attack and the prevention of subsequent attacks. The main focus for acute management is pain relief. Avoidance of subsequent attacks is based on lifestyle and pharmacotherapy with the aim of reducing plasma urate, which is monitored to assess therapeutic success. An evidence-based treatment target for prevention of gout recurrence is a serum urate concentration less than 0.36 mmol/L.

Kidney Disease

Kidney disease associated with hyperuricemia may take one or more of several forms: (1) gouty nephropathy with urate deposition in renal parenchyma, (2) acute intratubular deposition of urate crystals, and (3) urate nephrolithiasis.

Kidney Stones

Uric acid kidney stones occur in approximately one in five patients with clinical gout. Although serum and urinary uric acid should be measured in stone formers, many uric acid stone formers do not demonstrate hyperuricuria or hyperuricemia. Uric acid stone formation is promoted by the passage of a persistently acid urine with undissociated uric acid (pKa 5.57) being relatively insoluble, whereas urate at pH 7.0 is greater than 10 times more soluble. Pure uric acid stones account for approximately 8% of all urinary tract stones and, unlike many of the calcium-containing stones, are radiolucent. Allopurinol is the mainstay of treatment of uric acid stones.

Preeclampsia

Preeclampsia is pregnancy-induced hypertension associated with proteinuria (>0.3 g/day) and often edema and may become life threatening for the mother or the fetus (see Chapter 44). The role of uric acid measurement in the management of preeclampsia is uncertain: measurement recommendations appear in some guidelines but not others. If assessing urate concentrations in this setting, it is important to use pregnancy-specific reference intervals.

Inherited Diseases

Hyperuricemia can be a feature of several inherited disorders of purine metabolism, most of which are rare, and the diagnosis requires support from a specialist purine laboratory. Readers are referred to specialist textbooks for further information.

Hypouricemia

Hypouricemia, often defined as serum urate concentrations less than 2.0 mg/dL (0.12 mmol/L), is much less common than hyperuricemia. It may be secondary to severe hepatocellular disease with reduced purine synthesis or xanthine oxidase activity, or defective renal tubular reabsorption of uric acid. Defective reabsorption may be congenital, as in generalized Fanconi syndrome, or acquired. Overtreatment of hyperuricemia with allopurinol or uricosuric drugs and cancer chemotherapy with 6-mercaptopurine or azathioprine (inhibitors of de novo purine synthesis) may also cause hypouricemia.

Cardiometabolic Outcomes and Urate

Recent publications have highlighted positive associations between serum urate concentrations and a range of important cardiovascular and health outcomes that are generally

associated with the metabolic syndrome (e.g., obesity and insulin resistance, hypertension and renal disease, cardiovascular disease, type 2 diabetes, and all cause and cardiovascular mortality). More important, as well as indicating risk for patients with hyperuricemia, these relationships occur within the usually described "normal range" with the higher urate values associated with increasing risk. While these associations appear to be robust, it is less clear whether or not there is a causal association or whether urate-lowering therapy may be beneficial. The role of measurement of urate in the routine laboratory, validated clinical decision points, and appropriate management of patients based on the results remains to be determined; however, it is likely that there will be more roles for this analyte in the future.

POINTS TO REMEMBER

Urate

- Urate (uric acid) is the major end product of catabolism of the purine nucleosides.
- Excretion of urate is predominantly renal.
- Reference intervals for urate are influenced by many factors, including age, gender, menopausal status, diet, and waist circumference.
- Gout occurs when monosodium urate precipitates in joints and tissues from supersaturated body fluids, eliciting an intense inflammatory response.
- When treating gout, a target urate concentration of <6 mg/dL (<0.36 mmol/L) is used.

ASSESSMENT OF KIDNEY FUNCTION: GLOMERULAR FILTRATION RATE

The GFR is widely accepted as the best overall measure of kidney function. Progressive decrease in GFR is associated with the clinical signs and laboratory changes of kidney damage in all forms of kidney disease. Measuring GFR in established disease is useful in targeting treatment, monitoring progression, and predicting the point at which renal replacement therapy will be required. It is also used as a dosage guide to prevent toxicity by drugs excreted by the kidneys. There are a range of methods available to measure and estimate the GFR.

The Concept of Clearance

All tests for GFR are based on the clearance of a substance from the circulation by the kidneys. Renal clearance of a substance is defined as the volume of plasma from which the substance is completely cleared (removed) by the kidneys per unit of time. Provided a substance S is freely filtered in the glomerulus and neither secreted, resorbed, nor metabolized in the tubules, the renal clearance of S is equal to the GFR, and carries the same unit dimension (volume/time). Under these conditions the following equations apply.

$$GFR \times PS = US \times V \quad (21.1)$$

$$GFR = (US \times V)/PS \quad (21.2)$$

where GFR = the flow rate in mL per minute of plasma through the glomerular membranes, equivalent to a clearance in units of mL of plasma cleared of a substance per minute; US = urinary concentration of the substance; V = volumetric flow rate of urine in milliliter per minute; and PS = plasma concentration of the substance (in the same units as the urinary concentration of the substance).

The units of milliliter per minute is the most commonly used unit for GFR and is recommended for universal adoption to avoid confusion from the use of different units.

Kidney size and GFR are roughly proportional to body size, with larger people having a higher GFR and vice versa; to remove this effect it is conventional to adjust GFR to a standard body surface area (BSA) of 1.73 m^2. There are a number of formulae available to enable BSA estimation (e.g., Du Bois and Du Bois, Haycock). A variety of markers, both exogenous (radioisotopic and nonradioisotopic) and endogenous, have been used to measure or estimate GFR with the choice of marker and procedure based on cost and availability as well as accuracy.

Exogenous Markers of Glomerular Filtration Rate

Protocols for direct measurement of GFR are based on exogenous markers, with urinary clearance of inulin considered to be the "gold standard." There are many other exogenous markers and also different protocols for GFR determination. Markers may be given intravenously either by continuous infusion or single bolus and measured in serial urine or blood collections. The nonradioisotopic markers include inulin and iohexol, and the radiolabeled markers include labeled EDTA, DTPA, and iothalamate. Determination of GFR with exogenous markers is available in many hospitals and should be considered for clinical use when accurate GFR determination is required (e.g., prior to kidney donation, when dosing particularly toxic medications).

Endogenous Markers of Glomerular Filtration Rate

The most widely used endogenous marker of GFR is creatinine, expressed as its serum concentration, as renal clearance or included in equations for eGFR. Although a range of other low-molecular-weight compounds have been investigated for clinical use, only creatinine and cystatin C are recommended for routine use.

Creatinine

Serum creatinine concentration is a product of the rate of release into the circulation as well as the rate of removal, and awareness of the influence of both of these factors is required for an understanding of the use of creatinine measurements. Creatinine is produced in muscle from creatine at a fairly constant rate proportional to the amount of muscle present. Because muscle mass is affected by age, gender, race, nutrition, exercise, and other factors, the rate of production is variable between different people. The major route of creatinine removal from the circulation is by free filtration at the renal glomerulus although there is some additional loss into

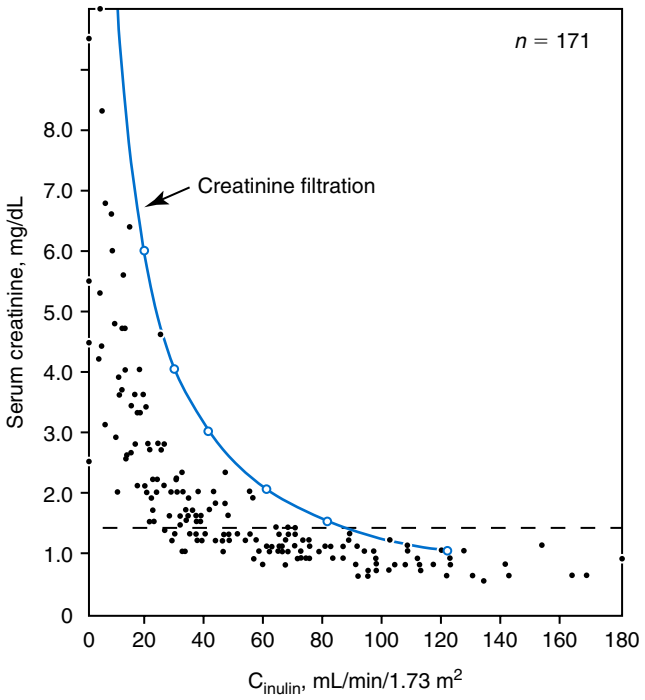


Fig. 21.5 The relationship between serum creatinine concentration and glomerular filtration rate *(GFR)*, measured as the clearance of inulin, in 171 patients with glomerular disease. The hypothetical relationship between GFR and serum creatinine is shown as a *continuous line*, assuming that only filtration of creatinine takes place. The *dashed horizontal line* represents an upper limit of normal for serum creatinine (1.4 mg/dL, 124 μmol/L), which is widely used at the time of publication. Because of creatinine secretion and/or a creatinine deficit through gut excretion, the serum creatinine consistently overestimates the GFR. (From Shemesh, O., Golbetz, H., Kriss, J. P., & Myers, B.D. [1985]. Limitations of creatinine as a filtration marker in glomerulopathic patients. *Kidney International, 28*, 830–838.)

the urine by tubular secretion and by loss through the GI tract. Due to the key role of glomerular filtration in creatinine removal, the concentration of creatinine in the circulation is inversely related to GFR—that is, for the same creatinine production, a halving of the GFR will lead to an approximate doubling of the serum creatinine concentration. Importantly, creatinine is convenient and cheap to measure and is widely available.

A range of factors affect creatinine metabolism and measurement, which are important when they are used to assess renal function. Drug effects can be due to blockage or tubular secretion (e.g., cimetidine, trimethoprim) or interference in creatinine assays. In patients with CKD, extrarenal clearance of creatinine becomes important when caused by degradation as a result of bacterial overgrowth in the small intestine, further blunting the anticipated increase in plasma creatinine in response to falling GFR. An important factor for any creatinine-based interpretation is that serum creatinine can remain within the reference interval until notable kidney function has been lost. Additionally, serum creatinine measurement will not detect people with mildly reduced GFR (60 to 89 mL/min per 1.73 m^2) and will fail to identify many patients with CKD and category 3A GFR (45 to 59 mL/min per 1.73 m^2) (Fig. 21.5). Thus, although an increased serum creatinine concentration generally equates with impaired kidney function, a normal serum creatinine does not necessarily equate with normal kidney function. However, while the interpretation of creatinine concentration against population-based decision points is insensitive for the detection of CKD, changes in serum creatinine within an individual may be used as a sensitive tool for detecting changes in kidney function, whether the results are within or outside population reference intervals. In this setting, changes greater than expected by chance (i.e., that due to biological and analytical variation) are likely to indicate a significant change in GFR in a patient.

Creatinine Clearance

A measured creatinine clearance, performed using a 24-hour urine collection and serum collection, ideally during this period, has been used for many years as a marker of GFR. However, the difficulty in obtaining a complete, timed 24-hour urine collection, together with a high within-subject biological variation in creatinine excretion and the known overestimation of GFR due to creatinine tubular secretion make this test both inaccurate and imprecise as a measure of GFR. In spite of these limitations, creatinine clearance remains in common use in recommendations for drug-dosing decisions (see later).

Estimating Glomerular Filtration Rate

The approximate mathematical relationship between serum creatinine and GFR (GFR α 1/serum creatinine) can be improved by correcting for some of the confounding variables that influence this relationship. Many different equations have been derived that estimate GFR using serum creatinine and inputs of some or all of gender, body size, race, and age. These generally produce a better estimate of GFR than serum creatinine alone or measured creatinine clearance, and professional societies throughout the world have recommended that such estimates should be used in association with serum creatinine. In adults, several such equations have been widely used, including the Cockcroft and Gault equation, the Modification of Diet in Renal Disease (MDRD) Study equations, and the CKD-EPI equations.

The Cockcroft and Gault equation is one of the earliest of these equations (see Table 21.2). It remains widely used in the context of assessing drug dosages for patients with kidney impairment (see later). However, the requirement for the measurement of body weight, the lack of a version validated for standardized creatinine assays, and the availability of demonstrably better equations have limited its ongoing use.

The era of GFR estimation as a widely used simple clinical tool began with publication of the MDRD Study equation in 1999, followed by its abbreviated four variable (creatinine, age, gender, race [black/white]) form in 2000. The original MDRD equation was updated in 2006 to a version suitable for IDMS traceable creatinine assays and this was widely endorsed by national organizations. Generally, the MDRD equation was found to provide a more accurate assessment of GFR than the Cockcroft and Gault equation, with the important advantage

of not requiring knowledge of weight, allowing automatic reporting of eGFR with the results for serum creatinine.

In 2009 the Chronic Kidney Disease Epidemiology Collaboration described a further equation (CKD-EPI$_{creat}$), again based on serum creatinine, gender, race, and age (see Table 21.2). This equation showed some improvement over MDRD, especially at higher levels of GFR (>60 mL/min per 1.73 m^2). KDIGO have recommended that the CKD-EPI$_{creat}$ equation should be used to estimate GFR.

Because the relationship between creatinine and kidney function varies among individuals, disease conditions, and populations, no GFR equation to date has found universal applicability. This remains an intense area of ongoing research. Other equations that have been developed are the CKD-EPI equations based on cystatin C, either alone or in combination with creatinine, and the Berlin Initiative Study (BIS) Group published equations with claimed superior performance in older people (see Table 21.2).

The relationship between creatinine and GFR has been shown to vary between some ethnicities, presumably due to differences in muscularity and dietary pattern. For example, the MDRD and CKD-EPI$_{creat}$ equation publications showed a clear distinction between African American and non-African Americans. The requirement to adjust GFR estimating equations for race is a potential limitation to their global application and local implementation.

Limitations to creatinine-based GFR estimating equations include interferences in creatinine measurement and patient-specific factors. Any factor affecting creatinine measurement, such as bias, imprecision, drug effects, hemolysis, icterus, or lipemia, will directly affect the eGFR. Any relevant way in which a patient is different from those included in the trials to establish the equations will also affect the accuracy of the equation at an individual level. These include patients with AKI, in whom serum creatinine concentrations may change rapidly and with extremes of muscularity (e.g., bodybuilders, amputees, and people with muscle wasting disorders). The equations are also not suitable for patients on dialysis and for use during pregnancy; adult equations should not be used with patients under 18 years of age. Notwithstanding this, in general, equations improve the estimation of GFR compared with serum creatinine alone.

Reference Intervals

An average GFR in young healthy individuals is 110 mL/min per 1.73 m^2 (5th to 95th percentiles approximately 90 to 140 mL/min per 1.73 m^2) with a gradual fall with increasing age. The age-related changes are also seen in outputs of eGFR equations. Age-related reference intervals are not widely used in clinical medicine with decision points based on CKD staging or advice for drug dosing providing the usual clinical support.

Glomerular Filtration Rate Measurement and Estimation: Future Considerations

Evaluation of kidney function remains an essential component of medical practice, and much has been achieved in recent years to improve and standardize estimates of GFR. However, there remains room for improvement, and research into new and better markers, equations, and procedures for the measurement of kidney function will continue. The challenge to the clinical lab is to use the best available tools in a coordinated manner so that evidence-based clinical decisions can be made. One particular focus should be collaboration between laboratories so that patients and doctors obtain the same clinical information irrespective the laboratory selected.

REVIEW QUESTIONS

1. The plasma concentration of creatinine is maintained within narrow limits predominantly by:
   a. the constant catabolism of purines.
   b. the constant rate of protein metabolism.
   c. the glomerular filtration rate.
   d. an individual's diet.
2. Symptoms of gout are caused by:
   a. decreased renal perfusion as in heart failure.
   b. precipitation of excessive uric acid in joints and the urinary tract.
   c. a low-protein diet.
   d. liver dysfunction as in hepatitis.
3. The main nitrogen-containing compound that accounts for more than 75% of excreted nonprotein nitrogen is:
   a. uric acid.
   b. creatinine.
   c. creatine.
   d. urea.
4. A deficiency of the enzyme xanthine oxidase due to severe hepatocellular disease will lead to:
   a. increased urea concentration.
   b. decreased creatinine concentration.
   c. hypouricemia.
   d. hyperuricemia.
5. During protein breakdown, the amino acid nitrogen groups are converted to urea through the urea cycle in which of the following?
   a. Liver
   b. Kidneys
   c. Heart
   d. GI tract
6. Compensated Jaffe assays for creatinine assessment:
   a. add a fixed value to the assay result to compensate for interferants.
   b. subtract a fixed value from each result to compensate for noncreatinine interference.
   c. use a special formula that involves body mass and urine production to "normalize" creatinine results.
   d. involve averaging of values to remove possible discrepancies between two samples obtained at different times of the day.

7. A value of 8.4 mg/dL of urea converts to what in mmol/L?
   a. 2.99
   b. 29.9
   c. 3.92
   d. 39.2
8. Ascorbic acid and bilirubin are interferants in which of the following assays?
   a. Jaffe reaction and uricase methods
   b. Jaffe reaction only
   c. Jaffe reaction, urease methods, and uricase methods
   d. Urease methods only
9. In the uricase method of uric acid measurement, the main interferants that must be minimized are:
   a. phenol and bilirubin.
   b. ascorbic acid and bilirubin.
   c. endogenous ammonia and oxidizing agents.
   d. ketone bodies and protein.
10. Uric acid is:
   a. the major product of protein catabolism.
   b. a derivative of muscle creatine.
   c. a metabolite of urinary nitrogen.
   d. the major product of purine catabolism.

SUGGESTED READINGS

McTaggart, M. P., Newall, R. G., Hirst, J. A., et al. (2014). Diagnostic accuracy of point-of-care tests for detecting albuminuria: a systematic review and meta-analysis. *Annals of Internal Medicine*, *160*, 550–557.

Price, C. P., Newall, R. G., & Boyd, J. C. (2005). Use of protein: creatinine ratio measurements on random urine samples for prediction of significant proteinuria: a systematic review. *Clinical Chemistry*, *51*, 1577–1586.

Miller, W. G., Bruns, D. E., Hortin, G. L., et al. (2009). Current issues in measurement and reporting of urinary albumin excretion. *Clinical Chemistry*, *55*, 24–38.

Kidney Disease Improving Global Outcomes. (2013). Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. *Kidney International*, *3*, 1–150.

National Institute for Health and Care Excellence. (2014). Chronic kidney disease. *Early identification and management of chronic kidney disease in adults in primary and secondary care*. <http://www.nice.org.uk/nicemedia/live/13712/66658/.pdf>.

Bachmann, L. M., Nilsson, G., Bruns, D. E., et al. (2014). State of the art for measurement of urine albumin: comparison of routine measurement procedures to isotope dilution tandem mass spectrometry. *Clinical Chemistry*, *60*, 471–480.

Lamb, E. J., Mackenzie, F., & Stevens, P. E. (2009). How should proteinuria be detected and measured? *Annals of Clinical Biochemistry*, *46*, 205–217.

Atkins, R. C., Briganti, E. M., Zimmet, P. Z., et al. (2003). Association between albuminuria and proteinuria in the general population: the ausdiab study. *Nephrology, dialysis, transplantation: official publication of the European Dialysis and Transplant Association European Renal Association*, *18*, 2170–2174.

Matsushita, K., van der Velde, M., Astor, B. C., et al. (2010). Association of estimated glomerular filtration rate and albuminuria with all-cause and cardiovascular mortality in general population cohorts: a collaborative meta-analysis. *Lancet*, *375*, 2073–2081.

Greenberg, N., Roberts, W. L., Bachmann, L. M., et al. (2012). Specificity characteristics of 7 commercial creatinine measurement procedures by enzymatic and Jaffe method principles. *Clinical Chemistry*, *58*, 391–401.

Myers, G. L., Miller, W. G., Coresh, J., et al. (2006). Recommendations for improving serum creatinine measurement: a report from the Laboratory Working Group of the National Kidney Disease Education Program. *Clinical Chemistry*, *52*, 5–18.

Ceriotti, F., Boyd, J. C., Klein, G., et al. (2008). Reference intervals for serum creatinine concentrations: assessment of available data for global application. *Clinical Chemistry*, *54*, 559–566.

Grubb, A., Blirup-Jensen, S., Lindstrom, V., et al. (2010). First certified reference material for cystatin C in human serum ERM-DA471/IFCC. *Clinical Chemistry and Laboratory Medicine*, *48*, 1619–1621.

Inker, L. A., Schmid, C. H., Tighiouart, H., et al. (2012). A new equation to estimate glomerular filtration rate from standardized creatinine and cystatin C. *The New England Journal of Medicine*, *367*, 20–29.

Peralta, C. A., Shlipak, M. G., Judd, S., et al. (2011). Detection of chronic kidney disease with creatinine, cystatin C, and urine albumin-to-creatinine ratio and association with progression to end-stage renal disease and mortality. *JAMA: Journal of the American Medical Association*, *305*, 1545–1552.

Soveri, I., Berg, U. B., Bjork, J., et al. (2014). Measuring GFR: a systematic review. *American Journal of Kidney Diseases: the official journal of the National Kidney Foundation*, *64*, 411–424.

Levey, A. S., Stevens, L. A., Schmid, C. H., et al. (2009). A new equation to estimate glomerular filtration rate. *Annals of Internal Medicine*, *150*, 604–612.

Earley, A., Miskulin, D., Lamb, E. J., et al. (2012). Estimating equations for glomerular filtration rate in the era of creatinine standardization: a systematic review. *Annals of Internal Medicine*, *156*, 785–795.

Maple-Brown, L. J., Hughes, J. T., Lawton, P. D., et al. (2012). Accurate assessment of kidney function in indigenous Australians: the estimated GFR study. *American Journal of Kidney Diseases: the official journal of the National Kidney Foundation*, *60*, 680–682.

Teo, B. W., Xu, H., Wang, D., et al. (2012). Estimating glomerular filtration rates by use of both cystatin C and standardized serum creatinine avoids ethnicity coefficients in Asian patients with chronic kidney disease. *Clinical Chemistry*, *58*, 450–457.

22

Carbohydrates

David B. Sacks

OBJECTIVES

1. Define the following:
   Carbohydrate
   Gluconeogenesis
   Glycogen
   Glycogen storage disease
   Glycogenesis
   Glycogenolysis
   Glycolysis
   Glycoprotein
   Hypoglycemia
   Ketone
   Monosaccharide, disaccharide, and polysaccharide
2. Provide two examples each of a monosaccharide, a disaccharide, and a polysaccharide; describe the chemical structure of each.
3. Summarize the regulation of glucose concentration in the blood, including two hormones involved in this process.
4. Explain the relationship between glucose, lactate, and pyruvate.
5. Discuss the brain's dependence on glucose, including how the brain is affected in a hypoglycemic state and the alternative energy source used.
6. State the blood glucose concentration at which hypoglycemia is typically diagnosed.
7. List three causes of fasting hypoglycemia and three causes of postprandial (nonfasting) hypoglycemia; describe the 72-hour fast test for diagnosing hypoglycemia.
8. List two causes of hypoglycemia in persons with type 1 diabetes.
9. Compare glucose values in fasting whole-blood, capillary, and venous blood specimens; state the reasons for the discrepant values.
10. State the effects of glycolysis on glucose in whole blood; list two ways in which glycolysis can be inhibited.
11. Describe the hexokinase method for measurement of glucose in blood; state the specimen requirements and the principle of the reaction, and list the known interferences.
12. Describe the glucose oxidase method for measurement of glucose in blood; state the specimen requirements and the principle of the reaction, and list the known interferences.
13. Describe the chemical reaction used to assess lactate concentration in blood; describe the chemical reaction used to assess pyruvate concentrations in blood.
14. State the specimen collection and storage requirements for blood lactate and pyruvate analyses.
15. Analyze and resolve case studies regarding carbohydrate disorders; correlate the results of carbohydrate analysis with carbohydrate disorders.

KEYWORDS AND DEFINITIONS

Aldehyde An organic compound with a carbonyl group (a carbon atom double-bonded to an oxygen) at the end of the carbon chain bonded to hydrogen and an R group (usually an alkyl group).

Carbohydrate Aldehyde or ketone derivatives of polyhydroxy alcohols composed of carbon, hydrogen, and oxygen in a ratio of 1:2:1.

Cori cycle The mechanism by which lactate produced by muscles is carried to the liver, converted back to glucose via gluconeogenesis, and returned to the muscles.

Diabetes mellitus A group of metabolic disorders of carbohydrate metabolism in which glucose is underutilized, producing hyperglycemia.

Glucagon A protein hormone that maintains blood glucose concentration by increasing blood glucose through glycogenolysis.

Glucose A six-carbon monosaccharide derived from the breakdown of carbohydrates in the diet or in body stores; can also be endogenously synthesized from protein or the glycerol moiety of triglycerides.

Glycogen An extensively branched polysaccharide containing many glucose residues and found particularly in muscle and liver cells for glucose storage.

Hypoglycemia Blood glucose concentration in the blood decreased below a healthy reference interval.

Insulin A protein hormone that maintains blood glucose concentration by decreasing blood glucose through cellular uptake.

Ketone An aldehyde that has a carbonyl group (carbon atom double-bonded to an oxygen atom) at any position other than at the end of the carbon chain.

Lactate An intermediary product in glucose metabolism that accumulates in the blood predominantly when tissue oxygenation is decreased, as during strenuous exercise; an increased blood lactate concentration is called lactic acidosis.

Pyruvate An organic acid formed from glucose through glycolysis.

Carbohydrates, including sugar and starch, are widely distributed in plants and animals. They perform multiple functions, such as serving as structural components in RNA and DNA (ribose and deoxyribose sugars) and providing a source of energy (glucose). Glucose is derived from (1) the breakdown of carbohydrates in the diet (grains, starchy vegetables, and legumes) or in body stores (glycogen), and (2) endogenous synthesis from protein or from the glycerol moiety of triglycerides. When energy intake exceeds expenditure, the excess is converted to fat and glycogen for storage in adipose tissue and liver or muscle, respectively. When energy expenditure exceeds caloric intake, endogenous glucose formation occurs from the breakdown of carbohydrate stores and from noncarbohydrate sources (e.g., amino acids, lactate, glycerol).

The glucose concentration in the blood is maintained within a fairly narrow interval under diverse conditions (feeding, fasting, or severe exercise) by hormones such as **insulin**, **glucagon**, and epinephrine (see details in Chapter 33). Measurement of glucose is one of the most commonly performed procedures in hospitals and other health care chemistry laboratories, and glucose measurement by point-of-care and home-use meters is a multi-billion-dollar industry. The most frequently encountered disorder of carbohydrate metabolism is high blood glucose concentration due to **diabetes mellitus**, which affects approximately 9% of the US adult population. The incidence of hypoglycemia (low blood glucose) is unknown but is low (excluding patients who use exogenous insulin to control blood glucose).

CHEMISTRY OF CARBOHYDRATES

Carbohydrates are aldehyde or ketone derivatives of polyhydroxy (more than one –OH group) alcohols or compounds that yield these derivatives on hydrolysis.

Monosaccharides

A monosaccharide is a simple sugar that consists of a single polyhydroxy aldehyde or ketone unit and is unable to be hydrolyzed to a simpler form. The backbone is made up of several carbon atoms. Sugars containing three, four, five, six, and seven carbon atoms are known as *trioses, tetroses, pentoses, hexoses,* and *heptoses,* respectively. One of the carbon atoms is double-bonded to an oxygen atom to form a carbonyl group. An **aldehyde** has the carbonyl group at the end of the carbon chain, whereas if the carbonyl group is at any other position, a **ketone** is formed (Fig. 22.1). The simplest carbohydrate is glycol aldehyde, the aldehyde derivative of ethylene glycol. The aldehyde and ketone derivatives of glycerol are, respectively, glyceraldehyde and dihydroxyacetone (see Fig. 22.1). Monosaccharides are termed *aldoses* and *ketoses* according to the position of the carbonyl group (Fig. 22.2).

Compounds that are identical in composition and differ only in spatial configuration are called *stereoisomers.* The carbon atoms in the unbranched chain are numbered 1 to 6, as shown by the numbers at the left of the formula for D-glucose in Fig. 22.2. The designation D- or L- refers to the position of the hydroxyl group on the carbon atom adjacent to the last (bottom) CH_2OH group. In general, designations of D- and L- for a sugar molecule refer to the stereoisomeric forms of the highest-numbered asymmetrical carbon atom.* By convention, the D-sugars are written with the hydroxyl group on the right, and the L-sugars are written with the hydroxyl group on the left (see Fig. 22.2). Most sugars in the human body are of the D-configuration. Several different structures exist, depending on the relative positions of the hydroxyl groups on the carbon atoms.

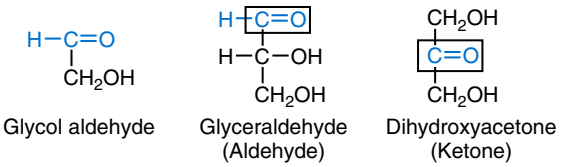


Fig. 22.1 Three-carbon carbohydrates.

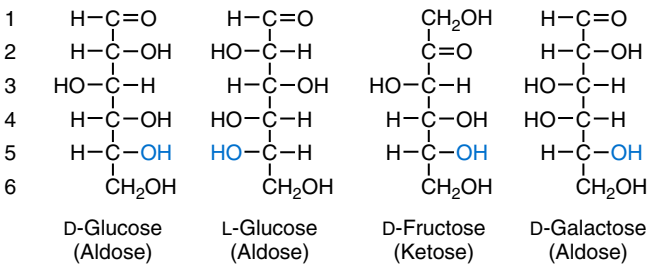
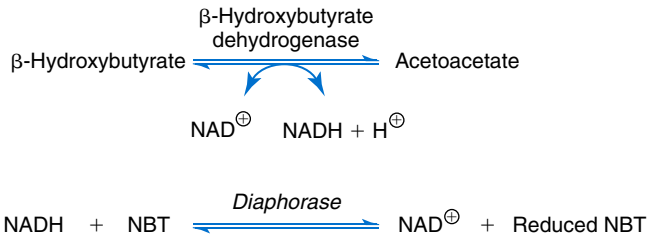


Fig. 22.2 Typical six-carbon sugars.

*Although the D and L designations are used in this chapter, readers should be aware that in the Cahn-Ingold-Prelog system, a series of rules determine configurations. In this new system, the symbols *R* and *S* are used to designate configurations.

The formula for glucose can be written in the form of aldehyde or enol, a short-lived reactive species. Shift to the enol anion is favored in alkaline solution, as follows:



The presence of a double bond and a negative charge in the enol anion makes glucose an active reducing substance that is oxidized by relatively mild oxidizing agents, such as cupric and ferric ions. Glucose in hot alkaline solution readily reduces cupric ions to cuprous ions. The color change has been used as a presumptive indication for the presence of glucose; for many years, blood and urine glucose concentrations were measured this way. Some other sugars also reduce cupric ions in alkaline solution; these are collectively referred to as *reducing sugars.*

The aldehyde group reacts with the hydroxyl group on carbon 5, represented by a symmetrical ring structure and depicted by the Haworth formula, in which glucose is considered as having the same basic structure as pyran (Fig. 22.3). In this formula, the plane of the ring is considered to be perpendicular to the plane of the paper, with the heavy lines pointing toward the reader. Hydroxyl groups in position 1 are then below the plane (α-configuration) or above the plane (β-configuration). A six-member ring sugar containing five carbons and one oxygen is a derivative of pyran and is called a *pyranose.* When linkage occurs with formation of a five-member ring containing four carbons and one oxygen, the sugar has the same basic structure as furan and is called a *furanose.* Fructose is shown in two cyclical forms. Fructopyranose is the configuration of the free sugar, and fructofuranose occurs whenever fructose exists in combination with disaccharides and polysaccharides, as in sucrose and inulin.

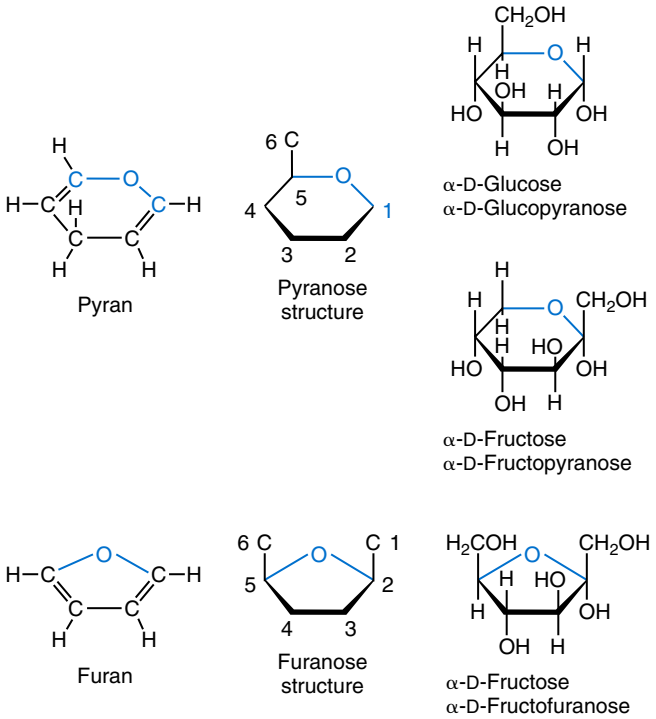


Fig. 22.3 The Haworth formula for sugars.

Disaccharides

Two monosaccharides join covalently by an *O*-glycosidic bond, with loss of a molecule of water, to form a disaccharide. The chemical bond between the sugars always involves the aldehyde or ketone group of one monosaccharide joined to an alcohol group (e.g., maltose) or an aldehyde or ketone group (e.g., sucrose) of the other monosaccharide (Fig. 22.4). The most common disaccharides are as follows:

Maltose = glucose + glucose
Lactose = glucose + galactose
Sucrose = glucose + fructose

If the linkage between two monosaccharides is between the aldehyde or ketone group of one molecule and a hydroxyl group of another molecule (as in maltose and lactose), one potentially free ketone or aldehyde group remains on the second monosaccharide. Consequently, the second glucose residue can be oxidized and is capable of existing in α- or β-pyranose form. Thus the disaccharide is a reducing sugar, but its reducing power is only approximately 40% of

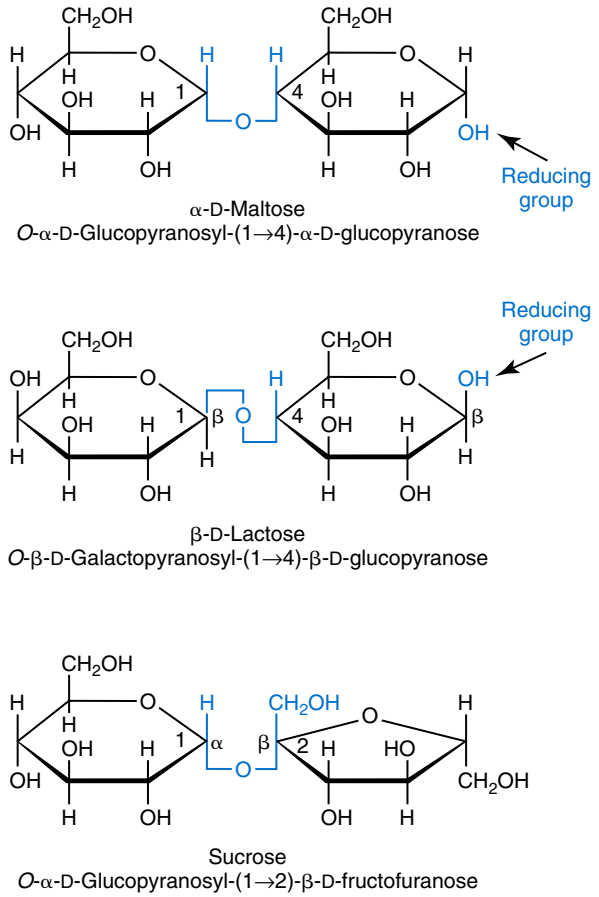


Fig. 22.4 Structural formulas of disaccharides.

the reducing power of the two single monosaccharides added together. Alternatively, if the linkage between two monosaccharides involves the aldehyde or ketone groups of both molecules (as in sucrose), a nonreducing sugar is formed because no free aldehyde or ketone group remains.

Polysaccharides

The linkage of multiple monosaccharide units results in the formation of polysaccharides. The major storage carbohydrates are starch in plants and glycogen in animals, both of which form granules inside cells. Polysaccharides can provide structural support. Cellulose is used by plants, whereas chitin is the principal component of the exoskeleton of arthropods (insects and crustacea).

Starch and Glycogen

Most starches are composed of a mixture of amyloses and amylopectins. Amylose consists of one long unbranched chain of glucose units linked together by α-1,4-linkages, with only the terminal aldehyde group free. In amylopectin, most of the units are joined by α-1,4-links, but α-1,6-glycosidic bonds also exist every 24 to 30 residues, producing sidechains. Amylopectin contains up to 1 million glucose residues. The structure of **glycogen** is similar to that of amylopectin, but branching is more extensive and is evident every 8 to 12 glucose residues. These branches enhance the solubility of glycogen and allow the glucose residues to be more readily mobilized. Glycogen is most abundant in the liver and is also found in skeletal muscle. The difference in structure between amylose and amylopectin is important in selection of the appropriate starch substrate for amylase determinations (see Chapter 19). The rate of hydrolysis is affected by structural differences in the starch.

Cellulose

Cellulose is an important structural polysaccharide in plants. It is an unbranched polymer of glucose residues joined by β-1,4-linkages. The β-configuration facilitates the formation of long straight chains, producing fibers of high tensile strength. The β-1,4-linkages are not hydrolyzed by α-amylases. Because humans do not have cellulases, they are unable to digest vegetable fiber.

Glycoproteins

Many integral membrane proteins have oligosaccharides covalently attached to the extracellular region, forming glycoproteins. In addition, most proteins that are secreted, such as antibodies, hormones, and coagulation factors, are glycoproteins. The number of attached carbohydrate residues varies among proteins and constitutes 1% to 70% of the weight of the glycoprotein. The oligosaccharides are attached by *O*-glycosidic linkages to the sidechain oxygen of serine or threonine residues or by *N*-glycosidic linkages to the sidechain nitrogen of asparagine residues.

One of the biological functions of the carbohydrate chains is to regulate the life span of proteins. Loss of sialic acid residues from the ends of oligosaccharide chains on erythrocytes results in the removal of red blood cells from the circulation. Carbohydrates have also been implicated in cell-cell recognition, in secretion, and in targeting of proteins to specific subcellular domains.

BIOCHEMISTRY AND PHYSIOLOGY

Glucose is the primary energy source for the human body. After absorption (see Chapter 38), the metabolism of all hexoses proceeds according to the body's requirements. This metabolism results in (1) energy production by conversion to carbon dioxide and water, (2) storage as glycogen in the liver or as triglyceride in adipose tissue, or (3) conversion to keto acids, amino acids, or protein.

The complete picture of intermediary metabolism of carbohydrates is complex and is interwoven with the metabolism of lipids and amino acids. For details, readers should consult a biochemistry textbook.

Regulation of Blood Glucose Concentration

The concentration of glucose in the blood is regulated by a complex interplay of multiple pathways, modulated by several hormones. *Glycogenesis* is the name for the conversion of glucose to glycogen, the most important storage polysaccharide in liver and muscle. The reverse process, namely, the breakdown of glycogen to glucose and other intermediate products, is termed *glycogenolysis*. The formation of glucose from noncarbohydrate sources, such as amino acids, glycerol, or lactate, is termed *gluconeogenesis*. The conversion of glucose or other hexoses into lactate or pyruvate is called *glycolysis*. Further oxidation to carbon dioxide and water occurs through the Krebs (citric acid) cycle and the mitochondrial electron transport chain coupled to oxidative phosphorylation, which generates the adenosine triphosphate that provides chemical energy for many bodily processes. Oxidation of glucose to carbon dioxide and water also occurs through the hexose monophosphate shunt pathway, which produces the reduced form of nicotinamide-adenine dinucleotide phosphate.

Hypoglycemia

Hypoglycemia is a blood glucose concentration below the fasting value, but the definition of a specific limit is difficult. The most widely used cutoff is 50 mg/dL (2.8 mmol/L), but some authors suggest 60 mg/dL (3.3 mmol/L). A transient decline may occur 1.5 to 2 hours after a meal, and it is not uncommon for plasma glucose concentration as low as 50 mg/dL (2.8 mmol/L) to be observed 2 hours after ingestion of an oral glucose load. Similarly, extremely low fasting blood glucose values may be occasionally noted without symptoms or evidence of underlying disease. Hypoglycemia is rare in persons who do not have drug-treated diabetes mellitus.

Symptoms of hypoglycemia vary among individuals, and none is specific. Epinephrine produces the classic signs and symptoms of hypoglycemia, namely, trembling, sweating, nausea, rapid pulse, light-headedness, hunger, and epigastric discomfort. These autonomic symptoms are nonspecific and

may be noted in other conditions, such as hyperthyroidism, pheochromocytoma, or even anxiety. Although controversial, some investigators have proposed that a rapid decrease in blood glucose concentrations may trigger the symptoms even though the blood glucose itself may not reach hypoglycemic values, whereas gradual onset to a similar glucose concentration may not produce symptoms.

The brain is completely dependent on blood glucose for energy production under physiological conditions, and approximately half of the glucose used in resting adults is in the central nervous system (CNS). Very low concentrations of plasma glucose (<20 or 30 mg/dL; 1.1 mmol/L or 1.7 mmol/L) cause severe CNS dysfunction. During prolonged fasting or hypoglycemia, ketones may be used as an energy source. The broad spectrum of symptoms and signs of CNS dysfunction range from headache, confusion, blurred vision, and dizziness, to seizures, loss of consciousness, and even death; these symptoms are known as *neuroglycopenia*. Restoration of plasma glucose usually produces prompt recovery, but irreversible damage may occur.

The age of onset of hypoglycemia is a convenient way to classify the disorder, but some overlap occurs among the various groups. For example, some glycogen storage disorders may arise in the third decade of life, and hormone deficiencies may occur in childhood.

Hypoglycemia in Neonates and Infants

Neonatal blood glucose concentrations are much lower than those of adults (mean, ~35 mg/dL; 1.9 mmol/L) and decline shortly after birth when liver glycogen stores are depleted. Glucose concentrations as low as 30 mg/dL (1.7 mmol/L) in a term infant and 20 mg/dL (1.1 mmol/L) in a premature infant may occur without clinical evidence of hypoglycemia. The more common causes of hypoglycemia in the neonatal period include prematurity, maternal diabetes, gestational diabetes mellitus (GDM), and maternal toxemia/preeclampsia. These are usually transient. Hypoglycemia with onset in early infancy is usually less transitory and may be due to inborn errors of metabolism or ketotic hypoglycemia; this type of hypoglycemia usually develops after fasting or a febrile illness.

Fasting Hypoglycemia in Adults

Hypoglycemia results from a decreased rate of hepatic glucose production or an increased rate of glucose use. Symptoms suggestive of hypoglycemia are fairly common, but hypoglycemic disorders are rare. However, true hypoglycemia usually indicates serious underlying disease and may be life threatening. An exact threshold for the establishment of hypoglycemia is not always possible, and values as low as 30 mg/dL (1.7 mmol/L) may be encountered in healthy, premenopausal women during the classic test—a 72-hour fast. Symptoms usually begin at plasma glucose concentrations below 55 mg/dL (3 mmol/L), and impairment of cerebral function begins when glucose is less than 50 mg/dL (2.8 mmol/L).

The 72-hour fast should be conducted in a hospital. During the fast, the patient should be allowed a liberal intake of calorie-free and caffeine-free fluids. Samples typically are drawn every 6 hours for analysis of plasma glucose, insulin, C-peptide, and proinsulin. When plasma glucose concentration is 60 mg/dL (3.3 mmol/L) or less, analyses are performed every 1 to 2 hours. The fast should be concluded when plasma glucose concentration falls to a predetermined concentration (such as 45 mg/dL [2.5 mmol/L] or less) or the patient exhibits signs or symptoms of hypoglycemia, or after 72 hours. Most patients with true hypoglycemia show an abnormally low value within 12 hours of beginning a fast. Women exhibit significantly lower glucose concentrations than men. Low plasma glucose alone is not sufficient to establish the diagnosis, and the absence of signs or symptoms of hypoglycemia during the fast excludes the diagnosis of a hypoglycemic disorder as the cause of such symptoms.

More than 100 causes of hypoglycemia have been reported. Drugs are the most prevalent cause, and many widely used drugs, including propranolol, salicylates, and disopyramide, produce hypoglycemia. Oral hypoglycemic agents, which have long half-lives (35 hours for chlorpropamide), are the most frequent cause of drug-induced hypoglycemia and may be directly measured in blood or urine. Surreptitious administration of insulin is detected by the discovery of low C-peptide concentrations with increased insulin concentrations.

Ethanol produces hypoglycemia by inhibiting gluconeogenesis, and this inhibition is aggravated by malnutrition (low glycogen stores) in individuals with chronic alcoholism. Individuals with hepatic failure (e.g., viral hepatitis, ingestions of toxins) have impaired gluconeogenesis or glycogen storage, which may result in hypoglycemia. Decreased hepatic glucose production requires dysfunction of more than 80% of the liver. Deficiencies of growth hormone (especially with coexistent adrenocorticotropic hormone [ACTH] deficiency), glucocorticoids, thyroid hormone, or glucagon may also produce hypoglycemia. Although a deficiency of glucocorticoids (e.g., Addison disease) is most consistently associated with hypoglycemia, most glucocorticoid-deficient adults are not hypoglycemic. Hormonal deficiency causes hypoglycemia in children more frequently than in adults.

Demonstration of a low plasma glucose concentration in the presence of a high plasma insulin value is highly suggestive of an insulin-producing pancreatic islet cell tumor. Because healthy people exhibit a wide range of insulin concentrations, absolute hyperinsulinemia occurs in less than 50% of individuals with insulinomas. Serum insulin concentrations inappropriately high for concurrent plasma glucose values denote autonomous insulin secretion. Provocative tests (glucagon, tolbutamide, or calcium) or suppression tests (infusion of insulin and measurement of C-peptide), although strongly recommended in the past, generally are not necessary.

Nonpancreatic neoplasms that cause hypoglycemia are rare, but when present, they are often extremely large mesenchymal neoplasms. No single mechanism explains the hypoglycemia in all cases, but secretion of a precursor of insulin-like growth factor 2 is the most commonly encountered finding. Glucose utilization is increased, especially in muscle, presumably reflecting activity of the hormone.

Hypoglycemia caused by septicemia should be relatively easy to diagnose. The mechanism is not well defined, but depleted glycogen stores, impaired gluconeogenesis, and increased peripheral use of glucose may all be contributing factors. Glucose tolerance is commonly depressed in individuals with renal disease, and hypoglycemia may occur in those with end-stage renal failure.

Some of the conditions that produce fasting hypoglycemia are readily apparent, but others require a lengthy diagnostic workup. Once fasting hypoglycemia is demonstrated, specific tests should be performed to establish the underlying cause. The oral glucose tolerance test (OGTT) is not an appropriate study for evaluation of a patient suspected of having hypoglycemia.

Postprandial Hypoglycemia

Drugs, antibodies to insulin or the insulin receptor, inborn errors (e.g., fructose-1,6-diphosphatase deficiency), and *reactive hypoglycemia* (also referred to as *functional hypoglycemia*) produce hypoglycemia in the postprandial (fed) state. It has been proposed that for individuals with vague symptoms after food ingestion, the preferred terminology should be *idiopathic reactive hypoglycemia* or *idiopathic postprandial syndrome.*

At the Third International Symposium on Hypoglycemia, reactive hypoglycemia was defined as a "clinical disorder in which the patient has postprandial symptoms suggesting hypoglycemia that occur in everyday life and are accompanied by a blood glucose concentration less than 45 to 50 mg/dL (2.5–2.8 mmol/L) as determined by a specific glucose measurement on arterialized venous or capillary blood, respectively." Patients complain of autonomic symptoms that occur approximately 1 to 3 hours after eating and seem to obtain relief, lasting 30 to 45 minutes, by food intake. These symptoms are rarely due to low blood glucose concentrations. A 5- or 6-hour glucose tolerance test had been the standard procedure to establish the presence of postprandial hypoglycemia, but that has been discredited. Consequently, an *OGTT should not be used in the diagnosis of reactive hypoglycemia.*

Postprandial hypoglycemia is infrequent, and demonstration of hypoglycemia during spontaneously occurring symptomatic episodes is necessary to establish the diagnosis. If this is not possible, a 5-hour meal tolerance test (which simulates the composition of a normal diet) or a "hyperglucidic" (high-glucose) breakfast test has been proposed.

The diagnosis of hypoglycemia has also been used to explain a wide variety of disorders that appear unrelated to blood glucose abnormalities. These nonspecific symptoms include fatigue, muscle spasms, palpitations, numbness, tingling, pain, sweating, mental dullness, sleepiness, weakness, and fainting. Behavioral abnormalities, poor school performance, and delinquency have been incorrectly attributed to low blood glucose concentrations. A diagnosis of hypoglycemia should not be made unless a patient meets the criteria of *Whipple's triad of low blood glucose concentration:* (1) symptoms known or likely to be caused by hypoglycemia, (2) low glucose measured when symptoms occur, and (3) relief of symptoms when glucose is increased to normal. Demonstration of normal plasma glucose concentration when the patient exhibits symptoms excludes the possibility of a hypoglycemic disorder.

Hypoglycemia in Diabetes Mellitus

Hypoglycemia occurs frequently in individuals with type 1 or type 2 diabetes. Patients using insulin experience approximately one to two episodes of symptomatic hypoglycemia per week, and severe hypoglycemia that requires assistance from others or is associated with loss of consciousness affects about 10% of this population per year. In patients practicing intensive insulin therapy (e.g., multiple injections, continuous subcutaneous insulin infusion), these figures are increased twofold to sixfold. The chief adverse event associated with intensive therapy in the Diabetes Control and Complications Trial (DCCT) was a threefold increase in the incidence of severe hypoglycemia. Similarly, hypoglycemia occurs in patients with type 2 diabetes (caused by oral hypoglycemic agents or insulin) but is less frequent than in type 1 diabetes. Defective glucose counterregulation (ability to increase glucose, counter to the effect of insulin) and ignorance of hypoglycemia are two pathophysiological mechanisms that contribute to hypoglycemia in patients with diabetes.

Defective Glucose Counterregulation Counterregulatory responses become impaired in patients with type 1 diabetes, increasing the risk of hypoglycemia. The secretion of glucagon in response to hypoglycemia is impaired by an unknown mechanism early in the course of type 1 diabetes. The secretory response of epinephrine to hypoglycemia becomes deficient later in the course of the disease. These defects are selective because other stimuli continue to elicit glucagon and epinephrine secretion. Glucose counterregulation does not appear to be notably defective in patients with type 2 diabetes.

Hypoglycemia Unawareness Up to 50% of patients with long-standing (longer than 30 years) type 1 diabetes do not experience neurogenic warning symptoms and are prone to more severe hypoglycemia. The mechanism is thought to be associated with a decreased epinephrine response to hypoglycemia. Intensively treated patients with type 1 diabetes require lower plasma glucose concentrations to elicit symptoms of hypoglycemia. Some authors have claimed that therapeutic use of human insulin rather than other insulins results in an increased incidence of ignorance of hypoglycemia, but analysis of 45 studies revealed no significant differences in hypoglycemic episodes among insulin species.

Lactate and Pyruvate

Lactic acid, an intermediary in carbohydrate metabolism, is predominantly derived from white skeletal muscle, brain, skin, renal medulla, and erythrocytes. The blood lactate concentration depends on the rate of production in these tissues and the rate of metabolism in the liver and kidneys. The liver uses approximately 65% (75 g/d) of the total basal lactate produced, predominantly in gluconeogenesis. The **Cori cycle** is the conversion of glucose to **lactate** in the periphery and reconversion of lactate to glucose in the liver. Extrahepatic

removal of lactate occurs by oxidation in red skeletal muscle and the renal cortex. A moderate increase in lactate production results in increased hepatic lactate clearance, but uptake by the liver is saturable when concentrations exceed 18 mg/dL (2 mmol/L). During strenuous exercise, for example, lactate concentrations may increase significantly, from an average concentration of about 8 mg/dL (0.9 mmol/L) to more than 180 mg/dL (20 mmol/L) within 10 seconds. No concentration of lactate is uniformly accepted for the diagnosis of lactic acidosis, but lactate concentrations exceeding 45 mg/dL (5 mmol/L) with pH less than 7.25 indicate significant lactic acidosis.

Measurement of **pyruvate** is useful in the evaluation of patients with inborn errors of metabolism who have increased serum lactate concentrations. A lactate-to-pyruvate ratio of less than 25 suggests a defect in gluconeogenesis, whereas an increased ratio (≥35) indicates reduced intracellular conditions found in hypoxia. Inborn errors associated with an increased lactate-to-pyruvate ratio include pyruvate carboxylase deficiency and defects in oxidative phosphorylation. Pyruvate is also measured in clinical studies evaluating reperfusion after myocardial ischemia.

Lactic Acidosis

Lactic acidosis occurs in two clinical settings: (1) type A (hypoxic), associated with decreased tissue oxygenation, such as shock, hypovolemia, and left ventricular failure; and (2) type B (metabolic), associated with disease (e.g., diabetes mellitus, neoplasia, liver disease), drugs/toxins (e.g., ethanol, methanol, salicylates), or inborn errors of metabolism. Lactic acidosis is not uncommon and occurs in approximately 1% of those admitted to the hospital. It has a mortality rate greater than 60%, which approaches 100% if hypotension is also present. Type A is much more common.

The mechanism of type B lactic acidosis is not known but is speculated to be a primary defect in mitochondrial function with impaired oxygen use. This leads to reduced stores of adenosine triphosphate (ATP) and nicotinamide adenine dinucleotide (NAD^+), with accumulation of NADH and H^+. In the presence of decreased liver perfusion or liver disease, lactate removal from the blood is reduced, thereby aggravating the lactic acidosis.

An uncommon but often undiagnosed cause of lactic acidosis is D-lactic acidosis. Absorption and accumulation of D-lactate from abnormal intestinal bacteria may cause systemic acidosis. This condition occurs after jejunoileal bypass surgery and manifests as altered mental status (from mild drowsiness to coma) with increased blood concentrations of D-lactate. Virtually all the commonly used laboratory assays for lactate use L-lactate dehydrogenase, which does not detect D-lactate. D-Lactate can be measured by gas-liquid chromatography or, enzymatically, with a specific D-lactate dehydrogenase.

Lactate in cerebrospinal fluid (CSF) normally parallels blood concentrations. With biochemical alterations in the CNS, however, CSF lactate values change independently of blood values. Increased CSF concentrations are noted in individuals with cerebrovascular accidents, intracranial hemorrhage, bacterial meningitis, epilepsy, and other CNS disorders. In aseptic (viral) meningitis, lactate concentrations in CSF are not usually increased; hence, CSF lactate has been used to help discriminate between viral and bacterial meningitis, but the clinical utility has been questioned. In a few children with inherited metabolic disease, CSF lactate concentrations may be increased despite plasma lactate in the reference interval.

Inborn Errors of Carbohydrate Metabolism

Deficiency or absence of an enzyme that participates in carbohydrate metabolism may result in accumulation of monosaccharides, which can be measured in the urine. Most of these conditions are inherited as autosomal recessive traits. Sugars frequently appear in the urine as a result of excessive consumption, without the presence of underlying disease.

Techniques used to separate and identify sugars have included fermentation, optical rotation, osazone formation with phenylhydrazine, specific chemical tests, and paper or thin-layer chromatography. The availability of glucose oxidase test strips, specific for glucose, has greatly simplified the differentiation of glucose from other reducing substances. For practical purposes, the urinary sugars of clinical interest are glucose and galactose. Urine from infants and children should be tested by both the glucose oxidase and copper reduction tests to identify individuals with inborn errors of metabolism. Reducing substances other than glucose should be further identified by chromatographic procedures.

Glycogen Storage Disease

Glycogen, although present in most tissues, is stored predominantly in the liver and skeletal muscle. During fasting, liver glycogen is converted to glucose to provide energy for the whole body. In contrast, skeletal muscle lacks glucose-6-phosphatase, and muscle glycogen can be used only locally for energy. Glycogen storage disease is a generic name encompassing at least 12 rare inherited disorders of glycogen storage in tissues. The different forms of glycogen storage disease are categorized by numerical type in the chronological sequence in which these defects were identified. Each form is due to a deficiency of a specific enzyme in glycogen metabolism, producing a quantitative or a qualitative defect of glycogen storage. Numerous mutations have been identified in patients with these conditions.

Because both the liver and skeletal muscle have the highest rates of glycogen metabolism, these tissues are most affected. The liver forms (types I, III, IV, VI, and IX), which comprise approximately 80% of the total, are marked by hepatomegaly (due to increased liver glycogen stores) and hypoglycemia (caused by inability to convert glycogen to glucose). Hypoglycemia is manifested by autonomic clinical symptoms (sweating, shakiness, and a light-headed feeling), growth retardation, and laboratory findings of decreased insulin and increased glucagon concentrations in the blood. The muscle forms (types II, V, and VII), in contrast, have mild symptoms

that usually appear in young adulthood during strenuous exercise as a result of the inability to provide energy for muscle contraction. Other muscle disorders may exhibit similar symptoms but are readily differentiated through evaluation of glycogen stores. The specific diagnosis of each type is made directly by demonstrating the enzyme defect in tissue.

ANALYTICAL METHODOLOGY

Analytical methods for measuring glucose, ketone bodies, lactate, and pyruvate are discussed in this section. Methods for measuring insulin, proinsulin, and glucagon are discussed in Chapter 33.

Measurement of Glucose in Body Fluids

Several methods are used to measure glucose in blood, serum, plasma, and urine. The College of American Pathologists (CAP) surveys demonstrate that all methods exhibit a coefficient of variation (CV) among laboratories that is ≤2.6% for glucose values on lyophilized serum.

Specimen Collection and Storage

In individuals with a normal hematocrit, fasting whole-blood glucose concentration is approximately 10% to 12% lower than plasma glucose. Although the glucose concentrations in the water phase of red blood cells and plasma are similar (the erythrocyte plasma membrane is freely permeable to glucose), the water content of plasma (93%) is approximately 11% greater than that of whole blood at a normal hematocrit. In most clinical laboratories, plasma or serum is used for most glucose determinations; methods for self-monitoring of glucose use whole-blood samples but may measure the glucose concentration in the plasma phase. During fasting, capillary blood glucose concentration is only 2 to 5 mg/dL (0.1 mmol/L to 0.3 mmol/L) higher than that of venous blood. After a glucose load, however, capillary blood glucose concentrations are 20 to 70 mg/dL (1.0 to 3.9 mmol/L; mean, ≈30 mg/dL [1.7 mmol/L]; equivalent to 20% to 25%) higher than the concentrations in concurrently drawn venous blood samples, probably due to glucose consumption in the tissues.

Glycolysis decreases serum glucose by approximately 5% to 7% in 1 hour (5 to 10 mg/dL; 0.3 mmol/L to 0.6 mmol/L) in normal uncentrifuged coagulated blood at room temperature. The rate of in vitro glycolysis is higher in the presence of leukocytosis or bacterial contamination. In separated, nonhemolyzed sterile serum, the glucose concentration is generally stable as long as 8 hours at 25°C and up to 72 hours at 4°C; variable stability is observed with longer storage periods. Plasma, removed from the cells after moderate centrifugation, contains leukocytes that also metabolize glucose—although cell-free sterile plasma has no glycolytic activity.

Sodium fluoride or, less commonly, sodium iodoacetate, is used to inhibit glycolysis. Fluoride ions prevent glycolysis by inhibiting enolase, an enzyme that requires Mg^{2+}. Inhibition is due to the formation of an ionic complex consisting of Mg^{2+}, inorganic phosphate, and fluoride ions; this complex interferes with the interaction of enzyme and substrate. Fluoride is also a weak anticoagulant because it binds calcium; however, clotting may occur after several hours, and it is therefore advisable to use a *combined fluoride-oxalate mixture,* such as 2 mg of potassium oxalate ($K_2C_2O_4$) and 2 mg of NaF/mL of blood, to prevent late clotting. Other anticoagulants (e.g., ethylenediaminetetraacetic acid, citrate, heparin) can also be used. Fluoride ions in high concentration inhibit the activity of urease and certain other enzymes; consequently the specimens may be unsuitable for determination of urea in procedures that require urease and for direct assay of some serum enzymes. $K_2C_2O_4$ causes loss of cell water, thereby diluting the plasma. Samples collected in these tubes therefore should not be used for measurement of other analytes.

Although fluoride has been widely used to inhibit glycolysis, the rate of decline in the first 1 to 2 hours after sample collection is not altered, and glycolysis may continue at a slower rate for up to 4 hours. Acidification of blood using citrate buffer inhibits in vitro glycolysis more effectively than fluoride.

To minimize glycolysis, the cells should be removed within minutes. Alternatively, the tube should be placed in an ice-water slurry and the cells separated within 30 minutes. Neither of these approaches is practical in routine analysis. It may not be necessary in routine analysis to use a fluoride-containing tube if plasma is separated from cells or if glucose is measured within 30 minutes of blood collection. However, inhibitors of glycolysis are necessary in patients with greatly increased leukocyte counts because differences of up to 65 mg/dL (3.6 mmol/L) have been observed between glucose values with and without glycolytic inhibitors after 1 to 2 hours of contact with the blood cells.

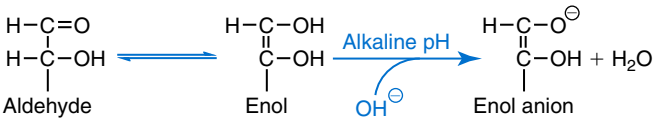
CSF may be contaminated with bacteria or other cells and should be analyzed immediately for glucose. If a delay in measurement is unavoidable, the sample should be centrifuged and stored at 4°C or −20°C.

In 24-hour collections of urine, glucose may be preserved by adding 5 mL of glacial acetic acid to the container before starting the collection. The final pH of the urine is usually between 4 and 5, which inhibits bacterial activity. Other preservatives that have been proposed include 5 g of sodium benzoate per 24-hour specimen, or chlorhexidine and 0.1% sodium nitrate with 0.01% benzethonium chloride. These may be inadequate, and urine should be stored at 4°C during collection. Urine samples may lose as much as 40% of their glucose after 24 hours at room temperature.

Measurement of Glucose in Blood

Hexokinase (HK), or glucose oxidase, is widely used in assays to measure the concentration of glucose in body fluids.

Hexokinase Methods. HK methods are based on a coupled enzyme assay that uses HK and glucose-6-phosphate dehydrogenase:



As indicated, glucose is first phosphorylated by ATP in the presence of HK and Mg^{2+}. The glucose-6-phosphate formed is oxidized by G6PD to 6-phosphogluconate in the presence of $NADP^+$ or NAD^+. The amount of reduced NADP (NADPH) or NADH produced is directly proportional to the amount of glucose in the sample and is measured by the increase in absorbance at 340 nm. $NADP^+$ is the cofactor when G6PD derived from yeast is used in the assay; NAD^+ is the cofactor when bacterial *(Leuconostoc mesenteroides)* G6PD is used. A reference method based on this principle has been developed and validated. In the reference method, serum or plasma is deproteinated by the addition of solutions of barium hydroxide and zinc sulfate. The clear supernatant is mixed with a reagent containing ATP, NAD^+, HK, and G6PD, incubated at 25°C until the reaction is complete, and NADH is measured. Calibrators and blanks are carried through the entire procedure, including the deproteination step.

Although highly accurate and precise, the reference method is too exacting and time-consuming for routine use in a clinical laboratory. An alternative approach is to apply the reaction directly to serum or plasma and use a specimen blank to correct for interfering substances that absorb at 340 nm.

Serum or plasma may be used. Sodium fluoride (NaF), with an anticoagulant such as ethylenediaminetetraacetic acid (EDTA), heparin, oxalate, or citrate, may be used. Hemolyzed specimens containing more than 0.5 g of hemoglobin/dL are unsatisfactory because phosphate esters and enzymes released from red blood cells interfere with the assay. Other sources of interference include drugs, bilirubin, and lipemia (triglycerides of 500 mg/dL [~5.7 mmol/L] or greater causing a positive interference).

Absorbances of sample or calibrator reaction mixtures are measured after the reactions have continued to completion (steady-state reaction, "end-point" method) or at a fixed time after initiation of the reaction (fixed-time kinetics). In the steady state, end-point methods for glucose concentrations may be calculated directly, based on the molar absorptivity of NADPH or NADH, but inclusion of a set of calibrators is recommended to detect possible deterioration of enzymes, ATP, $NADP^+$, or NAD^+—all of which are unstable. Reagents may also contain substances that react with the coenzymes. The presence of these substances is evaluated by measurement of the increase in absorbance observed in a reagent blank. The highest calibrator provides a check on the linearity of the response and the adequacy of the enzyme reagent. The procedures typically show a linear relation between absorbance and glucose concentrations of 0 to 500 mg/dL (0 to 28 mmol/L). Serum or plasma samples with glucose concentrations that exceed 500 mg/dL (28 mmol/L) should be diluted (usually with isotonic saline) and reassayed.

Also available are HK procedures in which indicator reactions produce colored products, enabling absorbance measurements in the visible range. An oxidation reduction system containing phenazine methosulfate and a substituted tetrazolium compound, 2-(*p*-iodophenyl)-3-*p*-nitrophenyl-5-phenyltetrazolium chloride (INT), is reacted with NADPH formed in the reaction. The reduced INT is colored, with maximal absorbance at 520 nm.

Glucose Oxidase Methods. Glucose oxidase catalyzes the oxidation of glucose to gluconic acid and hydrogen peroxide:

$$\text{Glucose} + H_2O + O_2 \xrightarrow[\text{Oxidase}]{\text{Glucose}} \text{Gluconic acid} + 2H_2O_2$$

Addition of the enzyme peroxidase and a chromogenic oxygen acceptor, such as *o*-dianisidine, results in the formation of a colored compound that is measured:

$$o\text{–Dianisidine} + H_2O_2 \xrightarrow{\text{Peroxidase}} \text{Oxidized } o\text{–Dianisidine} + H_2O \ \text{(Colored)}$$

Glucose oxidase is highly specific for β-D-glucose. Because 36% and 64% of glucose in solution is in the α- and β-forms, respectively, complete reaction requires mutarotation of the α-form to the β-form. Some commercial preparations of glucose oxidase contain an enzyme, mutarotase, which accelerates this reaction. Otherwise, extended incubation time allows spontaneous conversion.

The second step, involving peroxidase, is much less specific than the glucose oxidase reaction. Various substances, such as uric acid, ascorbic acid, bilirubin, hemoglobin, tetracycline, and glutathione, inhibit the reaction (presumably by competing with the chromogen for H_2O_2), producing lower values. Some glucose oxidase preparations contain catalase as a contaminant; catalase activity decomposes peroxide and decreases the intensity of the final color obtained. Calibrators and unknowns should be simultaneously analyzed under conditions in which the rate of oxidation is proportional to the glucose concentration.

Glucose oxidase methods are suitable for measurement of glucose in CSF. Urine, however, contains high concentrations of substances (such as uric acid) that interfere with the peroxidase reaction, producing falsely low results. Therefore, the glucose oxidase method should not be used for urine. A method in which the urine is first pretreated with an ion exchange resin to remove interfering substances has been described.

Some instruments use a polarographic oxygen electrode that measures the rate of oxygen consumption after the sample is added to a solution containing glucose oxidase. Because this measurement involves only the glucose oxidase reaction, interferences encountered in the peroxidase step are eliminated. To prevent formation of oxygen from H_2O_2 by catalase present in some preparations of glucose oxidase, H_2O_2 is removed by two additional reactions:

$$H_2O_2 + C_2H_5OH \xrightarrow{\text{Catalase}} CH_3CHO + 2H_2O$$

$$H_2O_2 + 2H^+ + 2I^- \xrightarrow{\text{Molybdate}} I_2 + 2H_2O$$

The latter reaction is effective even when catalase activity has diminished on storage of reagents. The procedure has been applied directly to urine, serum, plasma, or CSF. However, this approach should not be used for determination of glucose in whole blood because blood cells consume oxygen.

In dry, multilayer, slide automated systems, glucose is measured by a glucose oxidase procedure. A 10-μL sample of serum, plasma, urine, or CSF is placed on a porous film

on top of the layer containing the reagents. Glucose diffuses through the film and reacts with the reagents to produce a colored end product or dye. The intensity of this dye is measured through a lower transparent film by reflectance spectrophotometry. Advantages include small sample size, no liquid reagents, and improved stability on storage.

Glucose Dehydrogenase Method. The enzyme glucose dehydrogenase (β-D-glucose: NAD oxidoreductase, EC 1.1.1.47) catalyzes the oxidation of glucose to gluconolactone with concomitant reduction of NAD^+ to NADH. Mutarotase is added to shorten the time necessary to reach equilibrium. The amount of NADH generated is proportional to the glucose concentration. The reaction appears to be highly specific for glucose, shows no interference from common anticoagulants and substances normally found in serum, and provides results in close agreement with HK procedures. The glucose dehydrogenase procedure is not widely used in the United States, except in a glucose meter.

Measurement of Glucose in Urine

Examination of urine for glucose is rapid, inexpensive, and noninvasive and has been used to screen large numbers of samples. Monitoring of urine glucose lacks sensitivity and specificity and provides no information about blood glucose concentrations below the renal threshold (usually 180 mg/dL; 10 mmol/L). Older screening tests detect all sugars that reduce copper. Unfortunately, these tests also react with reducing substances other than glucose. Qualitative, quantitative, and semiquantitative methods are widely available for measuring glucose in urine and have essentially replaced the nonspecific tests in adults. Note that a reducing sugar method other than an enzymatic method for glucose must be used when neonates or infants are screened for inborn errors of metabolism that result in the appearance of reducing sugars other than glucose (e.g., galactose, fructose) in the urine.

Qualitative Method for Measurement of Total Reducing Substances. Benedict qualitative reagent contains cupric ion complexed to citrate in alkaline solution. Reducing substances convert cupric to cuprous ions, forming yellow cuprous hydroxide or red cuprous oxide. A convenient adaptation of the procedure marketed in tablet form (Clinitest) has been discontinued.

Semiquantitative Measurement of Glucose in Urine. Convenient paper test strips are commercially available from several manufacturers. Examples are Clinistix, Diastix, and Chemstrip. All use the glucose-specific enzyme glucose oxidase in a chromogenic assay. For example, Clinistix has filter paper impregnated with glucose oxidase, peroxidase, and the dye *o*-toluidine. Other dyes, such as tetramethylbenzidine, can be used. The test end of the strip is moistened with freshly voided urine and is examined after 10 seconds. A blue color develops if glucose is present at a concentration of 100 mg/dL or greater. Results are read by comparing the test color with a standard color chart. Automated urinalysis systems capable of analyzing 300 strips per hour are commercially available. The test is more sensitive for glucose than is the copper reduction test (Clinitest), which has a detection limit of 250 mg/dL (~14 mmol/L).

False-positive results may be produced by contamination of urine with H_2O_2 or a strong oxidizing agent, such as hypochlorite (bleach). False negative results may occur with large quantities of reducing substances, such as ketones, ascorbic acid, and salicylates. For routine examinations, a negative result by the strip test is usually interpreted to mean that the urine specimen is negative for glucose.

Other strip tests (such as Keto-Diastix and Diascreen 2GK) are designed for the semiquantitative estimation of both glucose and ketone bodies. The glucose portion of the strip uses the glucose oxidase–peroxidase method. The hydrogen peroxide produced oxidizes iodide to iodine, yielding various intensities of brown that correspond to increasing concentrations of glucose in the urine. The detection limit is 100 mg/dL (5.5 mmol/L).

Quantitative Methods for Determination of Glucose in Urine. Applications of various procedures for quantitative determination of glucose in urine were previously discussed in the section on the determination of glucose in body fluids. HK or glucose dehydrogenase procedures are recommended for greatest accuracy and specificity. Glucose oxidase procedures that depend only on the consumption of oxygen or the production of H_2O_2 are also reliable. Glucose oxidase procedures that include the H_2O_2-peroxidase reaction are not used for urine.

Reference Intervals

Although glucose is assayed by several different analytical procedures, reference intervals do not vary significantly among methods. The following values are representative of glucose assays:

Sample Reference Intervals for Fasting Glucose

Plasma/Serum	
Adults	74–99 mg/dL (4.1–5.5 mmol/L)
Children	60–100 mg/dL (3.5–5.5 mmol/L)
Premature neonates	20–60 mg/dL (1.1–3.3 mmol/L)
Term neonates	30–60 mg/dL (1.7–3.3 mmol/L)
Whole blood	65–95 mg/dL (3.6–5.3 mmol/L)
CSF	40–70 mg/dL (60% of plasma value) (2.2–3.9 mmol/L)
Urine	
24 h	1–15 mg/dL (0.1–0.8 mmol/L)

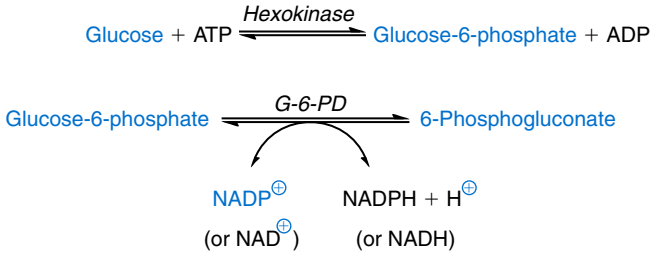
Note that the American Diabetes Association (ADA) criteria of fasting glucose of 126 mg/dL (7.0 mmol/L) or greater—not the reference interval—is used for the diagnosis of diabetes. Moreover, the threshold for diagnosis of hypoglycemia is variable and is considerably less than the lower limit of the reference interval. No sex difference exists. Plasma glucose concentrations increase with age—fasting glucose concentrations increase approximately 2 mg/dL (0.1 mmol/L) per decade; postprandial concentrations increase by 4 mg/dL (0.2 mmol/L) per decade; and concentrations after a glucose challenge increase by 8 to 13 mg/dL (0.4–0.7 mmol/L) per decade.

CSF glucose concentrations should be approximately 60% of the plasma concentrations and must always be compared with concurrently measured plasma glucose for adequate clinical interpretation.

Measurement of Lactate and Pyruvate in Body Fluids

Determination of Lactate in Whole Blood

Lactate is oxidized to pyruvate by lactate dehydrogenase in the presence of NAD^+. The NADH formed in this reaction is measured spectrophotometrically at 340 nm and serves as a measure of the lactate concentration:



The equilibrium of the reaction normally lies far to the left. However, by buffering the pH between 9.0 and 9.6, adding an excess of NAD^+, and trapping the reaction product pyruvate with hydrazine, the equilibrium can be shifted to the right. Pyruvate can also be removed through reaction with L-glutamate in the presence of alanine aminotransferase.

Because of its high specificity and simplicity, the enzymatic method is the method of choice for the measurement of lactate, although other methods may also be used (e.g., gas chromatography).

The Vitros analyzer, formerly the Ektachem, uses an assay in which lactic acid is oxidized to pyruvate by lactate oxidase. The H_2O_2 generated oxidizes a chromogen system, and the absorbance of the resulting dye complex, measured spectrophotometrically at 540 nm, is directly proportional to the lactate concentration in the specimen. Each mole of lactate that is oxidized produces 0.5 mol of dye complex.

Specimen Collection and Storage. Careful techniques are necessary to prevent changes in lactate concentration while blood is drawn and afterward. Patients should be fasting and at complete rest for at least 2 hours to allow lactate concentrations to reach steady state.

Venous specimens should be obtained without the use of a tourniquet or immediately after the tourniquet has been applied. Alternatively, the tourniquet should be removed after the puncture has been performed, and the blood should be allowed to flow for several minutes before the sample is withdrawn. Arterial blood sampling, which prevents these potential pitfalls, may also be used. Patients should avoid exercise of the hand or arm immediately before and during the procedure.

Both venous and arterial blood may be collected in heparinized syringes and immediately delivered into a premeasured amount of chilled protein precipitant, such as trichloroacetic acid, metaphosphoric acid, or perchloric acid. The clear supernatant, after centrifugation, is stable at 4°C for as long as 8 days. Meticulous attention to sample preparation is required. If blood is not preserved as directed, lactate rapidly increases in blood as a result of glycolysis. Increases may be as great as 20% within 3 minutes and 70% within 30 minutes at 25°C. Specimens collected as described in this section are also suitable for determination of pyruvate.

If a plasma specimen is required, blood should be collected in a tube containing 10 mg of NaF and 2 mg of $K_2C_2O_4$ per milliliter of blood. Ideally, the specimen should be immediately cooled, and the cells separated within 15 minutes, but reasonable stability of volunteers' lactate is seen at room temperature for 30 minutes in whole blood with NaF. Once the plasma is separated from the cells, lactate is stable.

Reference Intervals. The reference intervals for lactate in adults are as follows:

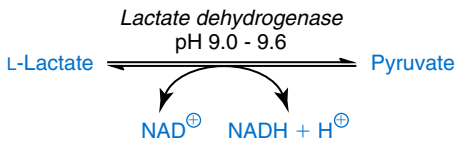
	LACTATE	
Specimen	**mmol/L**	**mg/dL**
Venous Blood		
At rest	0.3–2.0	2.7–18
Arterial Blood		
At rest	0.3–1.5	2.7–13.5

Laboratories should verify that these ranges are appropriate for use in their own settings.

Individuals in the hospital exhibit a wider range of values. Lactic acidosis occurs with blood lactate concentrations exceeding 5 mmol/L (45 mg/dL). Severe exercise dramatically increases lactate concentrations, and even movement of leg muscles by individuals at bed rest may result in significant increases. Plasma values are about 7% higher than those in whole blood, although differences depend on the procedure used. CSF values are usually similar to blood concentrations but may change independently in CNS disorders. Age-related reference intervals for CSF lactate (and lactate-to-pyruvate ratios) have been established in children. The upper limit of the reference interval (90th percentile) for CSF lactate in children in hospital from birth to 15.5 years varies continuously from 16 to 17 mg/dL (1.78 to 1.88 mmol/L). Normal 24-hour urine output of lactate is 5.5 to 22 mmol/dL.

Determination of Pyruvate in Whole Blood

The reaction involved in the determination of pyruvate is essentially the reverse of the reaction used in the lactate procedure:



At about pH 7.5, the equilibrium constant strongly favors the reaction to the right. The method is very specific, and 2-oxoglutarate, oxaloacetate, acetoacetate, and β-hydroxybutyrate do not interfere as with colorimetric methods. Pyruvate is extremely unstable in blood, more so than lactate; immediate use of a chilled protein precipitant is recommended.

Fasting venous blood, drawn when the individual is at rest, has a pyruvate concentration of 0.3 to 0.9 mg/dL (0.03 to 0.10 mmol/L). Arterial blood contains 0.2 to 0.7 mg/

dL (0.02 to 0.08 mmol/L). Values for CSF are 0.5 to 1.7 mg/dL (0.06 to 0.19 mmol/L). Urine output of pyruvate is normally 1 mmol/dL or less. Few clinical indications warrant the measurement of blood or urine pyruvate concentrations.

Measurement of Ketones in Body Fluids

Measurement of ketones in urine and blood is widely performed in patients with diabetes for both diagnosis and monitoring of diabetic ketoacidosis (see more details in the sixth edition of the *Tietz Textbook of Clinical Chemistry and Molecular Diagnostics*). None of the commonly used methods for the detection and determination of ketone bodies in serum or urine reacts with all three ketone bodies. The Gerhardt ferric chloride test reacts with acetoacetate only. Tests using nitroprusside are at least 10 times more sensitive to acetoacetate than to acetone and give no reaction at all with β-hydroxybutyrate. Traditional tests for β-hydroxybutyrate are indirect; they require brief boiling of the urine to remove acetone and acetoacetate by evaporation (acetoacetate first breaks down spontaneously to acetone), followed by gentle oxidation of β-hydroxybutyrate to acetoacetate and acetone with peroxide, ferric ions, or dichromate. The acetoacetate thus formed may be detected with the Gerhardt test or by one of the procedures in which nitroprusside is used. Quantitative enzymatic assays for β-hydroxybutyrate that can be performed directly on blood or serum are commercially available.

The semiquantitative Acetest and Ketostix are frequently used but are insensitive to β-hydroxybutyrate. It is important to bear in mind, therefore, that a negative nitroprusside test result does not rule out ketoacidosis.

Acetest

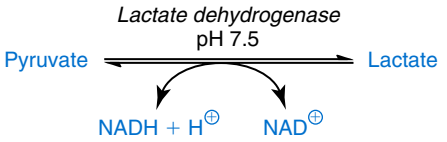
Acetest tablets contain a mixture of glycine, sodium nitroprusside, disodium phosphate, and lactose. Acetoacetate or acetone (to a lesser extent) in the presence of glycine forms a lavender-purple complex with nitroprusside. β-Hydroxybutyrate does not react with nitroprusside. Disodium phosphate provides an optimum pH for the reaction, and lactose enhances the color.

Ketostix

Ketostix is a modification of the nitroprusside test, in which a reagent strip is used instead of a tablet. The Ketostix test gives a positive reaction within 15 seconds with a specimen containing at least 50 mg of acetoacetate per liter. The color chart from the manufacturer gives readings for ketone concentrations of 50, 150, 400, 800, and 1600 mg/L. Acetone also reacts, but the test is less sensitive to it.

Determination of β-Hydroxybutyrate

In this test, β-hydroxybutyrate in the presence of NAD^+ is converted by β-hydroxybutyrate dehydrogenase to acetoacetate, producing reduced NAD (NADH). Diaphorase catalyzes the reduction of nitroblue tetrazolium (NBT) by NADH to produce a purple compound, and its absorbance is read in a special meter that provides a digital readout.



Serum β-hydroxybutyrate values vary from 0.21 to 2.81 mg/dL (0.02 to 0.27 mmol/L) in healthy people after an overnight fast. Ketone bodies in the blood can reach 20 mg/dL (2 mmol/L) with prolonged exercise. Patients with diabetic ketoacidosis usually have β-hydroxybutyrate concentrations greater than 20 mg/dL (>2 mmol/L).

Determination of Ketone Bodies in Urine

Acetest, Ketostix, Ketosis Test Strips, and Trueplus Ketone Test Strips are suitable for detecting ketone bodies in urine. The sensitivity and specificity of these tests are the same as outlined for serum.

POINTS TO REMEMBER

Disorders of Glucose Homeostasis

Hyperglycemia (increased blood glucose concentration), caused by diabetes mellitus, is the most frequently encountered disorder of carbohydrate metabolism.
Hypoglycemia is a blood glucose concentration below the fasting value.
The symptoms of hypoglycemia vary among individuals and none is specific.
Symptoms suggestive of hypoglycemia are fairly common, but true hypoglycemia is rare in individuals who do not have drug-treated diabetes mellitus.
Hypoglycemia occurs frequently in patients with diabetes and is the limiting factor in glycemic management.
True hypoglycemia usually indicates serious underlying disease.

Measurement of Glucose

Measurement of glucose is one of the most commonly performed analytical procedures.
Glucose may be measured in blood, urine, or cerebrospinal fluid.
Almost all common methods are enzymatic.
The glucose concentration in whole blood is 10%–12% lower than the plasma glucose concentration.
Glycolysis occurs in whole blood in a test tube and delay in separating cells from plasma (or serum) can significantly decrease the glucose concentration.

REVIEW QUESTIONS

1. The formation of glucose from noncarbohydrate sources occurs mostly in the liver and is referred to as:
   a. gluconeogenesis.
   b. glycogenesis.
   c. glycolysis.
   d. glycogenolysis.
2. Which of the following hormones decreases blood glucose?
   a. Epinephrine
   b. Glucagon
   c. Cortisol
   d. Insulin
3. Which anticoagulant is considered the best for serum glucose analysis because it inhibits glycolysis?
   a. Sodium oxalate
   b. EDTA
   c. Sodium fluoride
   d. Heparin
4. An example of a disaccharide is:
   a. glucose.
   b. starch.
   c. lactose.
   d. fructose.
5. The conversion of glucose into its storage form is referred to as:
   a. glycogenesis.
   b. glycolysis.
   c. glycogenolysis.
   d. glyconeogenesis.
6. Which of the following statements concerning carbohydrates is *incorrect*?
   a. Individuals diagnosed with type 1 diabetes mellitus can display hypoglycemic symptoms because of the impairment of glucagon secretion.
   b. Ethanol produces hypoglycemia by inhibiting gluconeogenesis.
   c. Monosaccharides are formed from the breakdown of starches and disaccharides within the small intestine.
   d. The brain functions normally with a low concentration of plasma glucose (<20 to 30 mg/dL).
7. The formation of 6-phosphogluconate with concomitant production of NADH is the final step in which of the following coupled-enzyme assays for glucose?
   a. Hexokinase method
   b. Glucose oxidase method
   c. Glucose dehydrogenase method
   d. Polarographic method
8. The typical cause of an inborn error of carbohydrate metabolism is:
   a. lack of insulin production.
   b. glucagonoma.
   c. absence of an enzyme involved in carbohydrate metabolism.
   d. chronic alcoholism with hepatic failure.
9. The most widely used cutoff blood glucose concentration that indicates hypoglycemia is:
   a. 100 mg/dL.
   b. 75 mg/dL.
   c. 50 mg/dL.
   d. 25 mg/dL.
10. Deficiency of a specific enzyme involved in glycogen metabolism will produce some type of:
   a. glycogen storage disease.
   b. lactic acidosis.
   c. insulin deficiency.
   d. glycolysis.

SUGGESTED READINGS

Adeva-Andany, M., López-Ojén, M., Funcasta-Calderón, R., Ameneiros-Rodríguez, E., Donapetry-García, C., Vila-Altesor, M., et al. (2014). Comprehensive review on lactate metabolism in human health. *Mitochondrion*, *17*, 76–100.

American Diabetes Association. (2012). Standards of medical care in diabetes—2012. *Diabetes Care*, *35*(Suppl.1), S11–S63.

Chan, A. Y., Swaminathan, R., & Cockram, C. S. (1989). Effectiveness of sodium fluoride as a preservative of glucose in blood. *Clinical Chemistry*, *35*, 315–317.

Chen, Y. T. (2001). Glycogen storage diseases. In C. R. Scriver, A. L. Beaudet, W. S. Sly, & D. Valle. (Eds.), *The metabolic and molecular bases of inherited disease* (Vol. 8) (pp. 1521–1551). New York: McGraw-Hill.

Cryer, P. E., Axelrod, L., Grossman, A. B., Heller, S. R., Montori, V. M., Seaquist, E. R., et al. (2009). Evaluation and management of adult hypoglycemic disorders: an endocrine society clinical practice guideline. *Journal of Clinical Endocrinology & Metabolism*, *94*, 709–728.

DCCT. (1993). The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *New England Journal of Medicine*, *329*, 977–986.

Gambino, R., Piscitelli, J., Ackattupathil, T. A., Theriault, J. L., Andrin, R. D., Sanfilippo, M. L., et al. (2009). Acidification of blood is superior to sodium fluoride alone as an inhibitor of glycolysis. *Clinical Chemistry*, *55*, 1019–1021.

Gerich, J. E. (2000). Physiology of glucose homeostasis. *Diabetes Obesity and Metabolism*, *2*, 345–350.

Koeberl, D. D., Kishnani, P. S., Bali, D., & Chen, Y. T. (2009). Emerging therapies for glycogen storage disease type I. *Trends Endocrinology Metabolism*, *20*, 252–258.

Martin-Rendon, E., & Blake, D. J. (2003). Protein glycosylation in disease: new insights into the congenital muscular dystrophies. *Trends Pharmacological Sciences*, *24*, 178–183.

Kraut, J. A., & Madias, N. E. (2015). Lactic acidosis. *New England Journal of Medicine*, *372*, 1078–1079.

Pithukpakorn, M. (2005). Disorders of pyruvate metabolism and the tricarboxylic acid cycle. *Molecular Genetics Metabolism*, *85*, 243–246.

Robinson, B. H. (1995). Lactic acidemia (disorders of pyruvate carboxylase, pyruvate dehydrogenase). In C. R. Shriver, A. L. Beaudet, W. S. Sly, et al. (Eds.), *The metabolic and molecular bases of inherited disease* (pp. 1479–1499). New York: McGraw-Hill.

Sacks, D. B. (2018). Carbohydrates. In C. Burtis, E. Ashwood, & D. Bruns (Eds.), *Tietz textbook of clinical chemistry and molecular diagnostics* (6th ed.). (pp. 518–538). St Louis: Saunders.

Sacks, D. B., Arnold, M., Bakris, G. L., et al. (2011). Guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes mellitus. *Clinical Chemistry*, *57*, e1–e47.

Seaquist, E. R., Anderson, J., Childs, B., Cryer, P., Dagogo-Jack, S., Fish, L., et al. (2013). Hypoglycemia and diabetes: a report of a workgroup of the American Diabetes Association and the Endocrine Society. *Diabetes Care*, *36*, 1384–1395.

Shapiro, E. T., Bell, G. I., Polonsky, K. S., Rubenstein, A. H., Kew, M. C., & Tager, H. S. (1990). Tumor hypoglycemia: relationship to high molecular weight insulin-like growth factor-II. *Journal of Clinical Investigation*, *85*, 1672–1679.

U.K. Prospective Diabetes Study (UKPDS) Group. (1998). Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). *Lancet*, *352*, 837–853.

23

Lipids, Lipoproteins, Apolipoproteins, and Other Cardiac Risk Factors

Sridevi Devaraj, Alan T. Remaley

OBJECTIVES

1. Define the following terms:
   - Apolipoprotein
   - Atherosclerosis
   - Cholesterol
   - Cholesterol ester
   - Emulsification
   - Fatty acid
   - Glycerol
   - Ketone bodies
   - Lipid
   - Lipoprotein
   - Lipoprotein lipase
   - Phospholipid
   - Prostaglandin
   - Sphingolipid
   - Triglyceride
2. List and describe the six classes of lipids, based on chemical structure and function.
3. Discuss the metabolism of cholesterol, fatty acids, and triglycerides, including biochemical structure, synthesis, esterification, absorption, and catabolism.
4. Compare and contrast the six lipoprotein classes on the basis of their physical characteristics, functions, and clinical significance; list the major apolipoproteins that are carried by these lipoproteins and their functions.
5. Describe the exogenous, endogenous, intracellular cholesterol transport, and reverse cholesterol transport pathways of lipoprotein metabolism, including the following:
   - Function
   - Lipoproteins and apolipoproteins involved
   - Enzymes involved
   - Cells/cellular components and organs involved
6. State the effects that the following disorders have on lipid metabolism, lipoproteins, apolipoproteins, and laboratory lipid values, and state known genetic defect:
   - Lipoprotein lipase deficiency
   - Type V hyperlipoproteinemia
   - Dysbetalipoproteinemia
   - Familial hypercholesterolemia
   - Familial combined hyperlipidemia
   - Familial hypertriglyceridemia
   - Familial defective apolipoprotein B-100
   - Hypoalphalipoproteinemia disorders
7. List five major risk factors for coronary heart disease; describe how lipoprotein disorders are managed in adults and children, including therapeutic lifestyle changes and use of cholesterol and triglyceride-lowering agents.
8. List five specific causes of secondary hyperlipidemia/dyslipoproteinemia, including examples of exogenous, endocrine, storage, and renal disorders.
9. Diagram the typical enzymatic reactions used in the laboratory to measure cholesterol and triglyceride; state specimen requirements and describe current assays used to assess the following analytes:
   - HDL
   - LDL
   - Apolipoprotein A-I and B-100
   - Lipoprotein(a)
10. State the uses for and the calculation of the Friedewald formula; calculate low-density lipoprotein cholesterol using this formula when given appropriate information.
11. Describe high-sensitivity C-reactive protein (CRP) and its measurement, and state its usefulness in predicting coronary heart disease events.
12. Analyze and solve case studies related to disorders of lipids and lipoproteins.

KEYWORDS AND DEFINITIONS

Acylglycerol (glycerol ester) A three-carbon alcohol that contains a hydroxyl group on each of its carbons and is classified by the number of fatty acyl groups present; triglycerides are the predominant form of glycerol ester in plasma.

Apolipoproteins The major protein components of lipoproteins.

Atherosclerosis A pathogenic process that is the underlying cause of the common cardiovascular disorders of (1) myocardial infarction, (2) cerebrovascular disease, and (3) peripheral vascular disease.

Cholesterol A steroid alcohol with 27 carbon atoms arranged in a tetracyclical sterane ring system, with a carbon–hydrogen (C–H) side chain and a polar hydroxyl group on its A-ring, making it an amphipathic molecule.

Coronary heart disease A narrowing of the small blood vessels caused by atherosclerotic plaque that impairs the supply of blood and oxygen to the heart.

Fatty acid Any straight-chain monocarboxylic acid with an alkyl chain generally classed as saturated fatty acids with no double bonds; monounsaturated fatty acids with one double bond; and polyunsaturated fatty acids—those with multiple double bonds.

Lipids A class of compounds that are soluble in organic solvents but are nearly insoluble in water and contain nonpolar carbon-hydrogen bonds.

Lipoprotein(a) A lipoprotein structurally similar to low-density lipoprotein but containing a plasminogen-like protein called apo(a). It carries only a relatively small fraction of total cholesterol but is considered to be particularly proatherogenic.

Lipoproteins Spherical micelle-like particles involved in the transport of lipids with nonpolar neutral lipids (triglycerides and cholesterol esters) in their core and more polar amphipathic lipids (phospholipids and free cholesterol) at their surface.

Phospholipid A polar amphipathic lipid located on the surface of a lipoprotein; phospholipids are also found at the aqueous interface of biologic membranes.

Prostaglandin Potent signaling molecules derived from unsaturated 20-carbon fatty acids (primarily arachidonic acid) via the cyclooxygenase pathway; these compounds are involved in a wide variety of physiological processes.

Triglyceride A glycerol ester consisting of three molecules of fatty acid esterified to glycerol and constituting 95% of tissue storage fat.

Lipids have important roles in virtually all aspects of life, including (1) serving as hormones, (2) serving as an energy source, (3) assisting digestion, and (4) acting as structural components in cell membranes. In addition, lipids and **lipoproteins**, the particles that transport lipids in the blood, are intimately involved in the development of **atherosclerosis**—a common disorder that occurs when fat, cholesterol, and other substances build up in the walls of arteries and form atherosclerotic plaques in the vessel wall. Atherosclerosis is the underlying cause of the common cardiovascular disorders of (1) myocardial infarction, (2) cerebrovascular disease, and (3) peripheral vascular disease. In this chapter, the basic biochemistry, clinical significance, and laboratory analysis of each of the major lipid and lipoprotein classes and other cardiovascular risk factors are discussed.

BASIC LIPIDS

The term *lipid* applies to a class of compounds that are soluble in organic solvents but nearly insoluble in water. Lipids primarily contain nonpolar carbon–hydrogen (C–H) bonds and typically yield **fatty acids** and/or complex alcohols after hydrolysis. Some lipids also contain charged or polar groups, which makes these lipids amphipathic with affinity for both water and organic solvents. Amphipathic lipids, such as **phospholipids**, are found at the aqueous interface of biologic membranes because they can form lipid bilayers. Most phospholipids contain a diglyceride, a phosphate group, and a simple organic molecule, such as choline, as the head group. One exception to this rule is sphingomyelin, which is derived from sphingosine instead of glycerol. Overall, lipids are broadly subdivided into six groups on the basis of their chemical structure, namely, (1) cholesterol, (2) fatty acids, (3) acylglycerols, (4) sphingolipids, (5) prostaglandins, and (6) terpenes.

Cholesterol

Cholesterol is found almost exclusively in animals and is a key membrane component of all cells. It is a steroid alcohol with 27 carbon atoms arranged in a tetracyclical sterane ring system, with a C–H side chain (Fig. 23.1). Although it is relatively hydrophobic, cholesterol does contain a polar hydroxyl (OH) group on its A-ring (see Fig. 23.1), thus making it amphipathic (i.e., possessing both hydrophilic and lipophilic properties) and accounting for surface orientation in cell membranes.

Cholesterol Absorption

The average Western diet contains approximately 300 to 450 mg of cholesterol per day, which is derived mostly from animal and dairy products, but only 30% to 60% of it is absorbed. Any dietary esterified cholesterol that contains a fatty acid attached to the hydroxyl group on the A-ring is rapidly hydrolyzed in the intestine to free cholesterol and fatty acids by cholesterol esterases secreted from the pancreas and the small intestine.

Before it is absorbed, cholesterol is first solubilized through a process called emulsification, which involves the formation of mixed micelles that contain (1) unesterified cholesterol, (2) fatty acids, (3) monoglycerides, (4) phospholipids, and (5) conjugated bile acids. Bile acids, by acting as detergents, are the most critical factor in micelle formation. Increased amounts of fat favor the absorption of cholesterol by the formation of more micelles. Most cholesterol absorption occurs in the middle jejunum and the terminal ileum of the small intestine and is mediated by the enterocyte protein NPC1L1 (Niemann-Pick C1-like 1), which is the target for the drug ezetimibe that blocks cholesterol absorption. Once cholesterol enters the intestinal mucosal cell, it is packaged with **triglycerides**, phospholipids, and a large protein called **apolipoprotein** (apo) B-48 into large lipoprotein particles called chylomicrons. Chylomicrons are secreted into the lymph

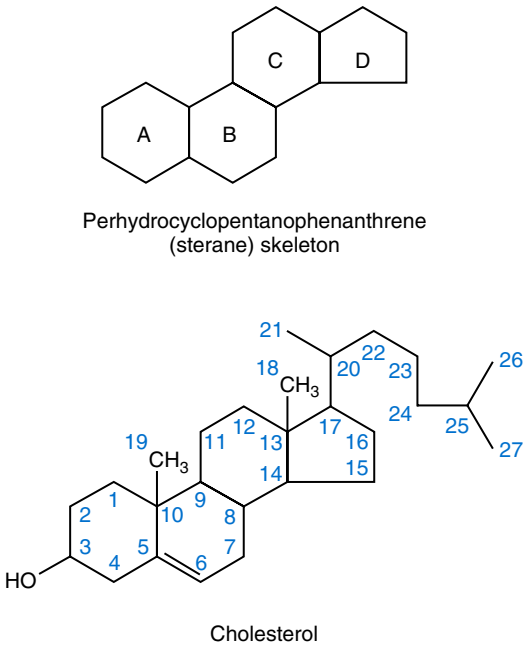


Fig. 23.1 Structure of cholesterol.

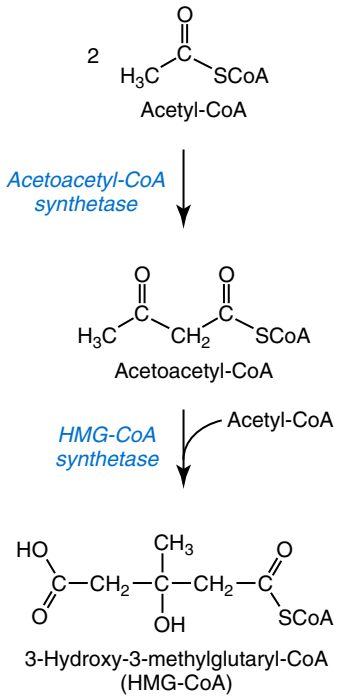


Fig. 23.2 Cholesterol biosynthesis (stage 1).

and eventually enter the circulation, where they deliver the absorbed dietary lipid to the liver and peripheral tissues.

Cholesterol Synthesis

Cholesterol is synthesized by all cells in the body, but particularly by the liver and the intestine. The biosynthesis of cholesterol occurs in three stages. In the first stage (Fig. 23.2), acetyl coenzyme A (CoA), a key metabolic intermediate, forms the six-carbon thioester 3-hydroxy-3-methyl-glutaryl (HMG)-CoA. In the second stage (Fig. 23.3), HMG-CoA is reduced to mevalonate and then is decarboxylated to a series of five-carbon isoprene units. These isoprene units are condensed to form first a 10-carbon (geranyl pyrophosphate) and then a 15-carbon intermediate (farnesyl pyrophosphate). Two of these C_{15} molecules then combine to produce the final product of the second stage, squalene, a 30-carbon acyclic hydrocarbon. The second stage is important because it contains the step involving the microsomal enzyme HMG-CoA reductase, which is the rate-limiting enzyme in cholesterol biosynthesis; it is inhibited by statins, the most effective class of current cholesterol-lowering drugs. The third stage (Fig. 23.4) occurs in the endoplasmic reticulum, with many of the intermediate products bound to a specific carrier protein. In a series of oxidation-decarboxylation reactions, a number of side chains are removed from the tetracyclical sterane ring structure to form the 27-carbon molecule of cholesterol.

Cholesterol Esterification

Cholesterol is esterified to a fatty acid to form a cholesteryl ester by two different enzymes. In cells, excess cholesterol is esterified by acylcholesterol acyltransferase (ACAT) (Fig. 23.5), which helps reduce the cytotoxicity of excess free cholesterol. Once esterified, cholesteryl esters are stored in intracellular lipid drops. Cholesteryl esters also are formed in the circulation by the action of a plasma enzyme called lecithin cholesterol acyltransferase (LCAT), which is bound to lipoproteins, particularly high-density lipoproteins (HDLs). The reaction involves the transfer of a fatty acid from the second carbon position of phosphatidylcholine (lecithin) to cholesterol (see Fig. 23.5). Cholesteryl esters account for approximately 70% of the total cholesterol in plasma. Once cholesterol is esterified, it loses its free hydroxyl group and becomes much more hydrophobic, moving from the surface of lipoprotein particles to the hydrophobic core.

Cholesterol Catabolism

Except for specialized endocrine cells that use cholesterol for the synthesis of steroid hormones, most peripheral cells have limited ability to further catabolize cholesterol. Cholesteryl esters are hydrolyzed to free cholesterol by various lipases in all cells, but thereafter cholesterol has to be returned to the liver to undergo any further catabolism. Approximately one-third of the daily production of cholesterol, or approximately 400 mg/day, is converted in the liver into bile acids (Fig. 23.6). Approximately 90% of these bile acids are reabsorbed in the lower third of the ileum and are returned to the liver by the enterohepatic circulation. Bile acids that enter the large intestine are partially deconjugated by bacterial enzymes to secondary bile acids. Cholic acid is converted, for example, to deoxycholic acid, and chenodeoxycholic acid is converted to lithocholic acid.

Not all cholesterol delivered to the liver is converted to bile salts. Much of it is resecreted into the circulation on lipoproteins, and the remainder is directly excreted into the bile unchanged, where it is solubilized into mixed micelles by bile acids and phospholipids. When the amount of cholesterol in bile exceeds the capacity of these solubilizing agents, it is possible for cholesterol to precipitate and form gallstones.

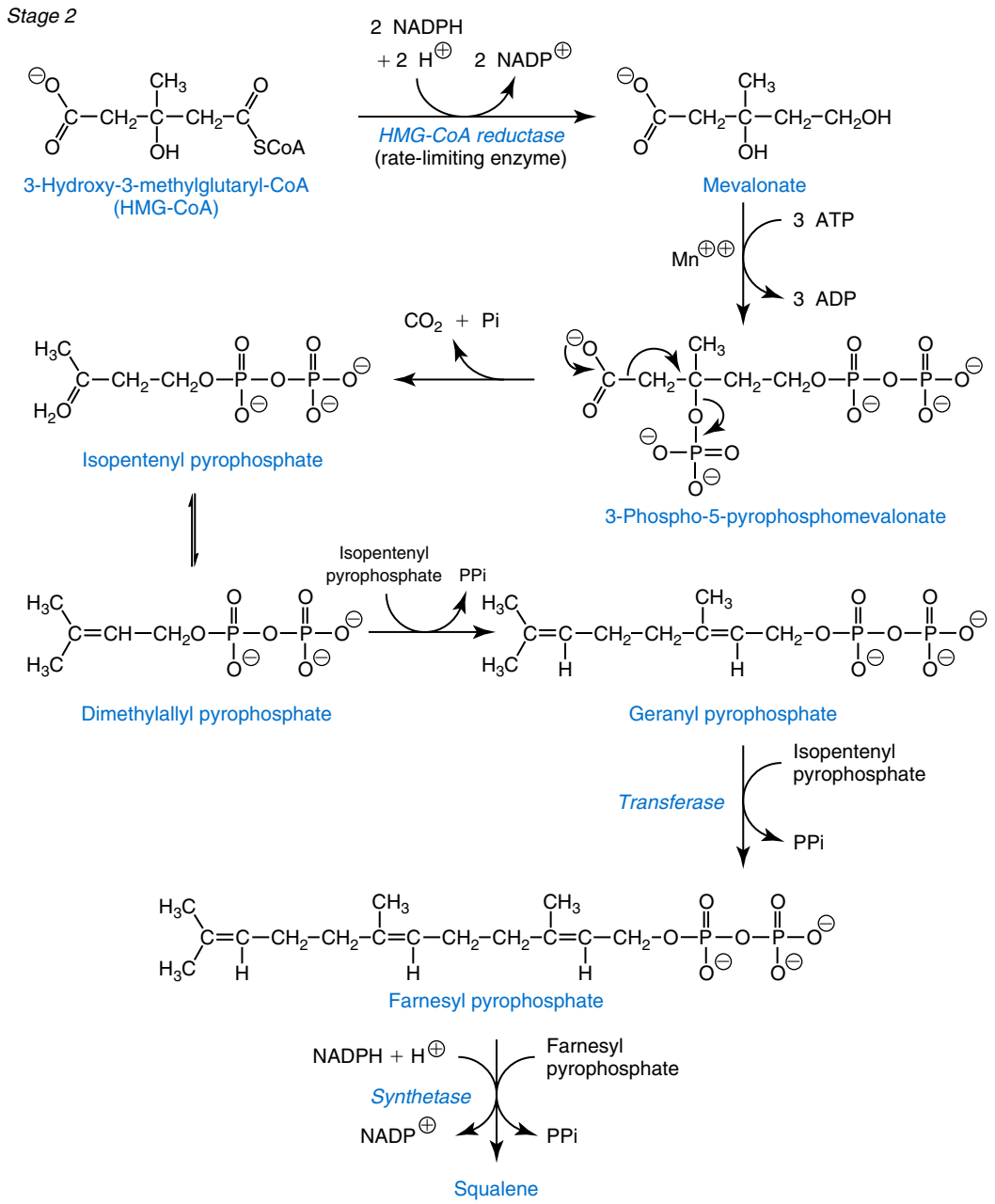


Fig. 23.3 Cholesterol biosynthesis (stage 2).

Fatty Acids

RCOOH is the general chemical formula for a fatty acid, where "R" is an alkyl chain. Fatty acid chain lengths vary and are commonly classified as short-chain (2 to 4 carbon atoms), medium-chain (6 to 10 carbon atoms), or long-chain (12 to 26 carbon atoms) fatty acids. Those of importance in human nutrition and metabolism are included in the long-chain class and contain an even number of carbon atoms.

Fatty acids are further classified according to their degree of saturation. Saturated fatty acids have no double bonds (C=C) between their carbon atoms; monounsaturated fatty acids contain one double bond; and polyunsaturated fatty acids contain multiple double bonds (Fig. 23.7). The double bonds in polyunsaturated fatty acids are usually three carbon atoms apart. Fatty acids from fish, such as salmon, possess up to six unsaturated double bonds and can be more than 20 carbon atoms long. Unsaturated fatty acids are prone to oxidation by the nonenzymatic reaction of oxygen with their double bonds. Numbering of the carbon atoms in fatty acids is done from the carboxyl terminal end (Δ-numbering system) or from the methyl terminal end (η- or ω-numbering system; Table 23.1). In addition, carbon atoms may be labeled with Greek symbols, with α being adjacent to the carboxyl group and ω being farthest away. In the Δ-system, fatty acids are abbreviated according to the (1) number of carbon atoms, (2) number of double bonds, and (3) position(s) of

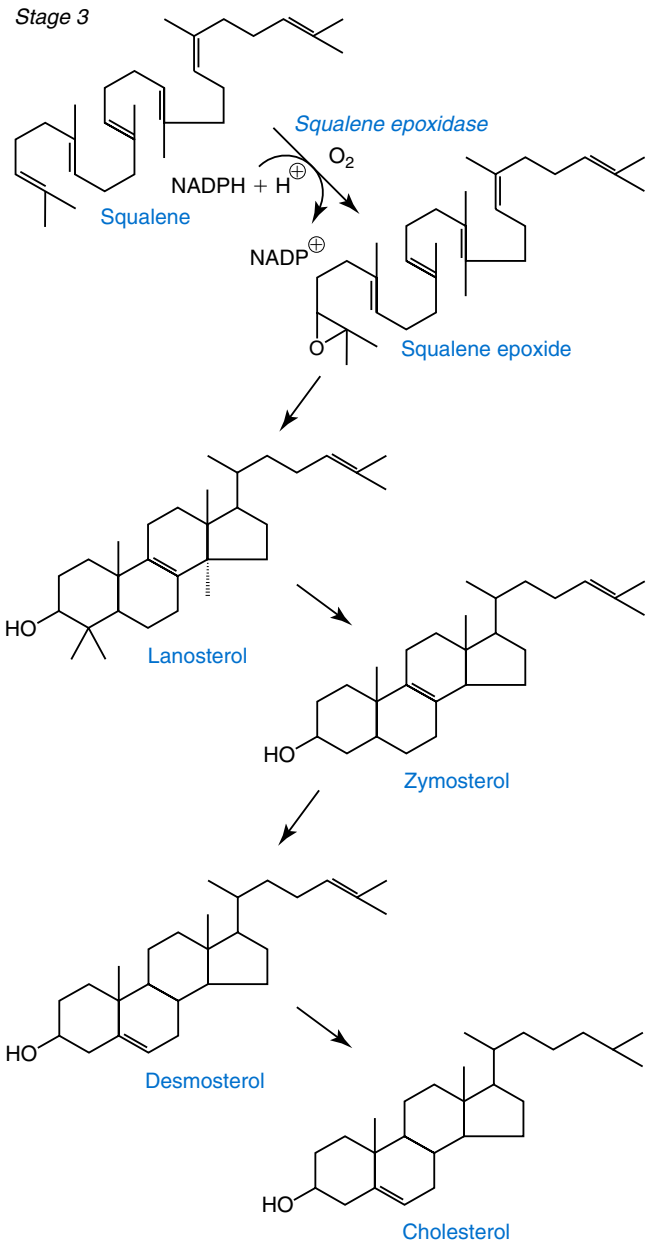


Fig. 23.4 Cholesterol biosynthesis (stage 3).

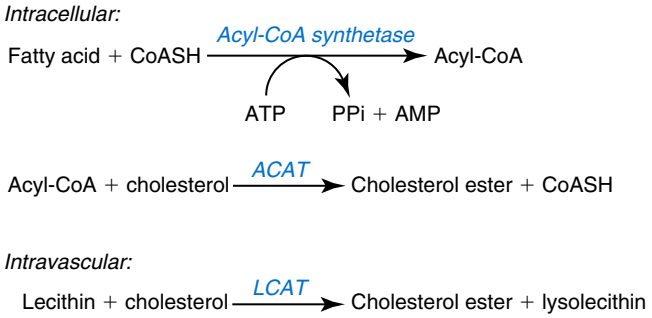


Fig. 23.5 Intracellular and intravascular esterification of cholesterol mediated by acylcholesterol acyltransferase and lecithin cholesterol acyltransferase, respectively. *ACAT,* Acylcholesterol acyltransferase; *ATP,* adenosine triphosphate; *CoA,* coenzyme A; *LCAT,* lecithin cholesterol acyltransferase.

double bond(s). For example, linoleic acid would be written as C_{18}:$2^{9,12}$ and contains 18 carbons and two unsaturated bonds between carbons 9 and 10 and carbons 12 and 13. When the η- or ω-system is used, linoleic acid would be abbreviated as C_{18}:2n–6, where only the first carbon forming the unsaturated pair is written. The *Geneva* or *systematic* classification, which is based on their chemical names, is a third common nomenclature system for fatty acids (see Table 23.1).

In saturated fatty acids, the chain is extended and flexible; the carbon atoms rotate freely around their longitudinal axis. However, unsaturated fatty acids have a fixed 30-degree bend in their chains at each double bond. Depending on the plane in which this bend occurs, the *cis* or the *trans* isomer is produced. In mammals, all naturally occurring unsaturated fatty acids are of the *cis* variety. *Trans* fatty acids in our diet come primarily from catalytic hydrogenation in which unsaturated double bonds are chemically reduced to raise their melting point. This process is used to "harden" or solidify fats in the manufacture of certain foods, such as margarine.

Most fatty acids are also synthesized by the body except for essential fatty acids, such as linoleic acid (C_{18}:$2^{9,12}$), which is made only by plants. Linoleic acid is also converted to arachidonic acid, which is a precursor for prostaglandin synthesis.

Fatty acids exist in the circulation in an unesterified or free state, the latter primarily bound to albumin, or in various esterified forms, such as triglycerides, phospholipids, or cholesteryl esters. The free fatty acid carboxyl group has a pK_a of approximately 4.8; thus free fatty acid molecules exist primarily in their ionized forms. The normal concentration of free fatty acids in human plasma is 0.3 to 1.1 mmol/L (8 to 31 mg/dL) and is very sensitive to physiological energy demands and the availability of alternative forms of metabolic fuel, such as glucose. Most of the free fatty acids in the circulation are derived from intracellular lipolysis of triglycerides by adipose tissue.

Fatty Acid Catabolism

Fatty acids are catabolized in the mitochondria and produce energy by a series of reactions known as β-oxidation. This process is repeated to shorten the fatty acid chain by two carbon atoms at a time from the carboxy terminal end of the molecule. For example, one mole of palmitic acid (C_{16}) is converted to eight moles of acetyl CoA. Acetyl CoA does not normally accumulate in the cell but is condensed enzymatically with oxaloacetate, derived largely from carbohydrate metabolism (Fig. 23.8), to yield citrate, a major component of the tricarboxylic acid cycle (Krebs cycle). The Krebs cycle is a common pathway for the final oxidation of nearly all metabolic fuels, whether derived from carbohydrate, fat, or protein, and ultimately results in the production of adenosine triphosphate (ATP), the main energy storage molecule in the body. Triglycerides contain three fatty acid molecules and are therefore a relatively efficient storage form of metabolic energy. Furthermore, energy storage by triglycerides is efficient in terms of space because it does not require any water for hydration, unlike carbohydrates.

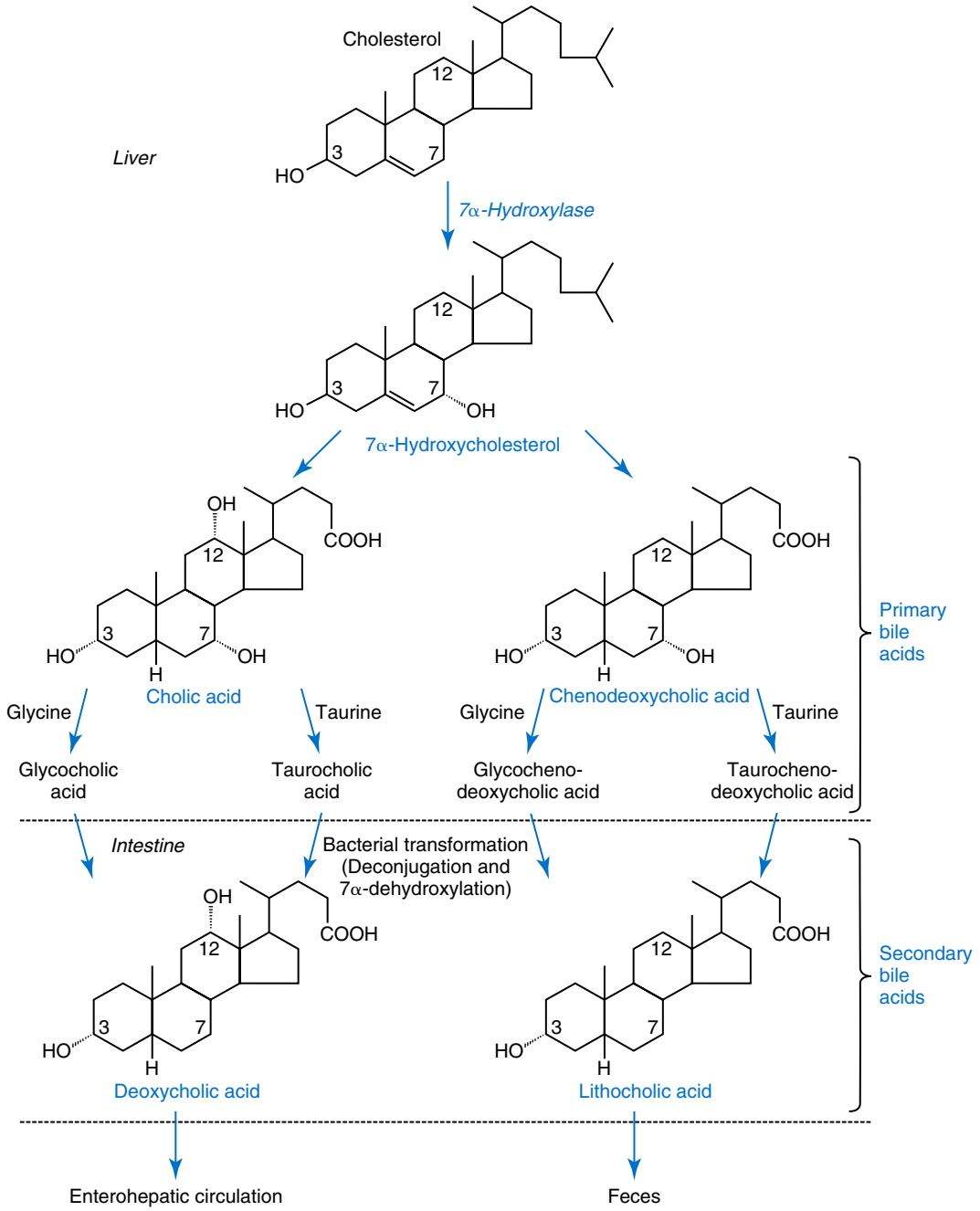


Fig. 23.6 Bile acid synthesis.

Ketone Formation

During prolonged starvation, or when carbohydrate metabolism is impaired, as in uncontrolled diabetes mellitus, the formation of acetyl CoA exceeds the supply of oxaloacetate. The resulting acetyl CoA excess is diverted to an alternative pathway in the mitochondria for the formation of (1) acetoacetic acid, (2) β-hydroxybutyric acid, and (3) acetone—the three compounds known collectively as ketone bodies (Fig. 23.9). Therefore ketosis develops from excessive production of acetyl CoA, as the body attempts to obtain necessary energy from stored fat in the absence of an adequate supply of carbohydrate metabolites (see Chapter 22).

Acylglycerols (Glycerol Esters)

Glycerol is a three-carbon alcohol that contains a hydroxyl group on each of its carbon atoms.

The two terminal carbon atoms in the glycerol molecule are chemically equivalent and are designated α and α′. The center carbon is labeled β. A common alternative labeling system uses the numeral 1 for the α-carbon, 2 for the β-carbon, and 3 for the α′-carbon. The class of **acylglycerol** is determined by the number of fatty acyl groups present: (1) one fatty acid—monoacylglycerols (monoglycerides); (2) two fatty acids—diacylglycerols (diglycerides); and (3) three fatty acids—triacylglycerols (triglycerides). For example,

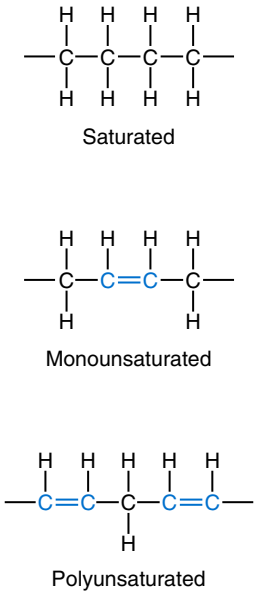


Fig. 23.7 Saturated and unsaturated fatty acids.

TABLE 23.1 Fatty Acids Commonly Found in Human Tissue

Common Name	Systematic Name	Δ-Numbering	-(ω) Numbering
Lauric	Dodecanoic	12:0	12:0
Myristic	Tetradecanoic	14:0	14:0
Palmitic	Hexadecanoic	16:0	16:0
Palmitoleic	9-Hexadecanoic	$16:1^{9}$	16:1n-7
Stearic	Octadecanoic	18:0	18:0
Oleic	9-Octadecanoic	$18:1^{9}$	18:1n-9
Linoleic[a]	9,12-Octadecadienoic	$18:2^{9,12}$	18:2n-6
Linolenic[a]	9,12,15-Octadecatrienoic	$18:3^{9,12,15}$	18:3n-3
Arachidic	Eicosanoic	20:0	20:0
Arachidonic	5,8,11,14-Eicosatetraenoic	$20:4^{5,8,11,14}$	20:4n-6

[a]Essential fatty acids.

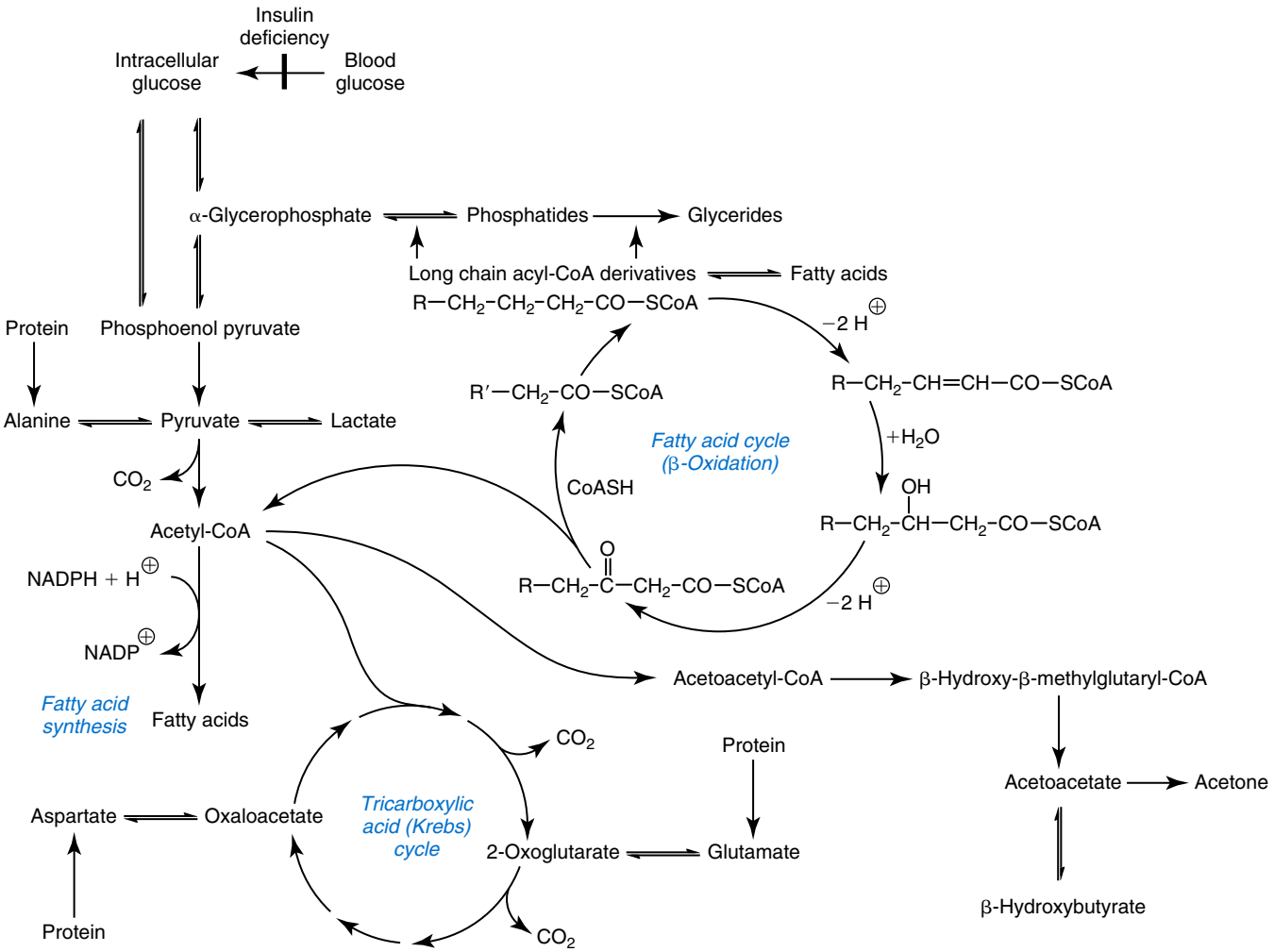


Fig. 23.8 Metabolic relations among intermediates of carbohydrate, fat, and protein metabolism. Note that acetyl-coenzyme A *(CoA)* is produced from both carbohydrate and fat. The glucogenic amino acids, derived from protein metabolism, enter glycolytic paths as α-keto acids. Ketogenic amino acids enter as acetyl-CoA.

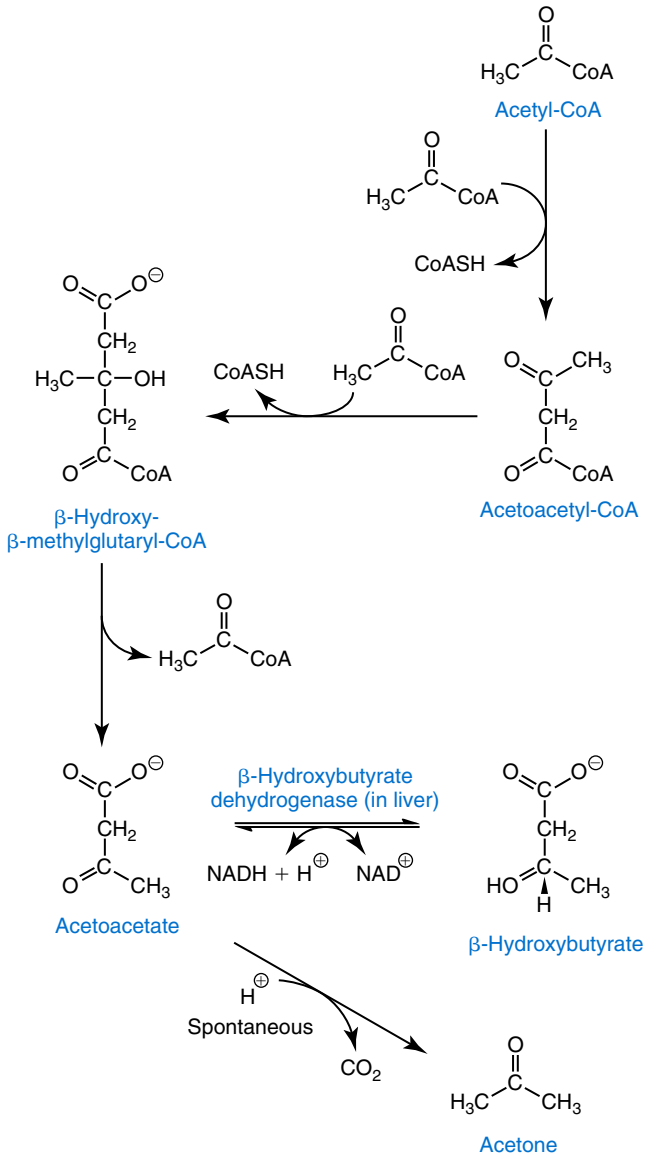


Fig. 23.9 Formation of ketone bodies. *CoA,* Coenzyme A.

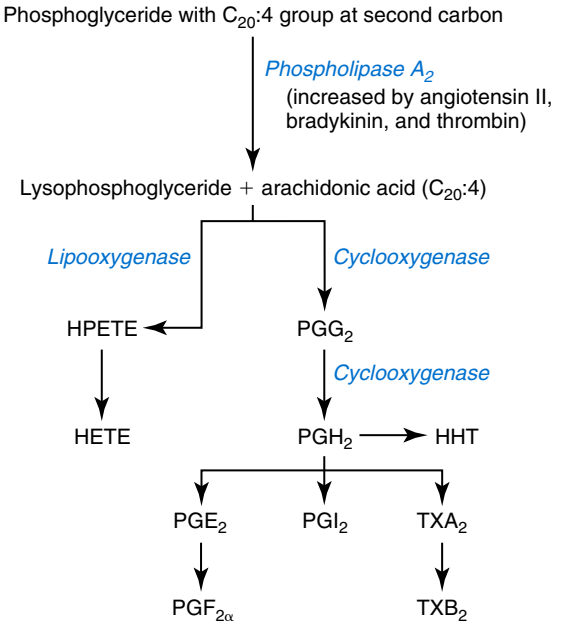


Fig. 23.10 Synthesis of prostaglandins from an arachidonic precursor. *HPETE, HETE, HHT,* 12-L-Hydroxy-5,8,10-heptadecatrienoic acid; *PG,* prostaglandin; *TX,* thromboxane.

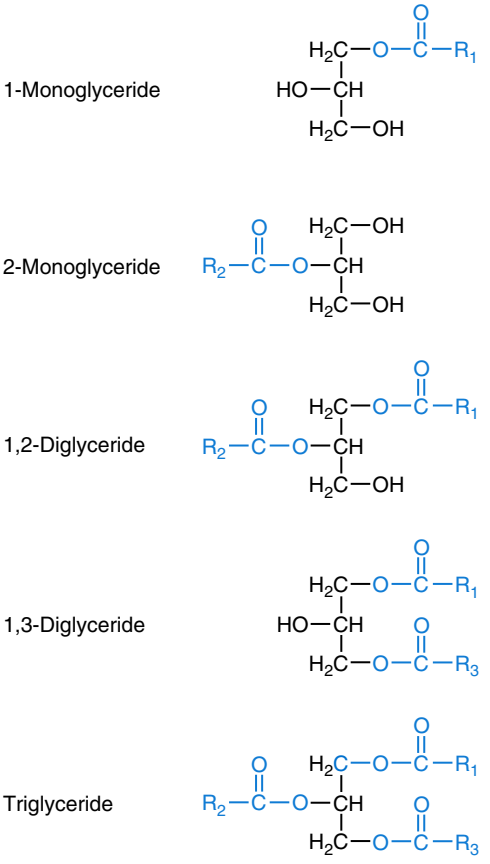


Fig. 23.11 Structure and classification of glycerol esters (acylglycerols). R_1, R_2, and R_3 are fatty acids of varying chain length.

1-monoglyceride indicates that a fatty acid is attached to the α-carbon. This numbering system applies to all acylglycerols, including the phosphoglycerides (Figs. 23.10 and 23.11).

Triglycerides constitute 95% of tissue storage fat and are the predominant form of glyceryl esters found in plasma. The fatty acid residues found in (1) monoglycerides, (2) diglycerides, or (3) triglycerides vary considerably and usually include different combinations of long-chain fatty acids (see Table 23.1). In general, triglycerides from plant sources, such as corn, sunflower, and safflower, tend to be enriched in unsaturated fatty acids and are liquid oils at room temperature. Triglycerides from animals, especially ruminants, tend to have saturated acids and are solids at room temperature.

Dietary triglycerides are digested in the duodenum and are efficiently absorbed in the proximal ileum. Through the action of pancreatic and intestinal lipases and in the presence of bile acids, they are first hydrolyzed to glycerol, monoglycerides, and fatty acids. After absorption, these components of triglycerides are reassembled as triglycerides in the intestinal epithelial cells and then are packaged with cholesterol and apo B-48 to form chylomicrons. Chylomicrons are secreted into the lymphatic system and eventually reach the circulation. Triglycerides are the main metabolic fuel carried by chylomicrons; they are delivered

to the liver and peripheral cells after they have been hydrolyzed to fatty acids by lipases.

Another major class of acylglycerols includes those containing phosphoric acid at the third (α′) carbon atom, which is referred to as phosphoglycerides. In their simplest form, the A group consists of a hydrogen atom and the molecule is called a diacylphosphoglyceride. Usually, the A group contains an alcohol, such as (1) choline, (2) serine, (3) inositol, or (4) ethanolamine. If the A group is choline, for example, the molecule is referred to as phosphatidylcholine (lecithin). Because the types of fatty acid residues R_1 and R_2 are varied, numerous types of phospholipids are formed. These phosphoglycerides are named according to the fatty acid acyl esters attached at C-1 and C-2 of the glycerol. Saturated fatty acids are typically esterified to the C-1 position, whereas polyunsaturated fatty acids are often attached to the C-2 position. In inner mitochondrial membranes, more complex phosphoglycerides, known as cardiolipins, are found. They are derived from two phosphoglyceride molecules joined by a glycerol bridge.

Sphingolipids

Sphingolipids make up a fourth class of lipids derived from the amino alcohol sphingosine (Fig. 23.12). This dihydric 18-carbon alcohol contains an amino group at C-17. A fatty acid containing 18 or more carbon atoms is attached to the amino group through an amide linkage to form *ceramide*. This is an intermediate structure in the formation of (1) sphingomyelin, (2) galactosylceramide, and (3) glucosylceramide (see Fig. 23.12). In addition, the sugar-containing ceramides have a sulfate group that is usually attached to the 2-position of the galactose residue to form the sulfatides. The glycosyl ceramides have additional monosaccharide moieties, such as (1) galactose, (2) *N*-acetylgalactosamine, and (3) *N*-acetylneuraminic acid, to form complex globosides and gangliosides. Gangliosides are especially abundant in the membranes of the gray matter of the brain, whereas glycosphingolipids have a more general role in cell membranes and can also be a source of blood group and tumor antigens.

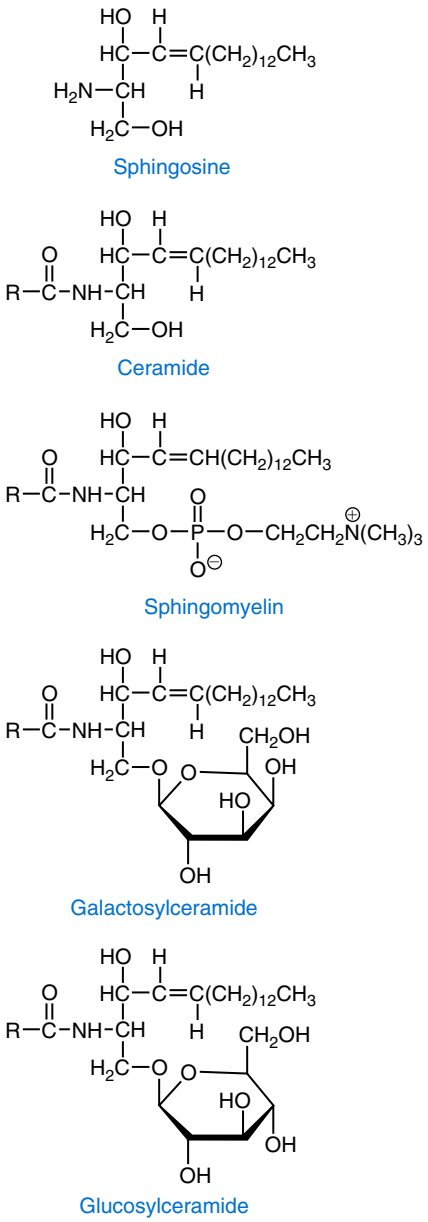


Fig. 23.12 Structures of sphingolipids.

Prostaglandins

Prostaglandins and related compounds, such as thromboxanes and leukotrienes, are derivatives of fatty acids, primarily arachidonate. These potent bioactive lipids exert diverse physiological actions (Table 23.2) at concentrations as low as 1 μg/L. The prostaglandins are a series of C20 unsaturated fatty acids that contain a cyclopentane ring; the parent fatty acid has been given the trivial name prostanoic acid. By convention, prostaglandins are abbreviated PG, with the class designated by a capital letter (A, B, E, F, G, H, and I), followed by a number and in some cases a Greek letter (see Figs. 23.10 and 23.13). The number after the capital letter is usually written as a subscript and is used to designate the number of unsaturated bonds in the PG side chains and not within the ring structure itself. Sixteen naturally occurring prostaglandins have been described (Table 23.3), but only seven, along with two thromboxanes, are commonly found in the body. These are termed primary prostaglandins.

TABLE 23.2 Prostaglandin-Mediated Effects

Site of Action	Physiological Response
Arterial smooth muscle	Alters blood pressure
Uterine muscle	Induces labor, therapeutic abortion
Lower gastrointestinal tract	Increases motility
Bronchial smooth muscle	Causes bronchospasm
Platelets	Increase coagulability
Capillaries	Increase permeability
Stomach	Enhances gastric acid secretion
Adipose tissue	Inhibits triglyceride lipolysis

Although prostaglandins have hormone-like action, they are different from conventional hormones in that they are synthesized at the site of action and are made in almost all tissues. Linoleic acid (C_{18}:$2^{9,12}$) is the precursor of two of the three

20-carbon fatty acids that form prostaglandins; linolenic acid (C_{18}:$2^{9,12,15}$) is the other precursor. Both of these fatty acids are considered essential because they are not synthesized in the body and therefore must be present in the diet. Once formed, prostaglandins have short-lived effects and are catabolized within seconds. Inactivation of prostaglandin appears to be mediated by two enzymes: 15α-hydroxy-prostaglandin dehydrogenase and Δ^{13}-prostaglandin reductase. Prostaglandins are not stored, but the precursor C_{20} fatty acids are present in tissue attached to the C-2 position of phosphoglycerides.

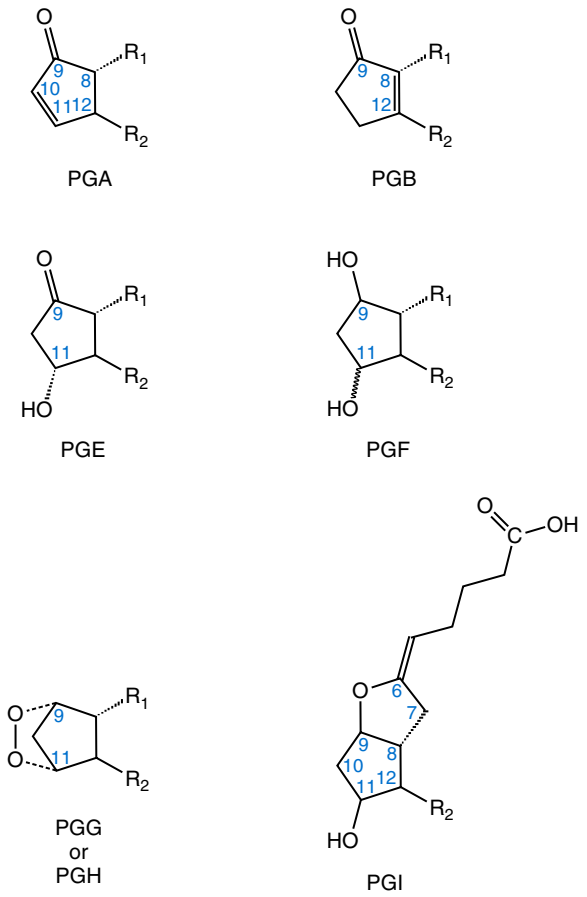


Fig. 23.13 Structures of phosphoglycerides and common alcohol groups associated with them. R_1 and R_2 are fatty acids of varying carbon atom lengths. *PG,* Prostaglandin.

TABLE 23.3 Naturally Occurring Prostaglandins

Primary PG	Other PGs
PGE_1	PGA_1
$PGF_1\alpha$	PGA_2
PGE_2	19α-$OHPGA_1$
$PGF_2\alpha$	19α-$OHPGA_2$
PGG_2	PGB_1
PGH_2	PGB_2
PGI_2	19α-$OHPGB_2$
Thromboxane A_2	PGE_3
Thromnane B_2	$PGF_3\alpha$

PG, Prostaglandin.

When prostaglandin synthesis is stimulated, the C_{20} precursor is hydrolyzed from phospholipids by phospholipase A_2. Release of the C_{20} fatty acid appears to be the rate-limiting step in prostaglandin synthesis and is stimulated by various mediators, such as bradykinin, thrombin, or angiotensin II.

Once released, arachidonic acid follows one of two pathways. The lipoxygenase route produces 12-L-hydroperoxy-5,8,10,14 eicosatetraenoic acid (HPETE); HPETE spontaneously decomposes to 12-L-hydroxy-5,8,10,14 eicosatetraenoic acid (HETE) (see Fig. 23.10). The alternative pathway is mediated by cyclooxygenase (COX) to produce the endoperoxides PGG_2 and PGH_2. Nonsteroidal antiinflammatory drugs (NSAIDs; such as aspirin, ibuprofen, naproxen, and indomethacin) inhibit the COX enzymes, thereby decreasing prostaglandin synthesis.

PGI_2, or prostacyclin, is derived from arachidonic acid (see Fig. 23.10) in the vascular endothelium. It has a powerful vasodilatory action, especially on the coronary arteries, and inhibits platelet aggregation. Thromboxane A_2 is synthesized from arachidonic acid but is also produced by platelets. It has the opposite effect of prostacyclin because it stimulates the contraction of arterial smooth muscle and enhances platelet aggregation. It has a half-life of approximately 30 seconds and is rapidly converted to its inactive metabolite, thromboxane B_2. Thromboxanes are slightly different in structure from the other prostaglandins in that they contain six-sided rings of five carbon atoms and one oxygen atom (Fig. 23.14).

Terpenes

Terpenes are polymers of the five-carbon isoprene unit and include vitamins A, E, and K and the dolichols, which play important roles in protein glycosylation.

LIPOPROTEINS

Lipids synthesized in the liver and in the intestine are transported in the plasma in macromolecular complexes known as lipoproteins. Lipoproteins are typically spherical micelle-like

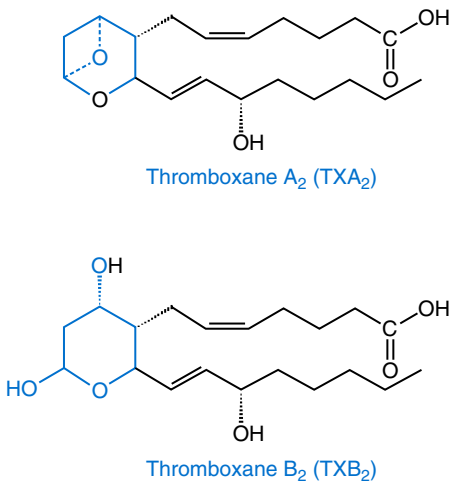


Fig. 23.14 Structures of thromboxanes.

particles with nonpolar neutral lipids (triglycerides and cholesterol esters) in their core and more polar amphipathic lipids (phospholipids and free cholesterol) at their surface (Fig. 23.15). They also contain one or more specific proteins, called apolipoproteins, on their surface.

Classification

Lipoproteins have different physical and chemical properties (Table 23.4) because they contain different proportions of lipids and proteins (Table 23.5). Traditionally, lipoproteins have been categorized on the basis of differences in their hydrated densities, as determined by ultracentrifugation. These categories include (1) chylomicrons, (2) very-low-density lipoprotein (VLDL), (3) intermediate-density lipoprotein (IDL), (4) low-density lipoprotein (LDL), (5) HDL, and (6) **lipoprotein(a)** [Lp(a)]. In general, the larger lipoproteins contain more core lipids in their core, such as triglycerides, and cholesteryl esters and hence are lighter in density. They also contain a smaller relative percentage of protein. In the fasting state, most plasma triglycerides are present in VLDL, but 2 to 6 hours after a meal, most triglycerides are transported on chylomicrons. LDL carries approximately 70% of total plasma cholesterol but very little triglyceride (see Table 23.5). HDL typically contains approximately 20% to 30% of plasma cholesterol.

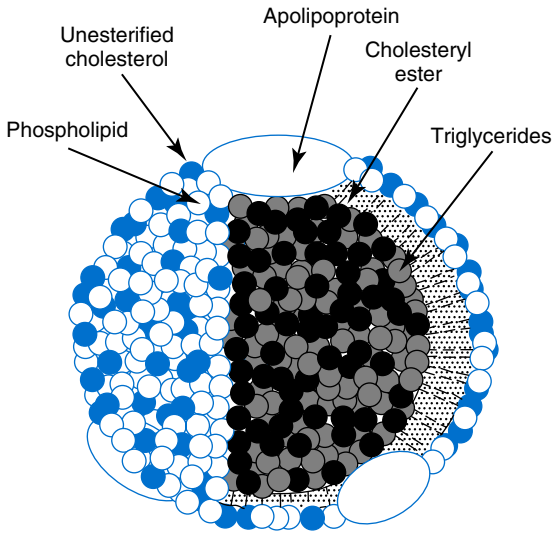


Fig. 23.15 Structure of a typical lipoprotein particle.

Lp(a) is a distinct class of lipoprotein (Table 23.6) that is structurally related to LDL because both lipoproteins possess one molecule of apo B-100 per particle, with similar lipid compositions. Unlike LDL, Lp(a) also contains a carbohydrate-rich protein called apo(a), which is covalently bound to the apo B-100 through a disulfide linkage. Apo(a) exhibits a significant sequence homology with plasminogen and a high degree of variation in polypeptide chain length (Fig. 23.16). Apo(a) contains a tandem array of a protein motif called a *kringle domain* with different-sized apo(a) molecules (polymorphisms) reflecting the inclusion of variable numbers of kringle 4 type 2 domains. The plasma level of Lp(a) is mostly genetically determined and is relatively invariant and thus does not need to be repeatedly measured. The mechanism is unclear, but Lp(a) is particularly proatherogenic, based on recent genetic studies, and has also been associated with aortic stenosis. There is currently no adequate drug or dietary therapy for lowering Lp(a), but a promising new therapy based on antisense oligonucleotides is being developed.

Lipoproteins can be electrophoretically separated, and this forms the basis for one system of nomenclature. For example, at pH 8.6, HDL migrates with the α-globulins, LDL with the β-globulins, and VLDL and Lp(a) between the α- and β-globulins, in the pre–β-globulin region. IDL forms a broad band between β- and pre–β-globulins. Chylomicrons remain at the application point. Traditionally, lipoprotein classes were referred to by their electrophoretic locations as pre–β-lipoprotein, VLDL; β-lipoprotein, LDL; and α-lipoprotein, HDL. Electrophoresis provided the foundation for an early original phenotypical classification system (types 1 to 5) for familial dyslipidemias (for further discussion on this technique, refer to Chapter 11).

TABLE 23.4 Characteristics of Human Plasma Lipoproteins

Variable	Chylomicron	VLDL	IDL	LDL	HDL	Lp(a)
Density (g/mL)	<0.95	0.95–1.006	1.006–1.019	1.019–1.063	1.063–1.210	1.040–1.130
Electrophoretic mobility	Origin	Pre-beta	Between beta and pre-beta	Beta	Alpha	Pre-beta
Molecular weight (Da)	$0.4–30 \times 10^9$	$5–10 \times 10^6$	$3.9–4.8 \times 10^6$	2.75×10^6	$1.8–3.6 \times 10^5$	$2.9–3.7 \times 10^6$
Diameter (nm)	>70	26–70	22–24	19–23	4–10	26–30
Lipid:lipoprotein ratio	99:1	90:10	85:15	80:20	50:50	75:26–64:36
Major lipids	Exogenous	Endogenous	Endogenous, cholesteryl esters	Cholesteryl esters	Phospholipids	Cholesteryl esters, phospholipids
Major proteins	A-I	B-100	B-100	B-100	A-I	(a)
	B-48	C-I	E	—	A-II	B-100
	C-I	C-II	—	—	—	—
	C-II	C-III	—	—	—	—
	C-III	E	—	—	—	—

HDL, High-density lipoprotein; *IDL,* intermediate-density lipoprotein; *LDL,* low-density lipoprotein; *Lp(a),* lipoprotein(a); *VLDL,* very-low-density lipoprotein.

TABLE 23.5 Chemical Composition (%) of Human Plasma Lipoproteins

		SURFACE COMPONENTS		CORE LIPIDS	
Core Lipid Esters	**Cholesterol**	**Phospholipids**	**Apolipoproteins**	**Triglycerides**	**Cholesteryl**
Chylomicrons	2	7	2	86	3
VLDL	7	18	8	55	12
IDL	9	19	19	23	29
LDL	8	22	22	6	42
HDL_2	5	33	40	5	17
HDL_3	4	25	55	3	13

Surface components and core lipids given as percentage of dry mass.
HDL, High-density lipoprotein; *HDL_2,* HDL syndrome 2; *HDL_3,* HDL syndrome 3; *IDL,* intermediate-density lipoprotein; *LDL,* low-density lipoprotein; *VLDL,* very-low-density lipoprotein.
From Havel, R. J., & Kane, J. P. (1995). Introduction: structure and metabolism of plasma lipoproteins. In: C. R. Scriver, A. L. Beaudet, W. S. Sly, & D. Valle. (Eds.), *The metabolic and molecular bases of inherited diseases* (7th ed., Vol. II) (1841–1850). New York: McGraw-Hill. Reproduced with permission of The McGraw-Hill Companies.

TABLE 23.6 Classification[a] and Properties of Major Human Plasma Apolipoproteins

Apolipoprotein	Molecular Weight (Da)	Chromosomal Location	Function	Lipoprotein Carrier(s)
Apo A-I	29,016	11	Cofactor LCAT	Chylomicron, HDL
Apo A-II	17,414	1	Not known	HDL
Apo A-IV	44,465	11	Activation of LCAT	Chylomicron, HDL
Apo B-100	512,723	2	Secretion of triglycerides from liver-binding protein to LDL receptor	VLDL, IDL, LDL
Apo B-48	240,800	2	Secretion of triglycerides from intestine	Chylomicron
Apo C-I	6630	19	Activation of LCAT	Chylomicron, VLDL, HDL
Apo C-II	8900	19	Cofactor LPL	Chylomicron, VLDL, HDL
Apo C-III	8800	11	Inhibition of apo C-II activation of LPL	Chylomicron, VLDL, HDL
Apo E	34,145	19	Facilitation of uptake of chylomicron remnant and IDL	Chylomicron, VLDL, HDL
Apo(a)	187,000–662,000	6	Unknown	Lp(a)

[a]Both Roman and Arabic numerals are used to label the individual classifications. In this chapter, Roman symbols are used.
HDL, High-density lipoprotein; *IDL,* intermediate-density lipoprotein; *LCAT,* lecithin cholesterol acyltransferase; *LDL,* low-density lipoprotein; *Lp(a),* lipoprotein(a); *LPL,* lipoprotein lipase; *VLDL,* very-low-density lipoprotein.

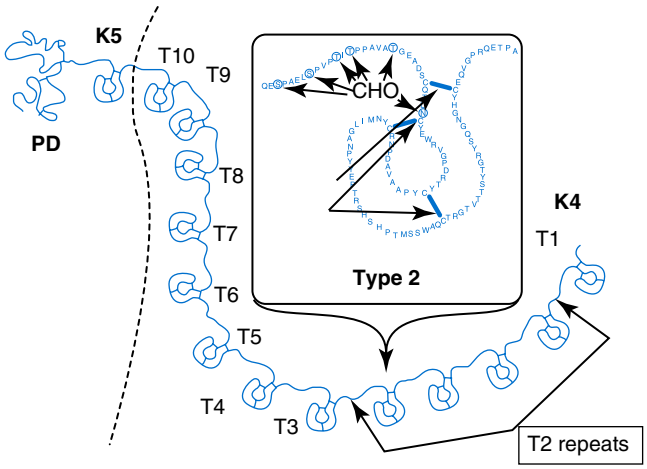


Fig. 23.16 Structure of lipoprotein(a).

APOLIPOPROTEINS

Apolipoproteins are the major protein components of lipoproteins (see Table 23.6). Each class of lipoprotein carries several apolipoproteins in differing proportions. Apo A-I is the major protein in HDL. Apo B-100 is the main protein on LDL and Lp(a), and apo B-48, which is produced from apo B-100 messenger RNA (mRNA) by an RNA editing process, is found on chylomicrons. Both apo B-100 and apo B-48 are found at one molecule per particle, are firmly bound, and do not exchange between particles, as the other apolipoproteins do. Apolipoproteins perform the following major functions: (1) modulate the activity of enzymes that act on lipoproteins, (2) maintain the structural integrity of the lipoprotein complex, and (3) facilitate the uptake of lipoprotein by acting as ligands for specific cell-surface receptors. Most apolipoproteins contain amphipathic helices, which are α-helices with one face containing hydrophobic amino acids and the other face containing polar or charged amino acids. This feature enables apolipoproteins to bind to lipids and still interact with the surrounding aqueous environment.

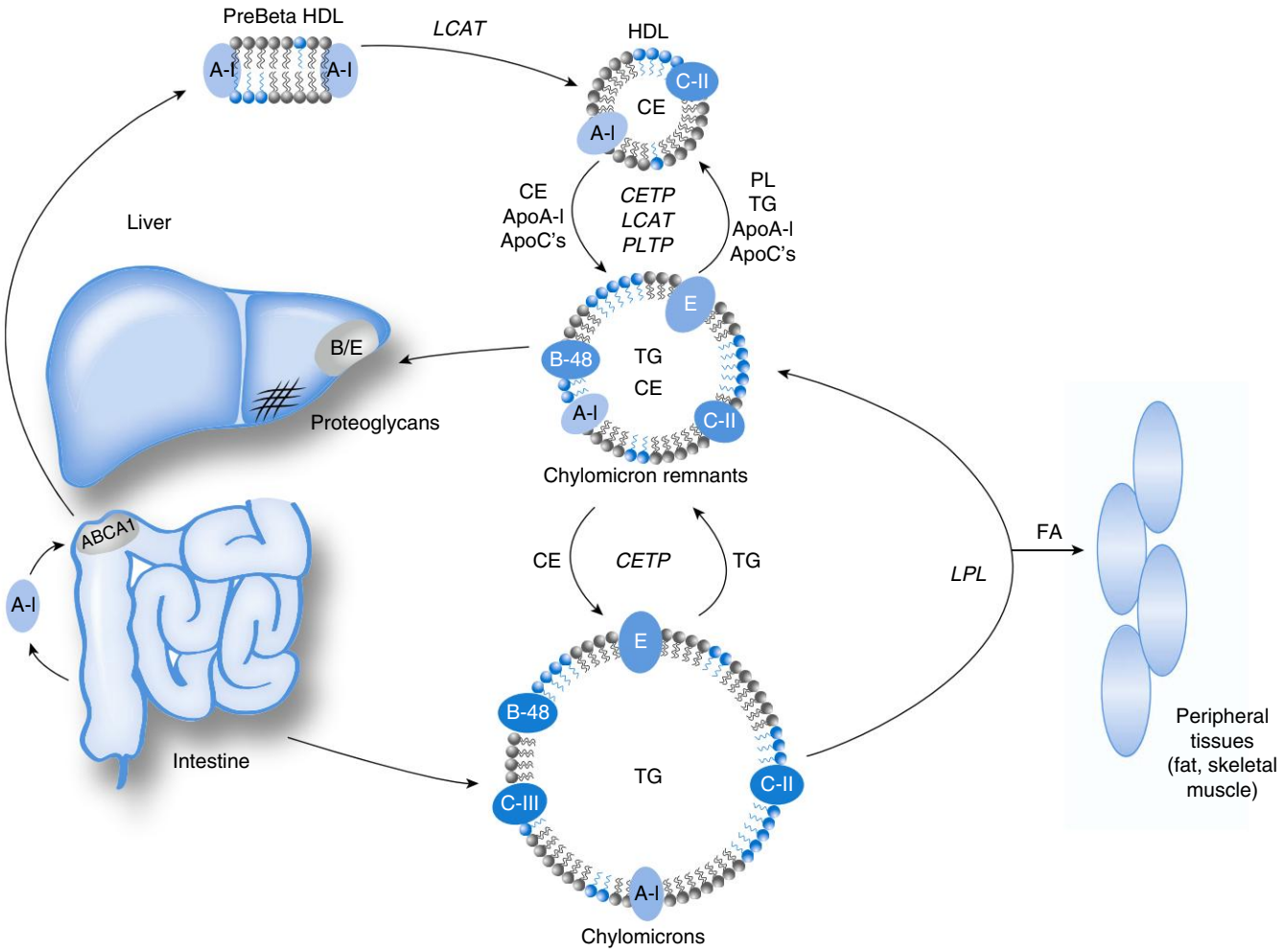


Fig. 23.17 Exogenous (intestinal) lipoprotein metabolism pathway. *A-1,* Apolipoprotein A-I; *ABCA1,* ATP-binding cassette transporter A1; *B-48,* apolipoprotein B-48; *B/E,* ApoB- and ApoE-dependent receptors; *C-II,* apolipoprotein C-II; *CE,* cholesterol ester; *CETP,* cholesterol ester transfer protein; *E,* apolipoprotein E; *FA,* fatty acid; *FC,* free cholesterol; *HDL,* high-density lipoprotein; *LCAT,* lecithin cholesterol acyltransferase; *LPL,* lipoprotein lipase; *PL,* phospholipid; *PLTP,* phospholipid transfer protein; *TG,* triglyceride.

METABOLISM OF LIPOPROTEINS

Lipoprotein metabolism is commonly divided into (1) exogenous, (2) endogenous, (3) intracellular cholesterol transport, and (4) reverse cholesterol transport pathways.

Exogenous Pathway

The roles of the exogenous pathway—transport of dietary lipids from the intestine to the liver and peripheral cells—are largely mediated by chylomicrons (Fig. 23.17). Nascent chylomicrons, which are 90% triglycerides by mass, are first assembled by the microsomal transfer protein (MTP) in the endoplasmic reticulum of enterocytes by combining triglycerides and other lipids with apo B-48. Chylomicrons are secreted into the lymph and, after entering the circulation, acquire from HDL additional apolipoproteins, such as apo E, apo C-III, and apo C-II. Apo E is a ligand for uptake by the liver, whereas C-III inhibits hepatic clearance of lipoproteins and is also an inhibitor of lipoprotein lipase (LPL). In contrast, apo C-II is a potent activator of LPL, which is attached to the luminal surface of endothelial cells and rapidly hydrolyzes the triglycerides on chylomicrons to free fatty acids. Liberated fatty acids are taken up by both a passive and active transport mechanism into muscle cells and other peripheral tissues as an energy source or by adipose cells for energy storage as triglycerides. As a consequence of lipolysis, chylomicrons are transformed into smaller chylomicron remnant particles, which are rapidly removed by the liver.

Endogenous Pathway

The function of the endogenous pathway is to transfer hepatically derived lipids, especially triglycerides, to peripheral cells for energy metabolism. It is mediated by the apo B-100–containing lipoproteins (Fig. 23.18). Hepatically derived lipids represent lipids that were synthesized by the liver or dietary lipids that were transferred to the liver by the exogenous pathway. VLDL, which contains approximately 55% triglycerides by mass and includes one molecule of apo B-100, is the principal apo B–containing lipoprotein that is secreted by the liver. As it

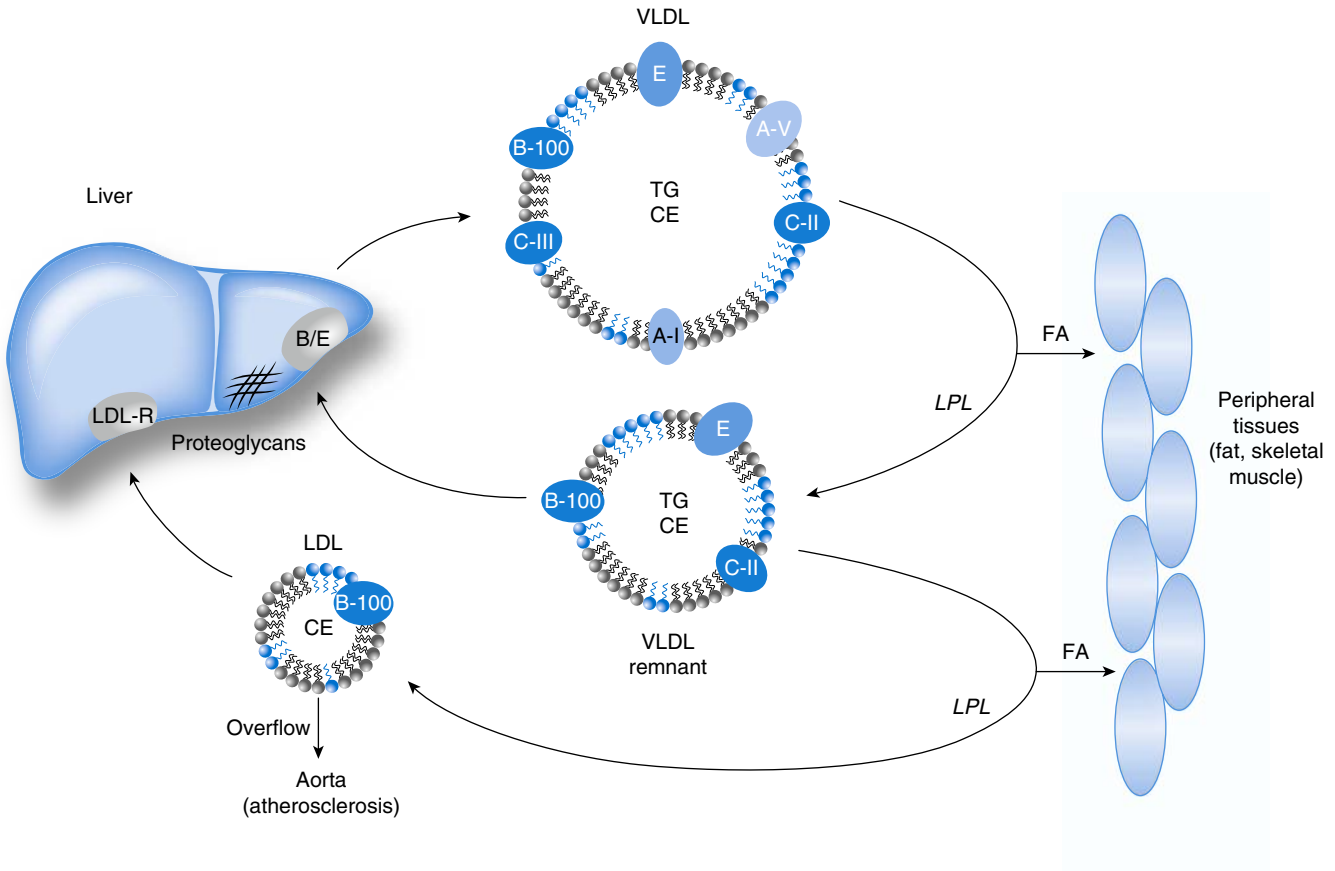


Fig. 23.18 Endogenous (hepatic) lipoprotein metabolism pathway. *A-1,* Apolipoprotein A-I; *A-V,* apolipoprotein A-V; *B-100,* apolipoprotein B-100; *B/E,* ApoB- and ApoE-dependent receptors; *C,* apolipoprotein C-II; *CE,* cholesterol ester; *E,* apolipoprotein E; *FA,* fatty acid; *FC,* free cholesterol; *HDL,* high-density lipoprotein; *IDL,* intermediate-density lipoprotein; *LCAT,* lecithin cholesterol acyltransferase; *LDL,* low-density lipoprotein; *LDL-R,* low-density lipoprotein receptor; *LPL,* lipoprotein lipase; *PL,* phospholipid; *TG,* triglyceride; *VLDL,* very-low-density lipoproteins.

does on chylomicrons, the apo C-II present on the surface of VLDL activates LPL on endothelial cells. This leads to hydrolysis of VLDL triglycerides and release of free fatty acids, which are taken up by cells. The progressive lipolysis of triglycerides from the core of VLDL transforms it to IDL and eventually to LDL. The triglycerides on LDL are further depleted by the cholesterol ester transfer protein (CETP), which removes them from LDL and exchanges them for cholesteryl esters from HDL. Although almost all cells express the LDL receptor, most LDL is eventually returned to the liver via the LDL receptor, which recognizes apo B-100. Cholesterol returned to the liver is (1) reused for the secretion of lipoproteins, (2) used in the production of bile salts, or (3) excreted directly into the bile. PCSK9 is a plasma protein that binds to the extracellular domain of the LDL receptor, promoting its cellular internalization and degradation (Fig. 23.19). Monoclonal antibody therapy against PCSK9 has been developed for inhibiting this process, thus leading to a lowering of plasma levels of LDL.

Intracellular Cholesterol Transport Pathway

The intracellular cholesterol transport pathway maintains cellular homeostasis of cholesterol. Although cholesterol is a necessary and critical component of all cell membranes, excess cholesterol will alter the biophysical properties of membranes and eventually will become toxic to the cell. Besides through biosynthesis, all cells receive cholesterol via uptake of extracellular lipoproteins by cell surface receptors, such as the LDL receptor (see Fig. 23.19). Most lipoprotein receptors deliver intact lipoprotein particles to lysosomes, where they are degraded. Cholesteryl esters are converted to free cholesterol by lysosomal acid lipase. Because most cells do not catabolize cholesterol further, any cholesterol delivered to the cell is (1) used for membrane biogenesis, (2) stored in intracellular lipid drops after reesterification by ACAT, or (3) removed from the cell by the reverse cholesterol transport pathway. In addition, cells have a complex mechanism involving both transcriptional and posttranscriptional regulation, so that any excess intracellular cholesterol will inhibit cholesterol biosynthesis and expression of the LDL receptor. However, when LDL is modified such as via oxidation, it is taken up by the scavenger receptor pathway in the macrophages, leading to unregulated cholesterol uptake, resulting in the earliest fatty streak lesion seen in atherosclerosis.

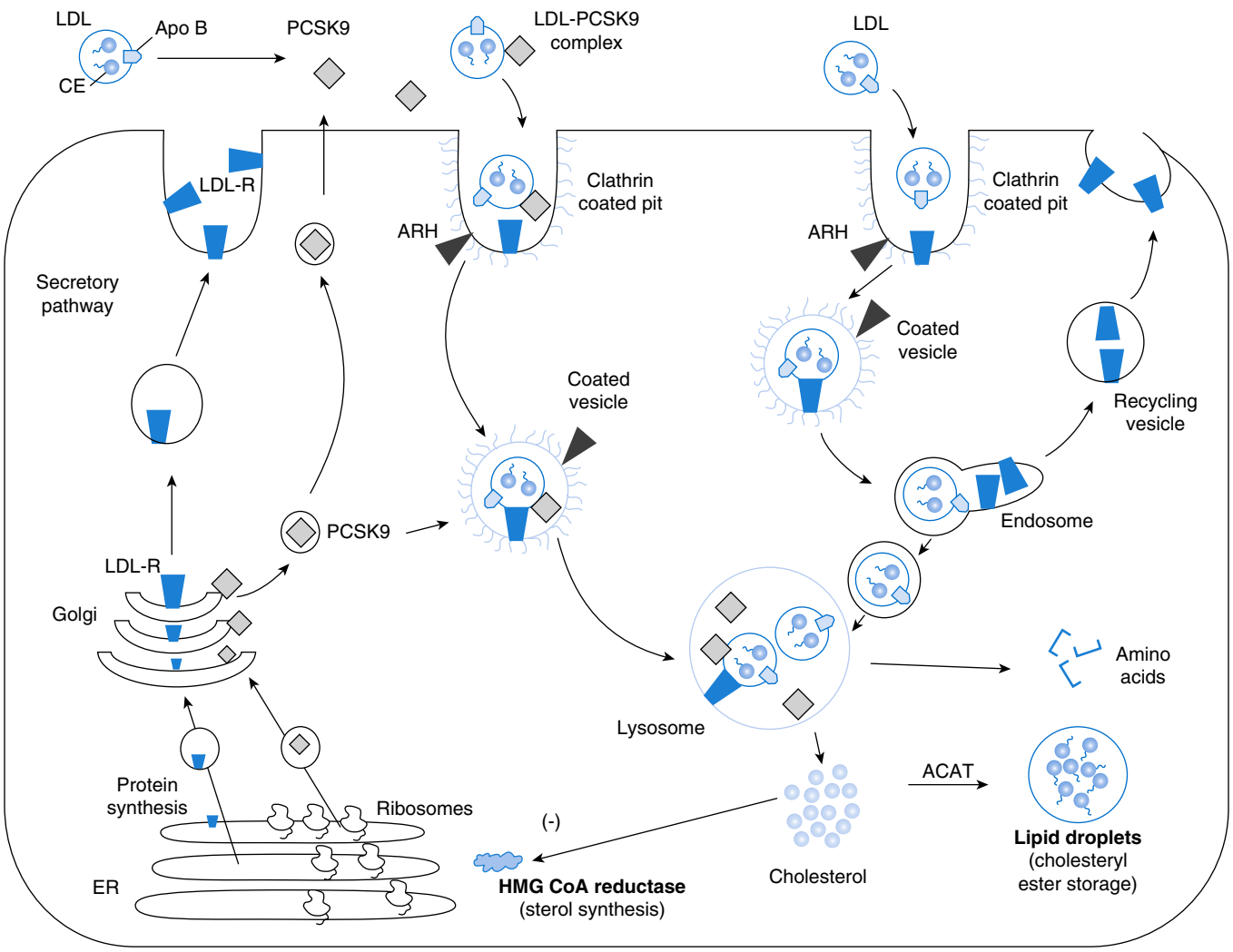


Fig. 23.19 Intracellular cholesterol transport pathway. *ACAT,* Acyl-CoA cholesterol acyltransferase; *Apo B,* apolipoprotein B-100; *ARH,* autosomal recessive hypercholesterolemia adaptor protein; *HMG-CoA reductase,* 3-hydroxy-3-methylglutaryl coenzyme A reductase; *LDL,* low-density lipoprotein; *LDL-R,* low-density lipoprotein receptor; *PCSK9,* proprotein convertase subtilisin/kexin type 9.

Reverse Cholesterol Transport Pathway

The function of the reverse cholesterol transport pathway is to remove excess cellular cholesterol from peripheral cells and return it to the liver for excretion. This process is largely mediated by HDL (Fig. 23.20). Cholesterol is actively pumped out of cells by the ABCA1 transporter onto lipid-poor apo A-I, which binds to lipid microdomains on the plasma membrane created by ABCA1. This process results in the formation of disc-shaped nascent HDL, which is made in the liver and in the intestine. Discoidal HDL then interacts with ABCA1 in peripheral cells, such as macrophages, and removes additional cholesterol. LCAT, which esterifies cholesterol on HDL, plays a key role in reverse cholesterol transport because cholesteryl esters are much more hydrophobic than cholesterol and remain trapped in the core of HDL until they are removed by the liver. The esterification of cholesterol on HDL converts the disc-shaped nascent HDL to spherical HDL. Spherical HDL, the main form of HDL in the circulation, acts as an extracellular acceptor for cholesterol that may be removed from cells by other mechanisms besides ABCA1.

In the next stage of the reverse cholesterol transport pathway, the liver selectively removes cholesteryl esters from the lipid-rich spherical HDL via binding of the HDL to the SR-BI receptor and lets the lipid-depleted HDL return to the circulation for additional rounds of cholesterol removal from peripheral cells. CETP also plays an important role in this pathway because a significant fraction of cholesterol that is removed from cells by HDL is transferred as cholesteryl esters onto LDL by CETP and is eventually removed from the circulation by hepatic LDL receptors.

CLINICAL SIGNIFICANCE

The main clinical significance of lipids is primarily associated with their role in the development of **coronary heart disease** (CHD). CHD is a disease in which plaque builds up inside the coronary arteries that supply oxygen-rich blood to heart muscle.

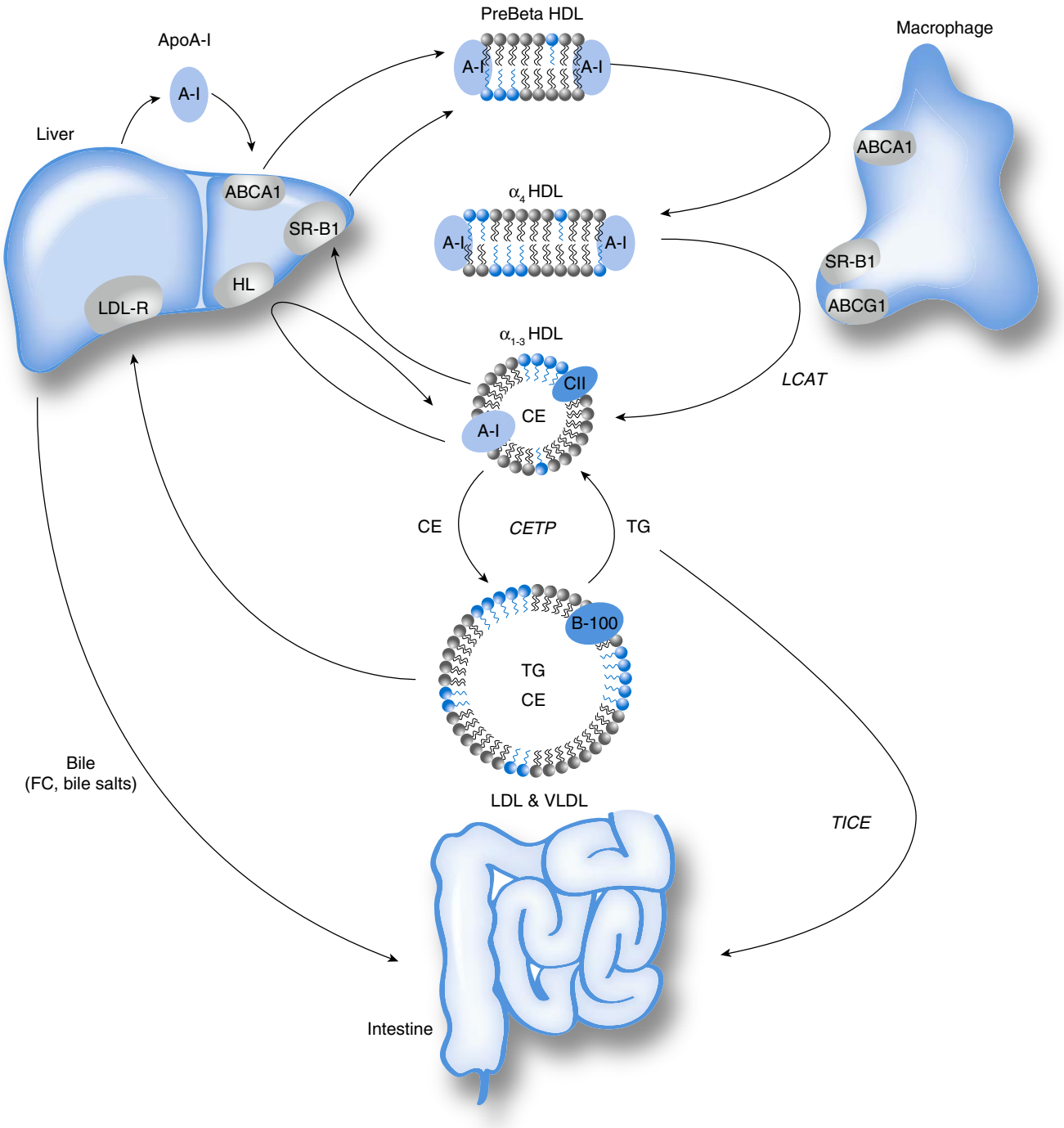


Fig. 23.20 Reverse cholesterol transport pathway. *ABCA1,* ATP-binding cassette transporter A1; *ABCG1,* ATP-binding cassette transporter GI; *A-I,* apolipoprotein A-1; *B,* apolipoprotein B-100; *CE,* cholesterol ester; *CETP,* cholesteryl ester transfer protein; *FC,* free cholesterol; *HL,* hepatic lipase; *HDL,* high-density lipoprotein; *LCAT,* lecithin cholesterol acyltransferase; *LDL,* low-density lipoprotein; *LDL-R,* LDL receptor; *SR-B1,* scavenger receptor B-I; *TG,* triglyceride; *TICE,* transintestinal cholesterol excretion; *VLDL,* very-low-density lipoproteins.

Numerous studies have established that when total cholesterol and LDL cholesterol (LDL-C) concentrations are high, the incidence and prevalence of CHD are also high. In contrast to LDL-C, increased HDL cholesterol (HDL-C) concentrations have been shown to be protective for CHD in epidemiologic studies. Because atherosclerosis begins at an early age and can take decades to manifest clinically, measurement of plasma lipids and lipoproteins is a valuable means of identifying individuals at risk for CHD and determining the most appropriate therapy.

Genetic Disorders of Lipoprotein Metabolism

Most patients with dyslipidemia do not have a single readily identifiable genetic mutation. Because of the complexity of lipoprotein metabolism, a multitude of environmental factors

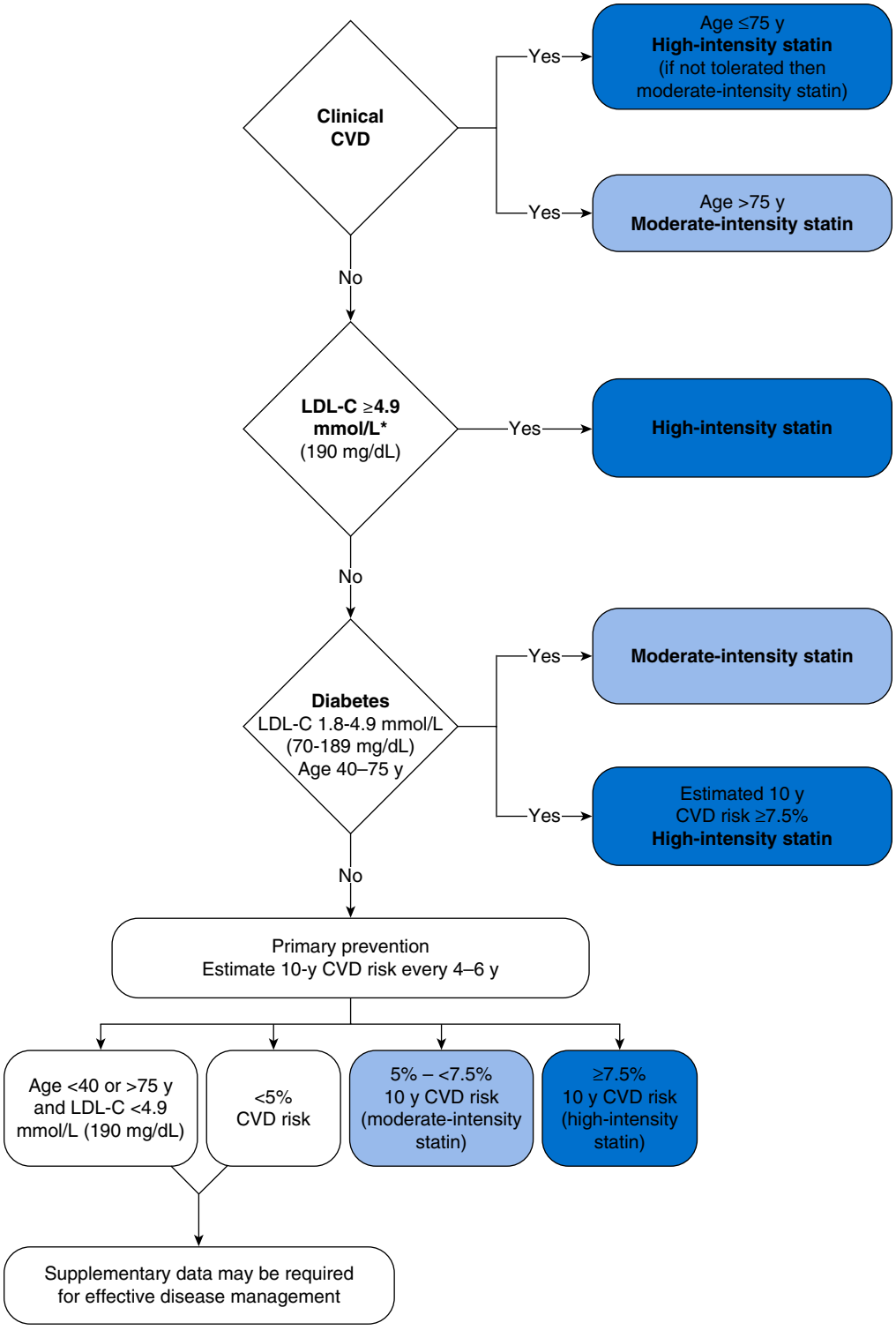


Fig. 23.21 Algorithm from the ACC/AHA 2013 guidelines on the assessment and treatment of cardiovascular disease in adults. *To convert mmol/L, divide by 38.66. (Modified from Stone, N. J., Robinson, J. G., Lichtenstein, A. H., et al. [2014]. 2013 ACC/AHA guidelines on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation*, *129*[Suppl. 2], S1–S45.)

and common genetic polymorphisms that vary in importance, depending on the individual, probably account for most cases of hypercholesterolemia. Many secondary causes of dyslipidemia are a consequence of relatively common disorders or conditions (Table 23.7). Although rare, established genetic causes of dyslipidemia have been identified and are illustrative of the role of lipid metabolism in the development of atherosclerosis.

Deficiency in Lipoprotein Lipase Activity

Deficient LPL activity due to mutations in the LPL gene is a rare autosomal recessive disorder characterized by

TABLE 23.7 Causes of Secondary Hyperlipidemia and Dyslipoproteinemia

Disorder	Cause
Exogenous	Drugs: corticosteroids, isotretinoin (Accutane), thiazides, anticonvulsants, β-blockers, anabolic steroids, certain oral contraceptives
	Alcohol
	Obesity
Endocrine and metabolic	Acute intermittent porphyria
	Diabetes mellitus
	Hypopituitarism
	Hypothyroidism
	Lipodystrophy
	Pregnancy
Storage disease	Cystine storage disease
	Gaucher disease
	Glycogen storage disease
	Juvenile Tay-Sachs disease
	Niemann-Pick disease
	Tay-Sachs disease
Renal	Chronic renal failure
	Hemolytic-uremic syndrome
	Nephrotic syndrome
Hepatic	Benign recurrent intrahepatic cholestasis
	Congenital biliary atresia
Acute and transient	Burns
	Hepatitis
	Acute trauma (surgery)
	Myocardial infarction
	Bacterial and viral infections
Others	Anorexia nervosa
	Starvation
	Idiopathic hypercalcemia
	Klinefelter syndrome
	Progeria (Hutchinson-Gilford syndrome)
	Systemic lupus erythematosus
	Werner syndrome

hyperchylomicronemia (type I pattern according to Fredrickson), with triglyceride concentrations reaching as high as 10,000 mg/dL (113 mmol/L). LPL is critical for the hydrolysis of triglycerides on chylomicrons and VLDL. This disorder is often first diagnosed in childhood, usually after recurrent episodes of severe abdominal pain and repeated attacks of pancreatitis. Eruptive xanthomas (a skin condition in which certain fats build up under the surface of the skin) and lipemia retinalis (a condition that occurs in patients when their plasma triglyceride concentrations exceed 2000 and 4000 mg/dL [22.6 to 45.2 mmol/L]) may be seen. The concentration of triglycerides often shows great fluctuation in response to diet and other factors that are not well understood. Individuals with this disorder are not predisposed to atherosclerotic disease, most likely because the chylomicron particles are too large to enter the vessel wall. The diagnosis is made by determination of LPL activity in plasma collected after heparin is injected into patients to release the LPL that is bound to heparin sulfate and other glycosaminoglycans on the surface of endothelial cells.

Mutations in apo C-II, the main activator of LPL, also result in impairment of chylomicron catabolism, although this usually is less severe than with LPL gene mutations. The diagnosis of this rare autosomal recessive condition is made by demonstrating low LPL activity in postheparin plasma that is restored after the addition of apo C-II to the LPL assay mixture.

Familial Combined Hyperlipidemia

Familial combined hyperlipidemia (FCHL) is an autosomal dominant disorder that accounts for as many as 10% to 15% of individuals with premature CHD. Families with FCHL often have increased plasma concentrations of total and LDL-C (type IIa) or triglyceride/VLDL (type IV), or both (type IIb). Lipoprotein patterns also vary within an individual over time. In all cases, apo B-100 concentrations are increased because of overproduction. LDL particles in these patients tend to be small and dense because of a decreased lipid-to-protein ratio. LDL-C is usually only modestly increased to approximately 190 mg/dL (2.14 mmol/L), which is less than what is typically observed in heterozygous familial hypercholesterolemia (FH) (350 mg/dL; 3.95 mmol/L). The concentration of triglycerides usually is between 200 and 400 mg/dL (2.26 and 4.52 mmol/L) but may be significantly higher. The underlying genetic basis of FCHL is not known; it is mostly likely a polygenetic disorder.

Familial Hypertriglyceridemia

Familial hypertriglyceridemia (FHTG) is an autosomal dominant disease characterized by a moderate increase in serum triglycerides. Overproduction of large VLDL particles with abnormally high triglyceride content is thought to be responsible for this disorder, but decreased lipolysis is sometimes a contributory factor. Because of the hypertriglyceridemia, plasma HDL-C is often notably decreased. This disorder is inherited in an autosomal dominant pattern with delayed expression, but it is a polygenic disorder with an estimated frequency in the population of approximately 1 in 500.

Type V Hyperlipoproteinemia

Type V hyperlipoproteinemia is characterized by an increase in both chylomicrons and VLDL and typically presents in adults. It is very common that acute pancreatitis develops subsequent to the massive hypertriglyceridemia (in the range of 1000 to 2000 mg/dL; 11.3 to 22.6 mmol/L). Type V HLP may also result from an inherited disorder that has an incidence of approximately 1 in 500. The disease appears to be an autosomal dominant disorder with increased production and/or decreased removal of VLDL. The activity of LPL in these individuals may be normal or low, and the plasma concentration of apo C-II is normal. Clinical presentations include (1) eruptive xanthomas, (2) lipemia retinalis, (3) pancreatitis, and (4) abnormal glucose tolerance. Most of these patients have a modestly increased risk of cardiovascular disease.

Dysbetalipoproteinemia (Type III)

Dysbetalipoproteinemia, also called *type III hyperlipoproteinemia* or remnant removal disease, is an autosomal recessive disorder caused by a defect in the removal of lipoprotein remnants from both chylomicrons and VLDL. Apo E present on the surface of lipoprotein remnant particles interacts with specific hepatic receptors and facilitates the removal of these particles. Apo E exists in three common polymorphisms or variants, designated E_2, E_3, and E_4. Some individuals with dysbetalipoproteinemia are homozygous for the apo E_2 isoform, which does not efficiently bind to hepatic remnant receptors, thus leading to accumulation of remnant particles. Although rare, genetic mutations in the *APOE* gene have been associated with the disorder. The remnant particles that accumulate are enriched in cholesterol, have a density less than 1.006 g/mL, and are commonly referred to as β-VLDL, or floating β-lipoprotein. Dysbetalipoproteinemia has a late onset and may manifest in patients with a secondary cause of dyslipidemia or obesity. The most distinctive clinical feature of dysbetalipoproteinemia is the presence of palmar xanthomas—yellow fat deposits in the creases of the palms. Tuberous and tuberoeruptive xanthomas may also occur but are not unique to this syndrome. Premature atherosclerosis commonly develops, particularly in the lower extremities. The incidence of dysbetalipoproteinemia is approximately 0.1% in the general population. However, apo E_2 homozygosity is seen in approximately 1% of the population in North America; thus the occurrence of defective alleles is necessary but not sufficient to produce the disorder.

Familial Hypercholesterolemia

FH is caused by defects in the LDL receptor pathway, which binds and removes LDL from the circulation. LDL thus accumulates in the plasma, resulting in its increased deposition in the skin, tendons, and arteries, where it causes atherosclerosis. LDL particles tend to be larger and carry increased amounts of cholesterol. Triglyceride concentration may be normal or only slightly increased, and HDL-C concentration is slightly decreased. Most of these patients have gene defects in the LDL receptor gene *(LDL-R)*. Less commonly, defects in two other genes, *RHOA* (formerly called *ARH-1*) and *PCSK9*, which code for proteins involved in internalization or processing of the LDL receptor, may cause FH. Mutations in the LDL receptor gene *(LDL-R)* and in *PCSK9* are inherited in an autosomal codominant pattern. Homozygous FH patients are severely affected, whereas heterozygotes usually have a milder phenotype but are still clinically affected usually at a later age. Defects in *RHOA* are inherited in an autosomal recessive pattern.

Heterozygous FH due to mutations in the LDL receptor gene is one of the most common genetic disorders, with an estimated incidence of 1 in 500 in the United States. The mean plasma LDL-C in children and adult heterozygotes is usually 2 to 3 times that of normal individuals, whereas LDL-C of homozygotes is usually fourfold to sixfold above normal. In heterozygotes, xanthomas appear toward the end of the second decade of life, and clinical manifestations of atherosclerotic disease often are noted during the fourth decade. In homozygotes, cutaneous xanthomas often develop by 4 years of age, if they are not already present at birth. Left untreated, death from myocardial infarction generally occurs in homozygotes before the end of the second or third decade of life.

Familial Defective Apolipoprotein B-100

Familial defective apo B-100 is an autosomal dominant disorder, the result of mutations in the gene *APOB*, which reduce its affinity for the LDL receptor. LDL-C is increased, but triglycerides and HDL-C are usually within reference intervals. Like those with FH, these individuals have an increased incidence of CHD, but overall, the dyslipidemia in this disorder is less aggressive than in FH. Clinical differentiation between this disorder and heterozygous FH is sometimes difficult, but management of the two disorders is similar. The frequency of this mutation is 1:500 to 1:600 in hypercholesterolemic individuals from populations of European descent, but it is very rare in non-Europeans.

Hypoalphalipoproteinemia

Hypoalphalipoproteinemia, or low HDL-C, is caused by several genetic defects and is often but not always associated with an increased incidence of CHD. Mutations or deletions of the *APOA1* gene are a rare cause of hypoalphalipoproteinemia. LCAT deficiency is also associated with low HDL. These patients often have cloudy corneas as a result of infiltration of lipid, as well as renal disease because of the production of abnormal lipoprotein particle called LpX that becomes trapped in the glomerulus.

Tangier disease is a rare autosomal recessive disorder that is associated with a notable reduction in HDL. The major clinical signs of Tangier disease are (1) hyperplastic orange tonsils, (2) splenomegaly, and (3) peripheral neuropathy. Deposition of cholesteryl esters is increased in various tissues of the body, particularly macrophages, which form foam cells. Tangier disease is caused by mutations in the *ABCA1* gene for the ABCA1 transporter, which mediates the first step of the reverse cholesterol transport pathway—efflux of cholesterol from cells (see Fig. 23.20).

Sitosterolemia

Sitosterolemia is a rare inherited disease characterized by overabsorption of dietary sitosterol and increased hepatic secretion of sitosterol. Consequently, cholesterol synthesis and elimination are altered and cholesterol levels are increased. Homozygotes for this defect have increased plasma sitosterol concentrations (requiring highly specialized laboratory measurement), xanthomas, and very early onset of atherosclerotic cardiovascular disease (ASCVD).

Lp(a) and LpX

There are two forms of hyperlipidemia not covered by the Fredrickson classification: increase in Lp(a) and the presence of LpX. The size variation in Lp(a) and consequently the level of Lp(a) is largely determined by the variant copy

numbers of "kringle" repeats in the APOA gene that encodes apo(a). Elevated Lp(a) concentrations impose increased risk of ASCVD and calcific aortic valve diseases. Populations of African descent are shown to have higher Lp(a) levels than do other populations, although it may not correspond to an increased risk of heart diseases. Markedly elevated Lp(a) also tends to run in certain families, but the inheritance pattern is not clearly defined.

LpX is produced in severe obstructive biliary tract diseases as an abnormal lipoprotein. It appears that when the neutral cholesteryl esters are insufficient to form a lipid core in lipoproteins, LpX is generated. It has a phospholipid bilayer instead of the single phospholipid layer in normal lipoproteins, and it lacks apo B-100 that is present on other lipoproteins in the endogenous pathway, which may confound the measurement or calculation of cholesterol fractions.

ADULT MANAGEMENT OF LIPOPROTEIN DISORDERS

The initial diagnosis of a dyslipoproteinemia and determination of the best treatment approach for a patient have been based largely on measurements of (1) total cholesterol, (2) triglyceride, (3) HDL-C, and (4) LDL-C; this is commonly referred to as a *lipid panel.* However, lipid and lipoprotein test results must be interpreted in the context of a medical history. Medical history and other laboratory test results are also important for determining whether a dyslipoproteinemia is the result of a primary lipoprotein disorder or is a consequence of one or more of the secondary causes of hyperlipidemia (see Table 23.7), which, if present, should also be addressed. In contrast to most other laboratory tests, hypercholesterolemia is defined on the basis of findings from epidemiologic studies that were used to establish a desirable concentration for reducing CHD risk (Table 23.8), rather than a reference study of normal individuals. Cholesterol screening is recommended every 5 years for all adults older than 20 years and should involve measurements of fasting (1) total cholesterol, (2) triglyceride, (3) HDL-C, and (4) calculated or measured LDL-C. If a fasting sample is not available initially, only total cholesterol and HDL-C should be considered.

Next, an assessment should be made of the risk for CHD. This is based on clinical evidence of existing CHD or the presence of conditions that are closely associated with CHD, such as (1) symptomatic carotid artery disease, (2) peripheral vascular disease, and (3) abdominal aortic aneurysm, which are called *CHD risk equivalents.* Risk assessment for CHD is also based on the presence of known risk factors for CHD, such as (1) hypertension, (2) smoking, and (3) family history (Box 23.1).

Therapeutic lifestyle changes (Box 23.2) serve as the cornerstone of therapy for lipid disorders. The concentration of LDL-C is used both to decide the most appropriate therapy and to monitor the effectiveness of treatment. Adult

TABLE 23.8 Adult Treatment Panel (ATP) III Classification of Low-Density Lipoprotein, Total, and High-Density Lipoprotein Cholesterol (mg/dL)

LDL cholesterol	<100	Optimum
	100–129	Near or above optimum
	130–159	Borderline high
	160–189	High
	≥190	Very high
Total cholesterol	<200	Desirable
	200–239	Borderline high
	≥240	High
HDL cholesterol	<40	Low
	≥60	High

HDL, High-density lipoprotein; *LDL,* low-density lipoprotein.
Modified from Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. (2001). Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). *Journal of the American Medical Association, 285,* 2486–2497.

BOX 23.1 Major Risk Factors (Exclusive of Low-Density Lipoprotein Cholesterol)

- Cigarette smoking
- Hypertension (blood pressure ≥140/90 mm Hg or on antihypertensive medication)
- Low HDL cholesterol (<40 mg/dL)[1]
- Family history of premature CHD (CHD in male first-degree relative <55 years; CHD in female first-degree relative <65 years)
- Age (men ≥45 years; women ≥55 years)

CHD, Coronary heart disease; *HDL,* high-density lipoprotein.
Modified from Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. (2001). Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). *Journal of the American Medical Association, 285,* 2486–2497.

BOX 23.2 Therapeutic Lifestyle Changes for the Prevention of Coronary Heart Disease

- Diet
  - Saturated fat <7% of calories, cholesterol <200 mg/day
  - Consider increased viscous (soluble) fiber (10–25 g/day) and plant stanols/sterols (2 g/day) as therapeutic options to enhance LDL lowering
- Weight management
- Increased physical activity

[a]HDL cholesterol ≥60 mg/dL counts as a "negative" risk factor; its presence removes one risk factor from the total count.
HDL, High-density lipoprotein; *LDL,* low-density lipoprotein.
Modified from Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. (2001). Executive Summary of the Third Report of the National Cholesterol Educational Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). *Journal of the American Medical Association. 285,* 2486–2497.

treatment guidelines for hypercholesterolemia are illustrated in Table 23.9. Note that aggressiveness of treatment depends on the risk category of the patient and his or her starting LDL-C concentration. Patients at highest risk for CHD (10-year risk >20%) and those who already have clinical evidence of CHD or a CHD risk equivalent have the lowest LDL-C treatment threshold and the lowest LDL-C target goal. Ideally, for such patients, LDL-C should be less than 100 mg/dL (2.59 mmol/L) after therapy; this will likely involve some type of drug treatment. Patients in an intermediate risk category (10-year risk <20%) have a higher concentration for their LDL-C target goal and in some cases, depending on their starting LDL-C, may be treated only by therapeutic lifestyle changes. For patients in the lowest risk category (0 to 1 risk factors), desirable LDL-C is less than 160 mg/dL (4.14 mmol/L), and drug therapy should be considered only if initial LDL-C is greater than 190 mg/dL (2.15 mmol/L). Specific recommendations have been put forth for the frequency of monitoring lipids and lipoprotein tests and for when drug therapy should be tried for patients who do not reach their LDL-C target treatment goals with just therapeutic lifestyle changes.

Increased triglycerides and low HDL-C are considered conditions that independently alter the risk for CHD and may alter the recommended treatment. However, in both cases the primary goal of therapy should be to lower LDL-C and secondarily to improve the increased triglycerides or low HDL, if these conditions persist after LDL-C is lowered. Each patient at risk for CHD should also be assessed for the presence of metabolic syndrome and should be treated for underlying causes and associated symptoms. Other newly developed lipoprotein tests, such as (1) Lp(a), (2) remnant lipoproteins, and (3) small dense LDL, and other markers, such as (4) C-reactive protein (CRP), and (5) homocysteine, may also be valuable in CHD risk stratification, particularly for patients who are at borderline or intermediate risk, based on conventional lipid and lipoprotein tests.

A wide variety of pharmacologic agents for cholesterol lowering in adults are prescribed individually or in combination to lower LDL-C, including (1) bile acid–binding resins (cholestyramine and colestipol), (2) niacin, (3) gemfibrozil, (4) ezetimibe, and (5) HMG-CoA reductase inhibitors (e.g., atorvastatin, fluvastatin, lovastatin, pravastatin, rosuvastatin, simvastatin, pitavastatin); agents in the latter group have been found to reduce LDL-C by as much as 40%. Some of these drugs are better tolerated by individual patients than others; all have demonstrated long-term safety in adults and have been shown to decrease CHD risk. Niacin has been the most effective therapy for raising HDL-C, but the value of raising HDL-C has been called into question by trials of the CETP inhibitors, which dramatically increase HDL-C but have not decreased CHD events. Monoclonal antibody therapy against PCSK9 has recently been approved and results in significant decreases of LDL-C even for patients already treated with statins. Because of the expense of this therapy, it is recommended only for statin-intolerant patients and or patients with severe dyslipidemias that are not adequately controlled on conventional drug therapy. Niacin, fibrates, and fish oil can also lower plasma triglycerides but so far have not been shown to be useful for reducing CHD events but should be considered in patients with severe hypertriglyceridemia who are at risk for pancreatitis.

In 2013 the National Heart, Lung, and Blood Institute (NHLBI) passed the mission of cardiovascular guideline production to a new task force from the American College of Cardiology/American Heart Association (ACC/AHA), which issued new recommendations on the "Treatment of Blood Cholesterol to Reduce Atherosclerotic Cardiovascular Risk in Adults" on "the Assessment of Cardiovascular Risk," and "for the Management of Overweight and Obesity in Adults." The new statement (Fig. 23.21), which for the most part considered only level I evidence from randomized, blinded, clinical trials and not expert opinion, eliminated numerical treatment targets for LDL-C and non–HDL-C and introduced a new pooled risk calculator for treatment initiation decisions based on cardiovascular disease, which includes not only CHD but also stroke and peripheral vascular disease. The new risk calculator also has a separate algorithm for African Americans. Patients deemed to be at high risk for CHD, including those with an LDL-C greater than 190 mg/dL (4.91 mmol/L), are advised to initiate lifestyle changes and drug therapy, defined as high-dose

TABLE 23.9 Low-Density Lipoprotein Cholesterol Goals and Cut Points for Therapeutic Lifestyle Changes and Drug Therapy in Different Risk Categories

Risk Category	LDL Goal (mg/dL)	LDL Concentration at Which to Initiate Therapeutic Lifestyle Changes (mg/dL)	LDL Concentration at Which to Consider Drug Therapy (mg/dL)
CHD or CHD risk equivalents (10-year risk >20%)	<100	≥100	≥300 (100–129; drug optional)
2+ risk factors (10-year risk ≤20%)	<130	≥130	10-year risk 10%–20%: ≥130 10-year risk <10%: ≥160
0–1 risk factor (10-year risk <10%)	<160	≥160	≥190 (160–189; LDL-lowering drug optional)

CHD, Coronary heart disease; *LDL,* low-density lipoprotein.

Modified from Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. (2001). Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). *Journal of the American Medical Association, 285,* 2486–2497.

statin in an attempt to reduce LDL-C by 50% from baseline (see Fig. 27.21). Other high-risk groups that are potential candidates for high-dose statins (or moderate dose if they cannot tolerate high-dose statins) are patients with preexisting cardiovascular disease (CVD), and diabetic patients (ages 40 to 75). Primary prevention patients, between age 40 and 75 years, who do not fall into one of these high-risk categories should have their 10-year CVD risk calculated and should be treated with high- or moderate-dose statins if the risk is greater than 7.5% and treated with moderate-dose statins if their 10-year CVD risk falls between 5% and 7.5%. Primary prevention patients with a 10-year risk of less than 5% or if they fall outside the age range for the risk calculator (<40 years or >75 years) are generally not recommended for statin therapy but could be considered for lifestyle changes and or statin therapy, depending on other considerations, such as a strong family history of premature CVD or the results of ancillary testing, such as high-sensitivity C-reactive protein (hsCRP). Compared with the previous Adult Treatment Panel guidelines, it is estimated that significantly more patients will be now treated with statins by the new ACC/AHA guidelines but some younger patients who would appear to be at risk, based on the older guidelines, would no longer be treated. Follow-up laboratory testing is recommended by the new guidelines only to judge compliance and to ensure that at least 50% reduction in LDL-C is achieved. Specific cholesterol treatment goals for LDL-C and non–HDL-C are no longer advised. Overall, use of nonstatin therapies is also discouraged by these new guidelines, although this may change in the future with new information, such as clinical trials that show that ezetimibe added to statin was more efficacious compared with statin monotherapy in reducing cardiovascular events in an acute coronary syndrome setting. There was no specific guidance on the management of patients with moderate to high triglycerides or increased Lp(a). The ACC/AHA statement also took no position for or against using apolipoprotein B, LDL-particle number, or Lp(a). However, other national and international organizations have included other CVD biomarkers in their lipid and lipoprotein recommendations, particularly for patients at intermediate risk. Furthermore, other organizations, such as National Lipid Association and American Association of Clinical Endocrinology did not endorse the ACC/AHA guidelines and have published their own guidelines.

Management of Lipoprotein Disorders in Pediatrics

Because the process of atherosclerosis begins in children, and CHD risk factors that start in childhood, such as obesity and dyslipidemia, often persist into adulthood, it is important to identify children at risk for CHD, with the expectation that early intervention will be the most effective way to prevent future disease. This goal has to be tempered by the concern that unlike in adults, clinical studies supporting the long-term effectiveness and safety of any type of therapy, particularly the use of lipid-lowering medications in children, are limited.

The 2001 guidelines for cardiovascular health and risk reduction in the pediatric population from NHLBI are integrated guidelines that make age-specific recommendations on the multiple risk factors that contribute to CHD. They are aimed at primary care physicians and other health professionals who treat children. They provide guidance not only on the diagnosis and management of dyslipidemias but also on how to monitor (1) family history, (2) tobacco exposure, (3) degree of physical activity, (4) diet, (5) growth charts, (6) blood pressure, and (7) diabetes, as well as (8) how to educate children and parents on healthy lifestyles.

Previous guidelines did not recommend universal lipid testing and instead relied upon identifying children at risk on the basis of family history of CHD or dyslipidemia. Although family history integrates (1) genetic, (2) behavioral, and (3) environmental factors, it is often inadequate and has been shown to miss as many as half of children at risk for CHD. Current guidelines recommend universal lipid screening between ages 9 and 11 years and then later between 18 and 21 years of age. Age 9 to 11 years was chosen because lipids often change after puberty and because this is the age at which children with severe lipid disorders should first be considered as possible candidates for drug therapy. Age 18 to 21 years was chosen as a second time to screen because some lipid disorders do not become manifest until late teenage years or young adulthood and because after this point in time, many individuals may not participate in lipid screening until many years later, when advanced disease has already developed. Because it is more convenient, particularly for young children, it is recommended that a nonfasting sample be analyzed to determine non–HDL-C; calculated as total cholesterol minus HDL-C and has been shown in both children and adults to be just as effective as LDL-C as a CHD risk marker. For other age groups, lipid testing is recommended on a fasting sample, only if children have a strong family history of CHD and/or have a parent with a dyslipidemia or a condition that puts them at high risk for CHD, such as diabetes.

As in adult guidelines, recommended interventions are calibrated to the degree of CHD risk; lifestyle changes, such as low-fat diet and increasing physical activity, are recommended first for all patients (Fig. 23.22). If, after a 3-month period, lifestyle changes prove ineffective in reaching the LDL-C target goal, which varies depending on the overall risk, then drug treatment with a statin and/or a bile acid sequestrant should be considered for children 10 years of age or older. All children, even those younger than age 10 years, who present with a marked increase in serum lipids (LDL-C >250 mg/dL [6.46 mmol/L]) may have a genetic lipid disorder, should be referred to a lipid specialist and may require additional types of drugs or other treatment approaches, such as LDL apheresis in the case of FH. Because of the high prevalence of obesity in the pediatric population, the most common pattern of dyslipidemia in children consists of (1) a moderate to severe increase in triglycerides, (2) a mild increase in LDL-C, and (3) decreased HDL-C. In general, patients respond well to changes in diet and weight loss and often do not require drug therapy. Children who present with a marked increase in triglycerides to greater than 500 mg/dL (>5.65

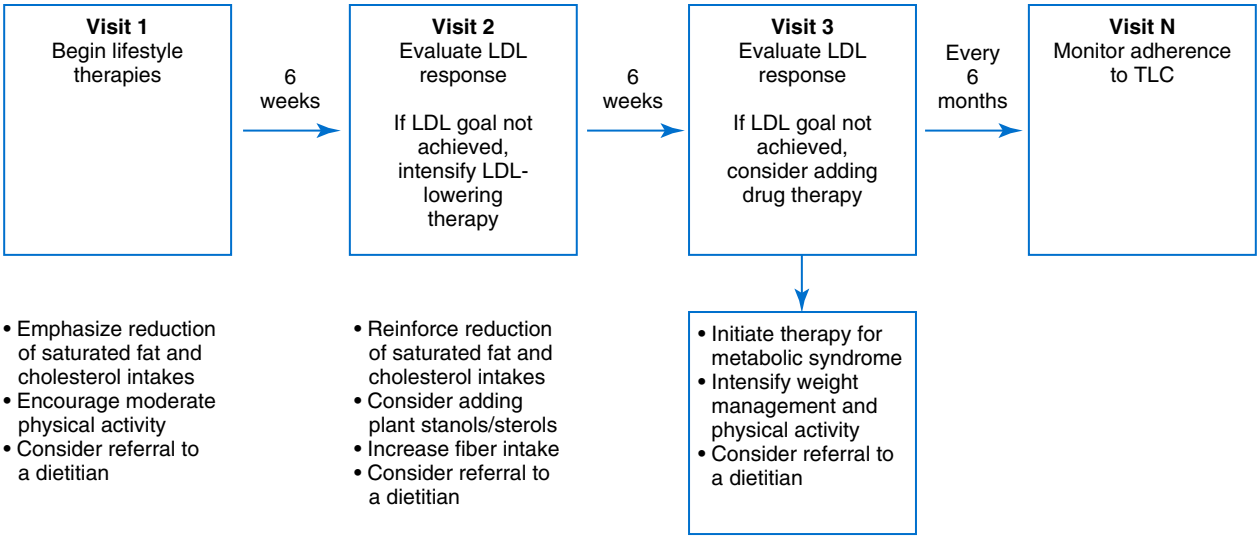


Fig. 23.22 Model of steps in therapeutic lifestyle change. *LDL*, Low-density lipoprotein. (Modified from Executive summary of the third report of the Expert Panel on Blood Cholesterol Levels in Children and Adolescents, National Cholesterol Education Program. Lipid Metabolism Branch, Division of Heart, Lung, and Blood Institute. NIH Publication No. 01-3670. US Department of Health and Human Services, Public Health Service, National Institutes of Health. Bethesda, MD: National Institutes of Health, 2003.)

mmol/L) may have a primary genetic defect, such as LPL deficiency, are at risk for pancreatitis and should be seen by a lipid specialist. It is recommended that children treated with lipid-lowering drugs start with the lowest possible dose; when placed on statins, hepatic transaminases and creatine kinase should be measured to monitor for hepatic and muscle toxicity (Fig. 23.23). Decisions about placing a child on a lipid-lowering drug should be based on an average of at least two fasting serum lipid profiles that are at least 2 weeks apart but no more than 3 months apart.

ANALYSIS OF LIPIDS, LIPOPROTEINS, AND APOLIPOPROTEINS

As risk-related cutoff points (see Table 23.9) were developed, much effort has gone into improvement and standardization of lipid and lipoprotein assays to ensure their proper performance (Table 23.10) and to ensure comparability of test results between different methods and different clinical laboratories, thus reducing misclassification of patients' CHD risk.

Currently, worldwide, the policy that most laboratories follow is an overnight fast of a minimum of 8 hours prior to obtaining samples for testing the lipid profile. The European Atherosclerosis Society/European Federation of Clinical Chemistry and Laboratory Medicine (EAS/EFLM) recently published a consensus statement promoting the use of nonfasting lipid profiles for the assessment of cardiovascular disease risk. The rationale for the EAS/EFLM stance is ease of collection and greater compliance in patient preparation for lipid testing. This statement addresses the perceived limitations of testing nonfasting samples by referencing data from studies that evaluated lipid variation in postprandial versus fasting states; showing that there is no change in HDL-C, apo A1, apo B, or Lp(a) and minimal changes in triglycerides, total cholesterol, calculated remnant cholesterol, non–HDL-C, and LDL-C. In addition, the publication addresses the influence of the postprandial state on the assessment of cardiovascular disease risk by citing studies where various nonfasting lipid parameters where monitored relative to clinical outcomes.

Cholesterol Assays

Enzymatic methods have become the assays of choice for the routine measurement of cholesterol. A commonly used three-step procedure is shown later. Enzyme reagents are mixed with serum or plasma and are incubated under controlled conditions for color development; absorbance is measured in the visible portion of the spectrum, generally at approximately 500 nm. The reagents typically use a bacterial cholesteryl ester hydrolase to cleave cholesteryl esters:

$$\text{Cholesteryl ester} + H_2O \xrightarrow{\text{Chloesteryl ester hydrolase}} \text{cholesterol} + \text{fatty acid} \quad (23.1)$$

The 3-OH group of cholesterol is then oxidized to a ketone in an oxygen-requiring reaction catalyzed by cholesterol oxidase:

$$\text{Cholesterol} + O_2 \xrightarrow{\text{Chloesteryl ester oxidase}} \text{cholest-4-en-3-one} + H_2O_2 \quad (23.2)$$

H_2O_2 is measured in a peroxidase-catalyzed reaction that forms a colored dye:

$$H_2O_2 + \text{phenol} + \text{4-aminoantipyrine} \xrightarrow{\text{Peroxidase}} \text{quinoneimine dye} + 2H_2O \quad (23.3)$$

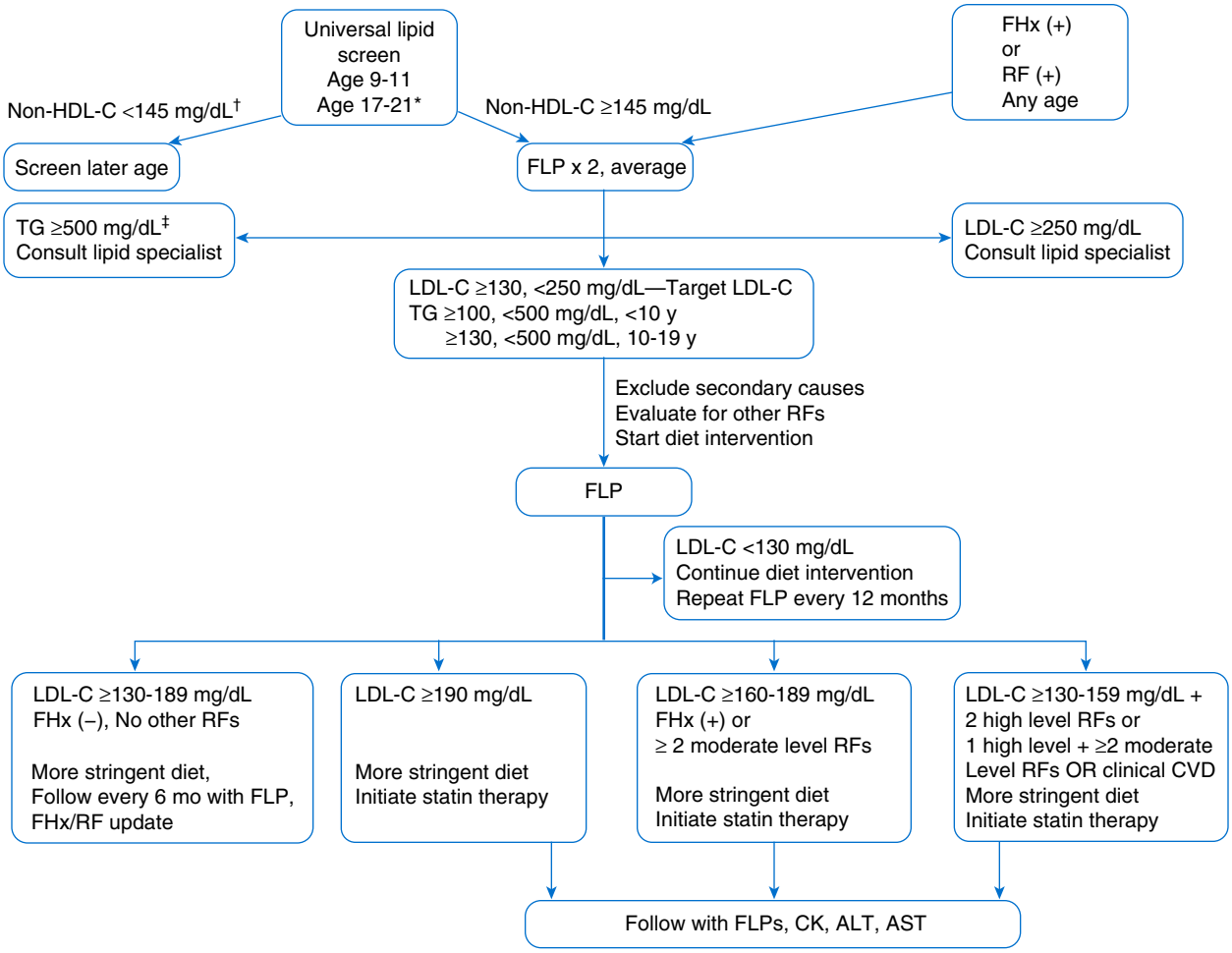


Fig. 23.23 Algorithm for assessment of coronary heart disease risk in children based on lipid screening. *For ages 19–21, non–HDL-C ≥190 mg/dL is the recommended cut point. *FHx(+)*, Positive family history; *FLP*, fasting lipid profile; *RF*, risk factor. Other positive risk factors or risk equivalents for cardiovascular disease include RF(+). *ALT*, Alanine aminotransferase; *AST*, aspartate aminotransferase; *CK*, creatine kinase. (Modified from Expert Panel on Integrated Guidelines for Cardiovascular Health and Risk Reduction in Children and Adolescents; National Heart, Lung, and Blood Institute. [2011]. Expert Panel on Integrated Guidelines for Cardiovascular Health and Risk Reduction in Children and Adolescents. Summary report. *Pediatrics*, *128*, S1–S44.)

TABLE 23.10 National Cholesterol Education Program Recommendations for Analytical Performance of Lipid and Lipoprotein Measurements

		CONSISTENT WITH	
	Total Error (%)	Bias (%)	CV (%)
Cholesterol	8.9	≤±3	≤3
Triglycerides	≤15	≤±5	≤5
HDL cholesterol	≤13	≤±5	≤4
LDL cholesterol	≤12	≤±4	≤4

CV, Coefficient of variation; *HDL*, high-density lipoprotein; *LDL*, low-density lipoprotein.

Enzymatic methods are subject to interference from other colored compounds or from those that compete with the oxidation reaction, such as (1) bilirubin, (2) ascorbic acid, and (3) hemoglobin. Assays are usually linear up to approximately 600 to 700 mg/dL (15.54 to 18.13 mmol/L).

Triglycerides

Triglycerides are also routinely measured with enzyme reagents. Several different enzyme sequences have been used. With all of these methods, the first step is the lipase-catalyzed hydrolysis of triglycerides to glycerol and fatty acids

$$\text{Triglycerides} + 3H_2O \xrightarrow{\text{Lipase}} \text{glycerol} + 3\ \text{fatty acids} \quad (23.4)$$

Glycerol is then phosphorylated in an ATP-requiring reaction catalyzed by glycerokinase:

$$\text{Glycerol} + \text{ATP} \xrightarrow{\text{Glycerokinase}} \text{glycerophosphate} + \text{ADP} \quad (23.5)$$

With the most commonly used methods, glycerophosphate is oxidized to dihydroxyacetone and H_2O_2 in a glycerophosphate oxidase–catalyzed reaction:

$$\text{Glycerophosphate} + O_2 \xrightarrow{\text{Glycerophosphate oxidase}} \text{dihydroxyacetone} + H_2O_2 \quad (23.6)$$

and H_2O_2 formed in the reaction is measured as described in reaction (3).

Enzymatic triglyceride methods are fairly specific in that they do not detect glucose or phospholipids. They are linear in the concentration range up to approximately 700 mg/dL (7.91 mmol/L) and, when automated, are operated with coefficients of variation (CVs) in the range of approximately 2% to 3%. Because glycerol is a product of normal metabolic processes, it is present in serum. Thus the measured quantity of triglycerides in serum is overestimated slightly, if not corrected for endogenous glycerol. In healthy individuals, endogenous glycerol represents the equivalent of less than 10 mg/dL (0.11 mmol/L) of triglyceride; therefore the error due to glycerol is not usually clinically significant. Some laboratories use an alternative triglyceride assay, in which the endogenous glycerol is first "blanked out" by adjustment of the calibrators to compensate for an average bias or by enzymatic consumption of glycerol in a prereaction step before triglycerides are measured.

High-Density Lipoprotein Cholesterol

The concentration of cholesterol associated with HDL is determined after the non-HDL lipoproteins are physically removed or effectively masked.

Traditionally, non-HDL lipoproteins—VLDL, IDL, Lp(a), LDL, and chylomicrons—were precipitated with polyanions. Polyanions react with positively charged groups on lipoproteins, and this interaction is further facilitated in the presence of divalent cations, such as magnesium, to create a precipitate. The precipitate is removed by centrifuging for at least 45,000 *g*-minutes, and cholesterol is measured enzymatically in the supernatant. Several polyanion-divalent cation combinations have been used, including (1) heparin sulfate–$MnCl_2$, (2) dextran sulfate–$MgCl_2$, and (3) phosphotungstate-$MgCl_2$. HDL-C assays are considered inaccurate in samples containing triglyceride concentrations greater than 400 mg/dL (4.52 mmol/L).

Direct HDL-C assays, also known as *homogeneous assays*, are most commonly used to measure HDL-C. In contrast to precipitation-based assays, no physical separation of HDL from non-HDL fractions occurs. Instead, HDL-C is selectively measured by effective masking or promoting the consumption of cholesterol from non-HDL fractions so that it does not react with the enzymes used to measure cholesterol on HDL. This is achieved by a variety of methods, depending on the type of assay. For example, some assays involve the use of antibodies or various polymers or complexing agents, such as cyclodextrin, that shield the cholesterol in the non-HDL fractions from reacting with cholesterol-measuring enzymes. Some assays depend on modifications of cholesteryl esterase and cholesterol oxidase that make them more selective for HDL-C. Finally, some assays use a blanking step that enzymatically consumes cholesterol from non-HDL fractions. Unlike the precipitation HDL-C assays, no sample pretreatment step is performed with direct assays and thus they can be fully automated. However, these assays sometimes show relatively poor performance compared with reference methods on dyslipidemic samples, particularly those with high triglycerides.

Low-Density Lipoprotein Cholesterol

Until recently, LDL-C was determined indirectly. In the most widely used indirect method, (1) total cholesterol, (2) triglycerides, and (3) HDL-C are measured in a fasting sample, and LDL-C is calculated from the primary measurements by the empirical Friedewald equation:

$$\text{LDL-C} = [\text{Total cholesterol}] - [\text{HDL Cholesterol}] - [\text{Triglycerides}]/5 \quad (23.7)$$

where all concentrations are given in milligrams per deciliter. Triglycerides/2.22 is used when LDL-C is expressed in millimoles per liter. The factor (triglycerides)/5 is an estimate of the VLDL cholesterol concentration and is based on the average ratio of triglycerides to cholesterol in VLDL.

In practice, the Friedewald equation should not be used (1) in samples that have triglyceride concentrations greater than 400 mg/dL (4.52 mmol/L), (2) in samples that contain significant quantities of chylomicrons (nonfasting specimen), (3) in patients with dysbetalipoproteinemia, and (4) in patients with low total cholesterol (LDL <70 mg/dL). In these cases the factor (triglyceride)/5 does not provide an accurate estimate of VLDL cholesterol, and this leads to large errors in calculated LDL-C. As with HDL-C, several methods are now used for the direct measurement of LDL-C. In general, these assays yield results similar to those obtained by calculating LDL-C and performing the β-quantification reference method, which is based on ultracentrifugation, but performance may not be as good on high-triglyceride samples. Evidence suggests that different direct assays may not be specific and may include some of the other apo B–containing lipoproteins or may miss some LDL subfractions, such as the more atherogenic small, dense LDLs.

Measurement of Apolipoproteins

Immunoturbidimetric and immunonephelometric assays are typically used in measuring apo A-I and apo B-100 (see Chapter 15 for additional information on these techniques). Alternatively, more sensitive techniques, such as enzyme-linked immunosorbent assay (ELISA) and radioimmunoassay (RIA), are usually more suitable for apolipoproteins that are present in significantly lower concentrations, such as apo C-I and apo C-II. Considerable effort has been expended in the past to standardize apo A-I and B-100 measurements, leading to significant improvement in the overall performance of these assays, but they are still not widely used for CHD risk assessment. Tests for apo B-100 and apo A-I are now well standardized and calibrated to World Health Organization reference materials. Immunoassays for total apo E are not commonly done, but genotyping of apo E is used to identify the apo E_4 isoform, which is associated with increased cholesterol and risk for Alzheimer disease.

Lipoprotein(a)

Although Lp(a) particles typically carry only a relatively small fraction of total cholesterol, these lipoprotein particles are

particularly proatherogenic. Therefore Lp(a) is often measured in individuals with a normal lipid panel but with a history of premature cardiovascular disease and/or a strong family history of cardiovascular disease. Turbidimetric, nephelometric, RIA, and ELISA methods are all currently used for Lp(a) measurement. Most of these assays, except for ELISA, are based on the use of polyclonal antibodies. Sandwich-type ELISAs are usually based on the use of a combination of monoclonal and polyclonal antibodies to apo B and apo(a). The immunoreactivity of antibodies directed to the repeat kringle 4 type 2 domain in apo(a) makes some Lp(a) assays vary as a function of apo(a) size. These assays tend to underestimate apo(a) concentration in samples with apo(a) of smaller size than the apo(a) present in the assay calibrator and to overestimate the apo(a) concentration in samples with larger apo(a). Traditionally, Lp(a) concentrations have been reported in terms of total Lp(a) particle mass or, alternatively, in terms of Lp(a) protein. If the purpose is to provide Lp(a) values that are independent of apo(a) size, it is recommended that the Lp(a) assay use antibodies directed to an apo(a) domain other than kringle 4 type 2, or to the apo B-100 component of Lp(a). This would allow the values to be expressed in nanomoles per liter. At present, Lp(a) measurements are not well standardized, but a value of approximately 30 mg/dL of total Lp(a) particle mass has traditionally been used as a cutoff, above which increased concentrations of Lp(a) are associated with increased risk of CHD. Because Lp(a) values vary among ethnic groups, reference intervals ideally should be population based. For example, African Americans generally have significantly higher Lp(a) concentrations than do whites.

Sources of Variation and Bias in Test Measurement

Concentrations of various lipoproteins and their constituent parts—lipids and apolipoproteins—vary within individuals (see Chapter 4). Analytical variations for such measurements are relatively small—generally approximately 2% to 3% CVs for cholesterol—but some direct measurements of LDL-C and HDL-C do not meet National Cholesterol Education Program (NCEP) guidelines for analytical performance (see Table 23.10) on dyslipidemic samples. In contrast, physiological variations are much larger and contribute to the vast majority of overall variation in lipid and lipoprotein concentrations. For this reason, a patient's usual lipid or lipoprotein concentration is not reliably established from a single measurement, and the NCEP Laboratory Standardization Panel has made several specific recommendations for reducing the effects of preanalytical factors and specimen-processing procedures on lipid and lipoprotein testing.

ADVANCED LIPID AND LIPOPROTEIN TESTING FOR CORONARY HEART DISEASE RISK ASSESSMENT

Despite the strong association of plasma lipid concentrations with CHD risk, it has been long recognized that half of all myocardial infarctions occur among individuals with a normal lipid panel. Consequently, a wide variety of alternative or emerging biochemical markers and assays have been identified for better assessing cardiovascular risk. For example, LDL-P and HDL-P, which represent measurement of LDL and HDL particle number, respectively, can be measured by nuclear magnetic resonance (NMR) or by mass spectrometry and are positively related to CHD risk in case of LDL-P and negatively related for HDL-P. Similarly, the size of lipoproteins can be measured by a variety of physical techniques, and in the case of LDL, smaller-size LDL has been shown to be positively related to cardiovascular risk, particularly in patients with diabetes and metabolic syndrome. Several studies have shown that both lipoprotein particle number and size may be superior to LDL-C and HDL-C as cardiovascular biomarkers. Other alternative tests are based on other factors besides conventional lipids or apolipoproteins, such as oxidized lipids and the plasma concentration of the amino acid homocysteine. However, at this time, no overall consensus has been reached on how most of these advanced lipid and lipoprotein tests are best used, but as discussed later, hsCRP, a marker of inflammation, may be useful for deciding how to best treat patients at intermediate risk of CHD based on conventional risk markers.

High-Sensitivity C-Reactive Protein

CRP was first discovered in 1930 and was subsequently shown to be an acute phase reactant. It is routinely monitored as an indication of infection and autoimmune disease (see Chapter 18).

Numerous epidemiologic studies have demonstrated that increased serum CRP concentrations are positively associated with risk of future CHD events (Fig. 23.24). These findings support the hypothesis that (1) inflammation, (2) atherothrombosis, (3) diabetes, and (4) hypertension are interrelated and share common pathogenic mechanisms.

Prospective epidemiologic studies have shown that a single hsCRP measurement is a strong predictor of (1) myocardial infarction, (2) stroke, (3) peripheral vascular disease, and (4) sudden cardiac death, even in individuals with no history of heart disease. It has also been shown in prospective studies that hsCRP can be used to more accurately classify patients at intermediate risk of CHD. National guidelines for the measurement of hsCRP as a marker of CHD risk have been issued jointly by the American Heart Association and the Centers for Disease Control and Prevention (AHA/CDC). The Canadian Cardiovascular Society guidelines included hsCRP in the risk assessment of CHD and as a secondary target for treatment. In addition, the AHA/ACC guidelines, following the 10-year risk assessment, added new risk markers such as hsCRP, coronary artery calcium, or ankle-brachial indices in the risk assessment tools.

Several pharmacologic agents that have demonstrated cardioprotective ability, such as aspirin and statins, lower the risk of CHD that is associated with increased hsCRP

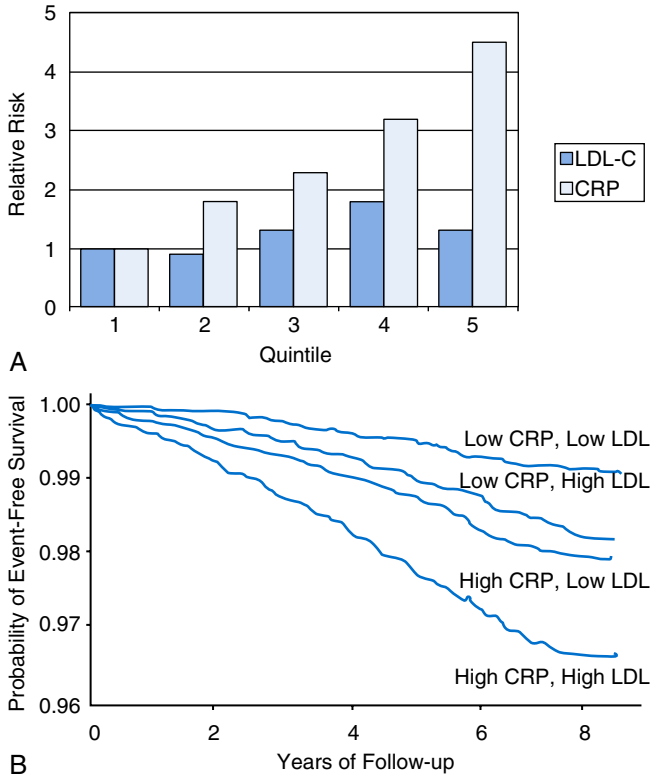


Fig. 23.24 (A) Relative risk of cardiovascular events with increase in low-density lipoprotein (LDL) and C-reactive protein *(CRP)*. (B) Probability of event-free survival decreases with increases in CRP and LDL levels. *LDL-C,* LDL cholesterol.

concentration. As a consequence, hsCRP is now often used in intermediate-risk individuals with low to normal lipids, but because of increased inflammation are at higher risk so that intervention with a statin can be implemented.

Methods have been developed to detect the low concentrations of CRP required for prediction of vascular risk, refer to as hsCRP. Of various techniques used to lower the limits of detection of CRP assays, the most successful approach has been to amplify the light-scattering properties of the antigen-antibody complex by covalently coupling latex particles to a specific antibody—a procedure that is easily automated on standard laboratory instrumentation. Many commercial hsCRP assays with low detection limits (<0.3 mg/L) are available, some with within-laboratory analytical imprecision of less than 10%.

Although it is an acute phase reactant, hsCRP exhibits a relatively low degree of intraindividual variability in clinically stable patients. It has been recommended to have two hsCRP measurements taken 2 or more weeks apart, with the average value used to estimate vascular risk. Furthermore, because hsCRP values are unaffected by food intake and exhibit almost no circadian variation, fasting specimens are not required.

Data from several large US and European cohorts indicate that the distribution of circulating hsCRP concentrations appears comparable among men and women not using postmenopausal hormone replacement therapy (HRT), with the 50th percentile for both genders seen at approximately 1.5 mg/L. Concentrations of hsCRP are higher in women who use oral HRT than in women who do not. Reference intervals of less than 1, 1 to 3, and greater than 3 mg/L, which correspond to approximate tertiles of the CRP distribution in healthy adults, are recommended for classification of individuals into low, moderate, and high cardiovascular risk groups in primary prevention settings.

REVIEW QUESTIONS

1. Which of the following formulas shows the correct calculation for indirectly measuring LDL-C (Friedewald formula)?
   a. LDL-C = HDL-C + (Triglyceride/5)
   b. LDL-C = Total cholesterol – (HDL-C) – (Triglyceride/5)
   c. LDL-C = Total cholesterol + HDL-C + (Triglyceride/5)
   d. LDL-C = HDL-C – (Triglyceride/5)
2. The protein component of a lipoprotein is referred to as:
   a. terpene.
   b. apolipoprotein.
   c. prostaglandin.
   d. phospholipid.
3. The lipoprotein that contains a carbohydrate-rich protein covalently bound to apo B-100 and a special protein motif called the "kringle" domain is:
   a. LDL.
   b. HDL.
   c. chylomicron.
   d. lipoprotein(a).
4. What lipoprotein transports mostly cholesteryl esters through the blood?
   a. LDL
   b. VLDL
   c. Chylomicrons
   d. Lipoprotein(a)
5. The enzyme that is critical for hydrolysis of triglycerides on chylomicrons for their conversion to chylomicron remnants is:
   a. cholesterol oxidase.
   b. glycerol kinase
   c. lipoprotein lipase.
   d. HMG-CoA reductase.
6. The rate-limiting enzyme in cholesterol biosynthesis that is inhibited by statin drugs is:
   a. cholesterol oxidase.
   b. glycerol kinase
   c. lipoprotein lipase.
   d. HMG-CoA reductase.

7. Which of the following lipid metabolic pathways has a role in transferring hepatically derived lipids, particularly triglyceride, to peripheral cells for energy metabolism?
   a. Exogenous pathway
   b. Endogenous pathway
   c. Intracellular cholesterol transport pathway
   d. Reverse cholesterol transport pathway
8. The formation of mixed micelles containing unesterified cholesterol, fatty acids, monoglycerides, phospholipids, and bile acids is referred to as:
   a. emulsification.
   b. denaturation.
   c. esterification.
   d. saturation.
9. With regard to lipids, a carboxyl (–COOH) with a long side chain (R) containing an even number of carbon atoms that is important in human nutrition and metabolism is a type of lipid referred to as a(n):
   a. acylglycerol.
   b. ester.
   c. fatty acid.
   d. terpene.
10. In the laboratory analysis of triglycerides, the initial step in all methods is:
   a. phosphorylation of glycerol catalyzed by glycerokinase.
   b. oxidation of cholesterol by cholesterol oxidase.
   c. hydrolysis of triglyceride by lipase.
   d. reduction of phenol and H_2O_2 by peroxidase.

SUGGESTED READINGS

Biggerstaff, K. D., & Wooten, J. S. (2004). Understanding lipoproteins as transporters of cholesterol and other lipids. *Advances in Physiology Education*, *28*, 105–106.

Davidson, M. H., Ballantyne, C. M., Jacobson, T. A., Bittner, V. A., Braun, L. T., Brown, A. S., et al. (2011). Clinical utility of inflammatory markers and advanced lipoprotein testing: advice from an expert panel of lipid specialists. *Journal of Clinical Lipidology*, *5*, 338–367.

Demacker, P. N. (2004). Laboratory-based assessment of plasma lipids and lipoproteins for the classification of familial hypercholesterolemic and hypertriglyceridemic states. *Seminars in Vascular Medicine*, *4*, 13–22.

Expert panel on integrated guidelines for cardiovascular health and risk reduction in children and adolescents: summary report. *Pediatrics*, *128*, (2011), S1–S44.

Havel, R. J., & Kane, J. P. Introduction: structure and metabolism of plasma lipoproteins. In C. R. Scriver, A. L. Beaudet, W. S. Sly, D. Valle (Eds.), *The metabolic and molecular bases of inherited diseases* (pp. 2705–2716). (8th ed). New York: McGraw-Hill.

Huxley, R., Lewington, S., & Clarke, R. (2002). Cholesterol, coronary heart disease and stroke: a review of published evidence from observational studies and randomized controlled trials. *Seminars in Vascular Medicine*, *2*, 315–323.

Mahley, R. W., Innerarity, T. L., Rall, S. C., Jr., & Weisgraber, K. H. (1984). Plasma lipoproteins: apolipoprotein structure and function. *Journal of Lipid Research*, *25*, 1277–1294.

Marcovina, S. M., Koschinsky, M. L., Albers, J. J., & Skarlatos, S. (2003). Report of the National Heart, Lung, and Blood Institute Workshop on Lipoprotein(a) and Cardiovascular Disease: recent advances and future directions. *Clinical Chemistry*, *49*, 1785–1796.

National Cholesterol Education Program (NCEP). (2002). Expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III). Third report of the National cholesterol education program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III): final report. *Circulation*, *106*, 3143–3421.

Recommendations on lipoprotein measurement: from the Working Group on Lipoprotein Measurement. National Cholesterol Education Program, NIH/NHLBI NIH Publication: Bethesda, MD.

Ridker, P., & Rifai, N. C-reactive protein and cardiovascular disease. MediEdition: St Laurent, Canada.

Ordovas, J. M. (2000). *Handbook of Lipoprotein Testing*. In N. Rifai, M. Dominiczak, & G. R. Warnick (Eds.) (2nd ed). Washington, DC: AACC Press.

Toth, P. P. Drug treatment of hyperlipidaemia: a guide to the rational use of lipid-lowering drugs. *Drugs*, 70, 1363–1379.

Vesper, H. W., Wilson, P. W., & Rifai, N. (2012). A message from the laboratory community to the National Cholesterol Education Program Adult Treatment Panel IV. *Clinical Chemistry*, *58*, 523–527.

24

Electrolytes and Blood Gases

Mark A. Cervinski, Mark D. Kellogg, Mitchell G. Scott

OBJECTIVES

1. Define the following terms:
   - Anion
   - Blood gas
   - Cation
   - Colligative properties
   - Electrolyte
   - Electrolyte exclusion effect
   - Osmolality
   - Oxygen dissociation curve
   - Oxygen saturation
   - pH
   - Partial pressure
   - Co-oximetry
   - Sweat chloride test
2. For each of the following electrolytes, state and describe the biochemistry, physiological functions, localization, clinical significance, method of analysis, and potential analytical interferences:
   a. Bicarbonate
   b. Chloride
   c. Potassium
   d. Sodium
3. Compare direct and indirect ion-selective electrode testing, including uses of and interferences in each method.
4. Discuss the measurement of sweat chloride, including reasons for performing, details of the phases of testing, and specimen requirements.
5. Describe the following methods of electrolyte/blood gas measurement, including which analytes are measured with each technique, the principle of the method, specimen requirements, and possible interferences:
   a. Coulometry-amperometry
   b. Osmometry
   c. Potentiometry
   d. Sweat testing
   e. Tonometry
6. List and describe the four colligative properties of a solution.
7. State the formula that demonstrates the relationship between total concentration of CO_2, bicarbonate, and hydrogen ion concentration.
8. State the Henderson-Hasselbalch equation and explain each of its components.
9. List the three essential properties of arterial blood required for adequate oxygen delivery to tissues.
10. For blood gas determinations, describe each of the following:
    a. Specimen requirements
    b. Preferred anticoagulant
    c. Collection technique
    d. Specimen transport conditions
11. List the analytical problems that arise in the analysis of pH, pO_2, and pCO_2 with inappropriate specimen collection or transport.
12. Diagram a generic blood gas analyzer; state and describe the methods or calculations used to measure pH, pO_2, and pCO_2, including noninvasive and continuous monitoring methods.

KEY WORDS AND DEFINITIONS

Anemic hypoxia Hypoxia (a reduced supply of oxygen to the tissues) resulting from a decrease in the amount of hemoglobin or number of erythrocytes in the blood.

Blood gases pCO_2 and pO_2 (partial pressures of carbon dioxide and oxygen), usually in whole blood.

Colligative properties Properties of solutions that depend on the number of particles in the solution; examples include (1) osmotic pressure, (2) boiling point elevation, (3) freezing point depression, and (4) vapor pressure lowering.

Cystic fibrosis (CF) Inherited disorder of a transmembrane conductance regulator protein (CFTR) that leads to chronic pancreatic and obstructive pulmonary disease.

Cystic fibrosis transmembrane conductance regulator (CFTR) A transmembrane protein produced by the *CFTR* gene.

Electrolytes Charged low-molecular-mass molecules present in plasma and cytosol, usually ions of (1) sodium, (2) potassium, (3) calcium, (4) magnesium, (5) chloride, (6) bicarbonate, (7) phosphate, (8) sulfate, and (9) lactate.

Electrolyte exclusion effect Electrolytes are excluded from the fraction of total plasma volume that is occupied by solids, which leads to underestimation of electrolyte concentration by some methods.

Hemoglobin (Hb) An oxygen-carrying, heme-containing protein abundant in red blood cells.

Henderson-Hasselbalch equation Equation that defines the relationship between pH, bicarbonate, and the partial pressure of dissolved carbon dioxide gas:

$$pH = pK + \log\left(\frac{cHCO_3^-}{\alpha + pCO_2}\right)$$

Ion-selective electrode (ISE) A type of special-purpose, potentiometric electrode consisting of a membrane selectively permeable to a single ionic species. The potential produced at the membrane–sample solution interface is proportional to the logarithm of the ionic activity or concentration.

Osmolal gap A difference between the observed and calculated osmolalities in serum analysis. The formula for calculated osmolality is 1.86 × Na (mmol/L) + glucose (mmol/L) + blood urea nitrogen (mmol/L) + 9; or 1.86 × Na (mmol/L) + glucose (mg/dL)/18 + urea (mg/dL)/2.8 + 9.

Osmometry Technique for measuring the concentration of dissolved solute particles in a solution.

Osmotic pressure The pressure required to stop osmosis through a semipermeable membrane between a solution and pure solvent.

Oximetry A technique used to determine the oxygen saturation of arterial blood.

Oxygen dissociation curve The sigmoidal curve obtained when SO_2 of blood is plotted against pO_2.

Oxygen saturation (SO_2) The fraction (percentage) of functional hemoglobin that is saturated with oxygen.

Oxyhemoglobin Hemoglobin that contains bound O_2.

P_{50} pO_2 for a given blood sample at which the hemoglobin of the blood is half saturated with O_2; P_{50} reflects the affinity of hemoglobin for O_2.

Partial pressure The substance (mole) fraction of gas times the total pressure; e.g., the partial pressure of oxygen, pO_2, is the fraction of oxygen gas times the barometric pressure.

pH The negative logarithm of hydrogen ion activity.

Pilocarpine iontophoresis Noninvasive method that uses electricity to force the drug pilocarpine into the skin for the purpose of inducing localized sweating at the site of iontophoresis.

Point-of-care testing (POCT) Clinical testing that occurs next to the patient, usually with a handheld device and an unprocessed specimen collected immediately before testing.

Sweat chloride The concentration of chloride in sweat; increased sweat chloride is characteristic of cystic fibrosis.

Water homeostasis The body process that maintains a balance of water intake and output.

Maintenance of **water homeostasis** is paramount to life for all organisms. In humans, the maintenance of water homeostasis in various body fluid compartments is primarily a function of the four major electrolytes: Na^+, K^+, Cl^-, and HCO_3^-. These electrolytes also have a role in acid-base balance and heart and muscle function, and serve as cofactors for enzymes. Abnormal electrolyte concentrations may be the cause or the consequence of a variety of medical disorders. Because of their physiologic and clinical interrelationships, this chapter discusses determination of (1) electrolytes, (2) osmolality, (3) sweat testing, (4) **blood gases** and **pH**, and (5) oxygen hemodynamics.

ELECTROLYTES

Electrolytes may be classified as *anions*, negatively charged ions that move toward an anode, or *cations*, positively charged ions that move toward a cathode. Important physiologic electrolytes include Na^+, K^+, Ca^{2+}, Mg^{2+}, Cl^-, HCO_3^-, $H_2PO_4^-$, HPO_4^{2-}, SO_4^{2-}, and some organic anions, such as lactate. At physiologic pH, proteins in serum are usually anions, and albumin accounts for most of the difference between the commonly measured cations (Na^+, K^+) and anions (Cl^-, HCO_3^-), also known as the anion gap. Hydrogen ion (H^+) concentration is routinely measured as pH, but its concentration is so low relative to other ions (10^{-9} vs. 10^{-3} mol/L) that its role as an electrolyte is negligible for clinical purposes. The major electrolytes (Na^+, K^+, Cl^-, HCO_3^-) occur primarily as free ions, whereas significant amounts (>40%) of Ca^{2+}, Mg^{2+}, and trace elements are bound by proteins, mainly albumin. Determination of body fluid concentrations of the four major electrolytes (Na^+, K^+, Cl^-, and HCO_3^-) is commonly referred to as an *electrolyte profile.*

SPECIMENS FOR ELECTROLYTE DETERMINATION

Serum and *plasma* are the usual specimens analyzed for electrolytes. Capillary blood, collected in microsample tubes or capillary tubes, or applied directly from a finger stick to point-of-care devices, are also common. Arterial or venous heparinized whole blood specimens obtained for blood gas and pH determinations may also be used with direct **ion-selective electrodes (ISEs)**. Differences in values between serum and plasma and between arterial and venous samples have been documented, but only the difference between serum and plasma K^+ can be considered clinically significant. Heparin, either lithium or ammonium salt, is required if plasma or whole blood is assayed. The use of plasma or whole blood has the advantage of shortening turnaround time as compared with serum, as these samples do not require sample clotting prior to analysis. Plasma or whole blood provides a

distinct advantage in determining K^+ concentrations, which are invariably higher in serum, depending on platelet count. Grossly lipemic blood can be a source of analytical error (see the section "Electrolyte Exclusion Effect"), and ultracentrifugation of lipemic serum or plasma is required prior to analysis via indirect-ISE methods. Hemolysis of red blood cells will cause erroneously high K^+ results in serum, plasma, and whole blood samples (pseudohyperkalemia); however, it is usually detected in serum and plasma due to its characteristic color. Hemolysis is, however, usually undetected when whole blood is analyzed. Pseudohyperkalemia may also develop in unhemolyzed specimens that are not promptly processed as K^+ will leak out of red blood cells when whole blood is stored at 4°C (for more details on preanalytical errors, refer to Chapter 3).

Urine collection for Na^+, K^+, or Cl^- assays should be done without the addition of preservatives. Body fluid aspirates, feces, or gastrointestinal (GI) fluid samples may also be submitted for electrolyte analysis.

SODIUM

Sodium is the major cation of extracellular fluid. Because it represents approximately 90% of the ≈154 mmol of inorganic cations per liter of plasma, Na^+ is responsible for almost one half the osmotic strength of plasma. It therefore plays a central role in maintaining the normal distribution of water and osmotic pressure in the extracellular fluid compartment (ECF).

Sodium is freely filtered by renal glomeruli. Seventy to 80% of the filtered Na^+ load is actively reabsorbed in the proximal tubules, with Cl^- and water passively following in an iso-osmotic and electrically neutral manner. Another 20% to 25% is reabsorbed in the loop of Henle, along with Cl^- and more water. In the distal tubules, interaction of the adrenal hormone aldosterone with the coupled Na^+-K^+ and Na^+-H^+ exchange systems results directly in the reabsorption of Na^+, and indirectly of Cl^-, from the remaining 5% to 10% of the filtered load. It is the regulation of this latter fraction of filtered Na^+ that primarily determines the amount of Na^+ excreted in the urine. These processes are discussed in detail in Chapter 35.

Specimens

Serum, plasma, and urine may be stored at 4°C or may be frozen. Erythrocytes contain only 1/10 of the Na^+ present in plasma, so hemolysis does not cause significant errors in serum or plasma Na^+ values. Lipemic samples should be ultracentrifuged, and the infranatant analyzed unless a direct ISE is used (see the section "Electrolyte Exclusion Effect").

Fecal and GI fluid specimens require preparation before assay. Because significant electrolyte loss in feces only occurs in diarrhea, only liquid stool samples may be justified for analysis. Immediately after collection, liquid stool specimens should be clarified of particulate matter by filtration through gauze or filter paper and by centrifugation. If not analyzed immediately, fecal and GI fluids should be stored frozen to prevent microbial growth.

Reference Intervals

A typical reference interval for *serum* or plasma Na^+ is 135 to 145 mmol/L. The interval for premature newborns at 48 hours is 128 to 148 mmol/L, and the value for umbilical cord blood from full-term newborns is ≈127 mmol/L.

Urinary sodium excretion varies with diet, but for an adult male consuming 7 to 14 g of NaCl per day, an interval of 120 to 240 mmol/L is typical. A large diurnal variation in Na^+ excretion has been noted, with the rate of Na^+ excretion during the night being only 20% of the peak rate during the day. The Na^+ concentration of cerebrospinal fluid is 136 to 150 mmol/L. Mean fecal Na^+ excretion is less than 10 mmol/L.

POTASSIUM

Potassium is the major intracellular cation with an average concentration in tissue cells of ~150 mmol/L, and in erythrocytes, the concentration is ~105 mmol/L. High intracellular concentrations are maintained by the Na^+, K^+ adenosine triphosphate (ATP)ase pump, which continually transports K^+ into the cell against a concentration gradient. Diffusion of K^+ out of the cell into the ECF and plasma occurs whenever pump activity is decreased because of (1) depletion of metabolic substrates such as glucose; (2) competition for ATP between the pump and other energy-consuming activities of the cell; or (3) slowing of cellular metabolism (as occurs with refrigeration). The importance of these considerations on sample integrity for analysis of K^+ is discussed later.

The body requirement for K^+ is satisfied by an average dietary intake of 2.4 to 4.4 g/d (60 to 120 mmol/d). Potassium absorbed from the GI tract is rapidly distributed, with a small amount taken up by cells, and most excreted by the kidneys. Potassium filtered through the glomeruli is almost completely reabsorbed in the proximal tubules and is then secreted in the distal tubules in exchange for Na^+ under the influence of aldosterone. Aldosterone enhances K^+ secretion and Na^+ reabsorption in the distal tubules by a Na^+-K^+ exchange mechanism. The kidneys respond to K^+ loading with an increase in K^+ output, so that urine collected during or after a period of high K^+ intake may have K^+ concentrations as high as 100 mmol/L. In contrast, the tubular response to conserve K^+ is very slow in the initial stages of depletion. Unlike the prompt response to conserve Na^+ in deficit states, it can take up to 1 week for the tubules to reduce K^+ excretion to 5 to 10 mmol/d from the typical 50 to 100 mmol/d.

Factors that regulate distal tubular secretion of K^+ include intake of Na^+ and K^+, mineralocorticoid concentration, and acid-base balance. Because renal conservation mechanisms are slow to respond, K^+ depletion can be an early consequence of restricted K^+ intake or loss of K^+ by extrarenal routes such as diarrhea. A diminished glomerular filtration rate is typical of renal failure, and the consequent decrease in distal tubular flow rate is an important factor in the hyperkalemia associated with renal failure.

Specimens

Comments made earlier on specimens for Na^+ analysis are generally applicable to those for K^+ analysis, with some caveats. Potassium concentrations in plasma and whole blood are 0.1 to 0.7 mmol/L lower than those in serum, and most reference intervals for serum K^+ are 0.2 to 0.4 mmol/L higher than those for plasma K^+. The extent of this difference depends on the platelet count, because additional K^+ in serum is primarily a result of platelet rupture during clotting. This variability in the amount of additional K^+ in serum makes plasma the specimen of choice.

Specimens for determining K^+ concentrations in serum or plasma must be collected by methods that minimize hemolysis, because the release of K^+ from as few as 0.5% of erythrocytes can increase K^+ values by 0.5 mmol/L. An increase in K^+ of 0.6% has been estimated for every 10 mg/dL of plasma **hemoglobin** caused by hemolysis. Thus slight hemolysis (Hb ≈ 50 mg/dL) can be expected to raise K^+ values by ≈3%, marked hemolysis (Hb ≈200 mg/dL) by 12%, and gross hemolysis (Hb > 500 mg/dL) by as much as 30%. The use of correction factors based on a hemolysis index has been suggested for estimating K^+ in hemolyzed samples, but their use has been questioned. Regardless, it is imperative that if the laboratory chooses to report K^+ concentrations on hemolyzed samples that the presence of hemolysis is noted with a comment that results are falsely elevated, whether an estimate of the extent of elevation is provided or not. It is important to remember that if K^+ concentrations are determined by ISE on whole blood specimens, hemolysis is not readily visible. When hemolysis is suspected, as in the case of an unexpected increase in K^+, a portion of the specimen should be centrifuged and the plasma visually inspected.

Clinically significant preanalytical errors can occur for K^+ determinations if blood samples are not processed quickly. Maintenance of the intracellular-extracellular K^+ gradient depends on the activity of the energy-dependent Na^+-K^+-ATPase. If a whole blood specimen is maintained at 4°C versus 25°C before separation, glycolysis is inhibited and the energy-dependent Na^+-K^+-ATPase cannot maintain the Na^+/K^+ gradient. An increase in plasma K^+ will occur as a result of K^+ leakage from erythrocytes and other cells. The increase of K^+ in serum is on the order of 0.2 mmol/L by 1.5 hours at 25°C, and is as high as 2 mmol/L after 4 hours at 4°C.

A falsely decreased K^+ value can be observed if an unseparated sample is stored at 37°C, because glycolysis occurs and K^+ shifts intracellularly. Even at room temperature, extreme leukocytosis can initially cause falsely decreased K^+ concentrations. The extent of this decrease depends on leukocyte count, temperature, and glucose concentrations, but has been reported to be as much as 0.7 mmol/L at 37°C. In addition to causing pseudohypokalemia, samples from leukemic patients with very high white blood cell counts (>300 × 10^9 cells/L) can result in a pseudohyperkalemia due to WBC rupture. Taken together, the recommendation for reliable K^+ determinations is to collect blood with heparin and to maintain it near 25°C, and then to separate the plasma within minutes by high-speed centrifugation. In practical terms, separation within 1 hour when samples are maintained at room temperature is unlikely to introduce great error.

Finally, skeletal muscle activity causes K^+ efflux from muscle cells into plasma and can cause a marked elevation in plasma K^+ values. A common example occurs when an upper arm tourniquet is not released before beginning to draw blood after a patient has repeatedly clenched their fist. This combination of tourniquet induced venostasis and muscular activity can artificially raise plasma K^+ values as much as 2 mmol/L.

Reference Intervals

Reported reference intervals for the serum of adults vary from 3.5 to 5.1 mmol/L and from 3.7 to 5.9 for newborns. For plasma, a frequently cited adult interval is 3.3 to 4.9 mmol/L. Urinary excretion of K^+ varies with dietary intake, but a typical range observed is 42 to 86 mmol/L for males and 33 to 70 mmol/L for females. In severe diarrhea, GI loss may be as much as 60 mmol/L.

METHODS FOR THE DETERMINATION OF SODIUM AND POTASSIUM

Sodium and potassium may be determined by (1) atomic absorption spectrophotometry (AAS), (2) flame ionization spectrophotometry (FES), (3) electrochemically with an ion-selective electrode, or (4) spectrophotometrically. Although AAS, FES, or spectrophotometric methods have been used for Na^+ and K^+ analyses in the past, most laboratories now use ISE methods. For example, in 2017, of the laboratories reporting proficiency data for Na^+ and K^+ to the College of American Pathologists (CAP), 98% were using ISE methods.

Ion-Selective Electrodes

ISEs measure Na^+ and K^+ using the principle of potentiometry. Potentiometry is the determination of change in electromotive force (*E,* potential) in a circuit between a measurement electrode (the ISE) and a reference electrode, as the selected ion interacts with the membrane of the ISE. ISE composition varies with the analyte measured; Na^+ ISEs are fitted with glass membranes and K^+ electrodes with liquid ion-exchange membranes incorporating valinomycin. In electrolyte analyzers and automated instruments, the system is calibrated with solutions containing defined amounts of Na^+ and K^+. The potentials of the electrodes in the presence of the calibrators are determined, and the $\Delta E/\Delta$ log concentration responses are stored in microprocessor memory for calculating unknown concentration when *E* of the unknown is measured.

Frequent calibration of the electrodes' response or slope factors, initiated by the user or by microprocessor-controlled uptake of calibrator, is typical of most current ISE systems. Some of the available ISE systems also check the potential

of one of the calibrators or of a third calibrator, immediately prior to the measurement of each patient specimen, to compensate for electrode drift and/or noise.

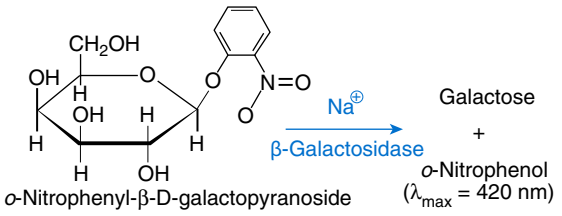
Two types of ISE methods, indirect ISE and direct ISE, are in use and must be distinguished. With *indirect ISE methods,* the sample is first mixed with a large volume of diluent of low ionic strength prior to introduction into the measurement chamber. Indirect ISE is most commonly used on automated, high-throughput clinical chemistry systems. Indirect methods were developed early in the history of ISE technology, when dilution was necessary to present a small sample in a volume large enough to adequately cover a large electrode and to minimize the concentration of protein at the electrode surface. With *direct ISE methods*, the sample is presented to the electrodes without dilution. Direct ISE methods became possible with electrode miniaturization and are available on a few automated laboratory instruments. Direct ISE electrodes are primarily used in blood gas analyzers and point-of-care devices where whole blood is directly presented to the electrodes.

Errors observed in the use of ISEs fall into three categories. First are errors caused by lack of selectivity. For instance, many Cl^- electrodes lack selectivity against other halide anions such as I^- or SCN^- from the metabolism of sodium nitroprusside. Second are errors introduced by repeated protein coating of the ISE membranes, or by contamination of the membrane or salt bridge by ions that compete or react with the selected ion and thus alter the electrode response. Such errors in ISE measurements necessitate periodic changes of the membrane as part of routine maintenance. Finally, the **electrolyte exclusion effect**, which applies only to indirect methods and is caused by the solvent-displacing effect of lipid and protein in the sample, results in falsely decreased values.

Spectrophotometric Methods

Spectrophotometric methods are based on Na^+ or K^+ specific enzyme activation. However, the higher cost of these methods and the fact that few problems exist with ISE methods have resulted in "niche" use of spectrophotometric analysis, primarily with smaller instruments used in physicians' offices, and more recently "isolation" laboratories for patients with emerging infectious diseases.

Kinetic spectrophotometric assays for Na^+ are based on activation of the enzyme β-galactosidase by Na^+ to hydrolyze *o*-nitrophenyl-β-D-galactopyranoside. The rate of production of *o*-nitrophenol (the chromophore) is measured at 420 nm.



Spectrophotometric K^+ assays rely upon the K^+ activation of pyruvate kinase for the conversion of phosphoenolpyruvate to pyruvate, which is then reduced to lactate by lactate dehydrogenase, in the process of oxidizing NADH to NAD^+.

ELECTROLYTE EXCLUSION EFFECT

The major plasma electrolytes, Na^+, K^+, Cl^-, and HCO_3^-, have minimal protein binding and are carried in the water phase of plasma. The composition of plasma is approximately 93% water and 7% solids such as proteins and lipids. When the fixed volume of plasma (e.g., 10 μL) is pipetted for dilution into a buffer of low ionic strength before indirect ISE analysis, only 9.3 μL of *plasma water* that contains the electrolytes is added to the diluent. However, when we report plasma electrolyte concentrations, we report the values in mmol/L of total plasma volume, not in plasma water volume. Thus in the absence of increased plasma protein or lipid concentrations, a reported Na^+ concentration of 140 mmol/L of plasma volume would equal a Na^+ concentration in plasma water of 150 mmol/L (140 × [100/93]). This negative "error" or bias in plasma electrolyte analysis has been recognized for many years.

All electrolyte reference intervals are based on the assumption that plasma is reliably ~93% water and reflect concentrations in total plasma volume and not in water volume. Even though it is the electrolyte concentration in plasma water that is physiologically relevant (the Na^+ concentration of normal saline is 150 mmol/L), it was assumed that the volume fraction of water in plasma is sufficiently constant that this difference could be ignored. This electrolyte exclusion effect, however, becomes problematic when pathophysiologic conditions are present that alter the plasma water volume, such as hyperlipidemia or hyperproteinemia. In these settings, falsely low electrolyte values are obtained whenever samples are diluted before analysis.

In certain settings, such as ketoacidosis with severe hyperlipidemia or multiple myeloma with severe hyperproteinemia, the negative exclusion effect may be so large that laboratory results lead clinicians to believe that electrolyte concentrations are normal or low, when in fact the concentration in the water phase may be high or normal, respectively, leading to pseudohyponatremia (Fig. 24.1).

In severe hypoproteinemia, the effect works in reverse, resulting in falsely high (2% to 4%) Na^+ or K^+ values. As with the indirect ISE, the original method for measuring plasma electrolytes, flame photometry, also required dilution of the sample prior to analysis, and as such is also subject to pseudohypo- and hypernatremia.

Direct ISE methods do not require sample dilution, and as there is no dilution, ion activity is directly proportional to the concentration in the water phase, not the concentration in the total plasma volume. Therefore direct ISE methods are free of electrolyte exclusion effects. To make results from direct ISEs equivalent to those from indirect ISEs (and flame photometry), direct ISE methods operate in what is commonly referred to as the "flame mode." In this mode, the directly measured concentration in plasma water is multiplied by the average water volume fraction of plasma (0.93). Although the actual water volume fraction of plasma may vary widely, as long as the activity of the specific ion is constant, the concentration of the ion in the water phase becomes *independent* of

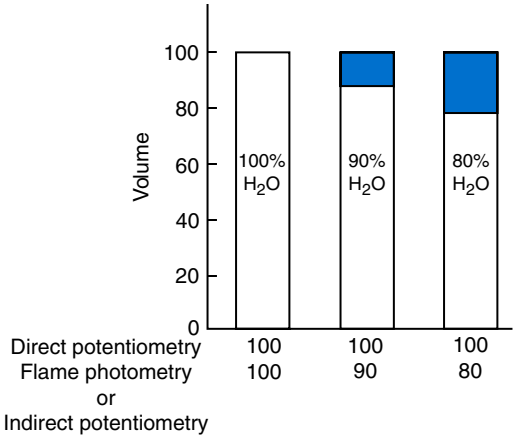


Fig. 24.1 Predicted influence of water content on sodium measurements for a 100 mmol/L NaCl solution by direct ion-selective electrode (ISE) versus flame emission photometry or indirect ISE. *Blue areas* represent the volumes of nonaqueous components that are suspended in the solution. In plasma these consist of lipids or proteins. (Modified from Apple, F. S., Koch, D. D., Graves, S., & Ladenson, J. H. [1982]. Relationship between direct-potentiometric and flame-photometric measurement of sodium in blood. *Clinical Chemistry, 28*, 1931–1935.)

the relative proportions of water and total solids if the ion is not bound by proteins. Most clinical chemists and physicians have reached the conclusion that direct ISE methods for electrolyte analysis are the methods of choice. However, results from direct ISE methods will continue to be converted to total plasma volume concentrations by use of the "flame mode," as greater than 80% of laboratories continue to use indirect ISE methods.

Chloride

Chloride is the major extracellular anion and, similar to Na^+, is significantly involved in the maintenance of water distribution, osmotic pressure, and anion-cation balance in the ECF. In contrast to its high ECF concentrations (≈103 mmol/L), the concentration of Cl^- in the intracellular fluid of erythrocytes is 45 to 54 mmol/L, and in the intracellular fluid of most other tissue cells, it is only ≈1 mmol/L. In gastric and intestinal secretions, Cl^- is the most abundant anion.

Chloride ions are almost completely absorbed from the intestinal tract. In the kidneys, they are filtered from plasma at the glomeruli and are passively reabsorbed, along with Na^+, in the proximal tubules. In the thick ascending limb of the loop of Henle, Cl^- is actively reabsorbed by the chloride pump, which promotes passive reabsorption of Na^+. Loop diuretics such as furosemide inhibit the chloride pump. Chloride concentrations are not homeostatically controlled, and passively reflect the concentration of the major ions sodium and bicarbonate. In addition, chloride concentrations fall, to maintain the electroneutrality of body fluids when pathological levels of other anions (e.g., ketoacids and lactate) are present.

Specimens

Chloride is most often measured in serum or plasma, urine, and sweat. Fecal determination of Cl^- may also be useful for the diagnosis of congenital hypochloremic alkalosis with hyperchloridorrhea (increased Cl^- in stool). Chloride is stable in serum and plasma, and hemolysis does not significantly alter serum or plasma Cl^- concentration because the erythrocyte concentration of Cl^- is approximately half of that in plasma.

Reference Intervals

Reported reference intervals for Cl^- in the serum or plasma vary from 98 to 107 mmol/L to 100 to 108 mmol/L. Urinary excretion of Cl^- varies with dietary intake, but an interval of 110 to 250 mmol/24 hours is typical.

METHODS FOR DETERMINATION OF CHLORIDE IN BODY FLUIDS

Chloride is determined largely by ISE, with some laboratories performing coulometric-amperometric titration for sweat chloride testing.

Ion-Selective Electrode Methods

Chloride ISEs are composed of solvent polymeric membranes incorporating quaternary ammonium salt anion exchangers. Although they are by far the most common methods for measuring Cl^- in clinical laboratories, these electrodes may suffer from membrane instability and lot-to-lot inconsistency in terms of selectivity to other anions. Anions that tend to be problematic include other halides and organic anions (e.g., SCN^-), which can be particularly problematic because of their ability to solubilize in the polymeric organic membrane of these electrodes.

SWEAT CHLORIDE

Sweat Collection

Sweat collection is a two-step procedure involving (1) localized sweating stimulated by iontophoresis of the cholinergic drug pilocarpine nitrate into an area of the skin, and (2) collection of sweat from the stimulated area into gauze or plastic microbore tubing. Iontophoresis uses a small electric current to deliver pilocarpine into the sweat glands from the positive electrode, and an electrolyte solution at the negative electrode completes the circuit. There are two methods for **pilocarpine iontophoresis** and sweat collection: using filter paper or gauze, or using plastic microbore tubing. The filter paper/gauze method, sometimes referred to as the Gibson-Cooke method, is the original method. Iontophoresis is achieved with electrodes over reagent-soaked gauze, followed by collection of sweat from the collection site into clean gauze. Alternatively, electrodes attached to pilocarpine-containing agar disks can be used for stimulation, with sweat collected into plastic capillary tubing. The inner flexor surface of the forearm is the preferred site for sweat testing. Sweat stimulation can be performed on the inner thigh; however, this site yields lower sweat volume.

Sufficient Sample Collection

The CFF requires 75 mg or 15 μL of sweat, if collected by Gibson-Cooke or Macroduct methods, respectively. CFF accredited centers are required to monitor the rate of their

"quantity not sufficient" (QNS) specimens and must review sample collection steps, should the QNS rate exceed the mandated acceptable limits. On average, the percentage of insufficient samples should not exceed 5% for patients older than 3 months of age, and 10% for patients less than 3 months of age. Insufficient sweat samples can be the result of several factors such as age, race, and skin condition. Collecting an adequate amount of sweat can be more challenging in patients younger than 1 month of age. For this reason, it is recommended that sweat testing in asymptomatic individuals be performed when the infant is at least 2 weeks of age, is more than 36 weeks' gestation at birth, and weighs more than 2 kg. To increase the chances of an adequate sample volume, two sites can be stimulated at one patient encounter. If the laboratory collects sweat from two sites (bilateral testing), the collection is considered QNS only when both sites are inadequate.

Quality of Sweat Collection and Analysis

Sweat collection is a manual process with many opportunities for error. Most errors result in QNS collections (QNS by CFF guidelines). Quality of sweat collection can be maintained by having a sufficient test volume to maintain competency, and limiting the testing personnel to a few well-trained individuals. The CFF recommends following CLSI-C34-A2 guidelines for sweat collection and analysis.

The Centers for Disease Control and Prevention does not list sweat as a potentially infectious material unless it is visibly contaminated with blood. However, laboratory personnel should practice the same standard precautions they would use with any other body fluid. All equipment used in iontophoresis should be disinfected between patients in accordance with procedures consistent with the institution's infection control policies. Disinfectants must not contain bleach, which could contaminate the sweat sample.

It is important to have the patient physically prepared for the sweat collection. The patient should be physiologically and nutritionally stable, well hydrated, free of acute illness, and not receiving mineralocorticoids. The patient should avoid using lotions on the skin prior to sweat collection. Possible adverse events during sweat collection include redness and itching at the stimulation site. Pilocarpine urticaria associated with sweat testing is rare but has been reported. Burns to the patient's skin after iontophoresis are extremely rare but can occur.

Reference Intervals for Sweat Chloride

For all individuals, including newborns with a positive newborn screen for CF, a sweat chloride concentration of less than 30 mmol/L indicates that CF is unlikely. This is a recent change to the previous practice of reporting separate reference intervals for infants up to and including 6 months of age. The following reference intervals are recommended for all patients regardless of age.

≤29 mmol/L: CF unlikely
30 to 59 mmol/L: intermediate
≥60 mmol/L: indicative of CF

The functional upper limit for sweat chloride is 160 mmol/L. A sweat chloride concentration greater than 160 mmol/L represents specimen contamination or analytical error. A sweat chloride concentration less than 30 mmol/L alone is insufficient to rule out the diagnosis; the value should be interpreted in regard to the clinical presentation and with the knowledge that "normal" concentrations have been associated with CF. Some mutations of the CF gene are associated with intermediate or normal sweat chloride concentrations. For example, according to the Cystic Fibrosis Foundation (CFF) registry, 3.5% of CF patients had sweat chloride concentrations less than 60 mmol/L, and 1.2% had concentrations less than 40 mmol/L.Clinical practice regarding follow-up of intermediate sweat chloride results varies widely. The CFF recommends a repeat sweat chloride test within 2 months of the intermediate result.

Reasons for Falsely Elevated Chloride in Sweat

There are numerous physiological conditions in addition to CF that may increase sweat chloride concentrations (see the CLSI document C34-A3). However, the most common cause of falsely elevated sweat chloride is evaporation. The 2007 CFF guidelines for sweat testing were established with the goal of reducing preanalytical variations in sweat testing. Testing procedures should minimize the opportunity for evaporation and contamination of the sweat sample. Sweat samples should be analyzed soon after collection on the same day. However, the laboratory can store the sweat samples if necessary. If sweat is collected into gauze, it should be reweighed promptly and can be stored with a tightly fitting lid up to 72 hours at 4°C. If sweat is collected in microbore tubing coils, it should be transferred to a 0.2 mL microcentrifuge tube with a tight-fitting cap and may be stored for up to 72 hours. The sweat should not be stored or transported in the microbore tubing.

Quantitative Analysis of Sweat Chloride

Coulometric-Amperometric Titration

Sweat chloride concentration is determined by coulometric titration as this method can measure chloride concentrations as low as 10 mmol/L, and is the most precise method for measuring Cl^- over the entire range of concentrations found in sweat. ISEs are not used to measure sweat chloride, as they lack precision at the chloride concentrations found in normal subjects.

Coulometric-amperometric determinations of Cl^- (also called the *Cotlove chloridometer technique*) depend on the generation of Ag^+ from a silver electrode at a constant rate, and on the reaction of Ag^+ with Cl^- in the sample to form insoluble silver chloride (AgCl):

$$Ag^+ + Cl^- \rightarrow AgCl$$

After the stoichiometric point is reached, excess Ag^+ in the mixture triggers shutdown of the Ag^+ generation system. A timing device records elapsed time between the start and stop of Ag^+ generation. Because the time interval is proportional to the amount of Cl^- in the sample, the concentration of

Cl^- can be determined by comparing the time elapsed against a calibration curve.

This method is subject to interferences by other halide ions, by CN^- and SCN^- ions, by sulfhydryl groups, and by heavy metal contamination. Maintenance of the systems is crucial for proper operation.

Sweat chloride ion concentration is used to confirm the diagnosis of **cystic fibrosis (CF)**. CF is caused by dysfunction of the **cystic fibrosis transmembrane conductance regulator (CFTR)**. Functional abnormalities in CFTR result in impaired electrolyte transport in secretory and absorptive epithelia, including the reabsorptive duct of the sweat gland. This leads to high salt loss via the sweat gland in patients with CF. In 1989, the *CFTR* gene was identified as the genetic basis of CF and over 1900 mutations in the 168 kb gene are known today. However, genetic testing is not always definitive, as the majority of mutations are not associated with clinical symptoms. Quantitative analysis of chloride in sweat remains the gold standard for CF diagnosis.

Sweat Conductivity Screening

Sweat conductivity is not diagnostic for CF but might be used as a screening test for CF. Compared to sweat chloride analysis, sweat conductivity is easier to perform and requires a smaller sample (minimum of 6 µL). Sweat conductivity is measured by comparing electrical conductance of sweat to that of NaCl standard solutions. Sweat conductivity is expressed as mmol/L (equivalent NaCl), a unit that represents the molar concentration of a solution of NaCl that has the same conductivity as the sweat sample at the same temperature, not the actual sodium or chloride concentration in the sweat sample. The CFF has approved the Macroduct Sweat-Chek (Wescor) for screening at clinical sites, such as community hospitals, using sweat conductivity ≥50 mmol/L as a cutoff for referral to an accredited CF care center.

BICARBONATE (TOTAL CARBON DIOXIDE)

Total carbon dioxide is used here to describe the quantity that is measured most often in automated clinical chemistry analyzers by acidification of a serum or plasma sample and measurement of carbon dioxide released by the process, or by alkalinization and measurement of total bicarbonate. Under certain conditions of collection and specimen handling, total carbon dioxide values determined in this manner will be almost identical to values for the calculated concentration of total carbon dioxide obtained in blood gas analysis (see the later section in this chapter on blood gas methods).

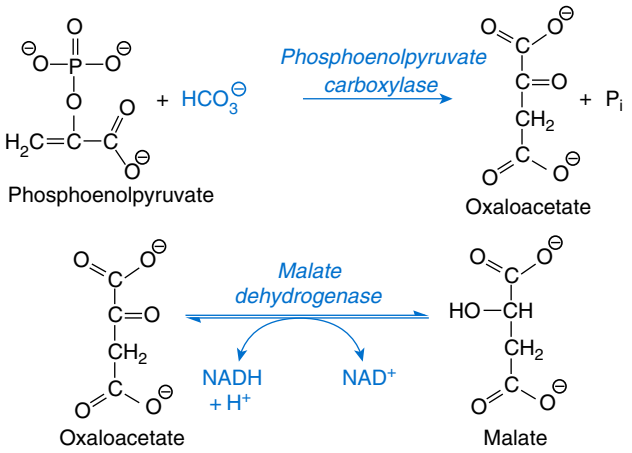
Specimens

The same sample types used for Na^+ or K^+ may be assayed. Given a specimen in a vacuum-draw tube, the concentration of total CO_2 is most accurately determined when the assay is done as promptly as possible after collection and centrifugation of the blood in a fully filled, unopened tube. Ambient air contains far less CO_2 than does plasma, and gaseous dissolved CO_2 will escape from the specimen into the air, with a consequent decrease in the CO_2 value of up to 2 to 3 mmol/L in the course of 1 hour. In practical terms, the logistics of high-volume processing and automated analysis of specimens almost ensures that most CO_2 measurements are done on specimens that have lost some dissolved gaseous CO_2, simply because preservation of anaerobic conditions is not practical between the time plasma is placed on an instrument and the time it is sampled. Thus the term *bicarbonate* may be preferable to *total CO_2*.

Methods for Determination of Serum or Plasma Total Carbon Dioxide

Methods for total CO_2 measurement with modern automated instruments may be electrode based or enzymatic. In indirect electrode-based methods, the plasma sample is acidified to convert the various forms of CO_2 present in the sample to gaseous CO_2. The amount of gaseous CO_2 released after acidification is determined by a pCO_2 electrode in the reaction chamber of the CO_2 module. *Direct ISE methods* for total CO_2 are no longer common on automated analyzers due to problems with specificity.

In *enzymatic methods* for CO_2, the specimen is first alkalinized to convert all CO_2 and carbonic acid to HCO_3^-. The enzymatic reactions are as follows:



The decrease in absorbance at 340 nm, as NADH is oxidized to NAD^+, is proportional to the total CO_2 content. The enzymatic method is more common than the electrode-based method, with approximately 80% of laboratories utilizing an enzymatic method.

Reference Intervals

The reference interval for total carbon dioxide in adults is 22 to 28 mmol/L but can be method and instrument dependent. The reference interval for CO_2 also varies with age, with pediatric patients exhibiting lower values due to a higher respiratory rate.

POINTS TO REMEMBER

- Sodium, potassium, and chloride concentrations in biological fluids are predominantly quantified using ion-selective electrodes.
- Sweat chloride concentrations are quantified utilizing coulometric-amperometric titration, and elevated concentrations are used to diagnose cystic fibrosis.
- Indirect ion-selective electrodes used for measuring serum/plasma sodium concentrations are prone to the electrolyte exclusion effect in samples containing high concentrations of lipids and/or proteins.
- Total serum/plasma CO_2 concentrations will decrease when samples are uncapped and exposed to room air.

PRINCIPLES OF OSMOTIC PRESSURE AND OSMOSIS

Osmometry is a technique for measuring the concentration of solutes that contribute to the osmotic pressure of a solution. Osmotic pressure governs the movement of solvent (water in biological systems) across semipermeable membranes that separate two solutions. Examples of biologically important selective membranes are those enclosing the glomeruli and capillary vessels that are permeable to water and to essentially all *small* molecules and ions, but not to large protein molecules. Differences in the concentrations of osmotically active molecules that cannot cross a membrane cause those molecules that can cross the membrane (such as water in the descending loop of Henle) to move to equalize the concentrations on both sides of the membrane. The force driving this movement of solute and permeable ions is known as **osmotic pressure**. As a simple example, consider an aqueous solution of sucrose placed within a sac made up of a membrane permeable only to water, with an open vertical glass tube attached to the sac. If the sac is placed into a beaker of distilled water, water will move from the beaker across the membrane into the sucrose solution. The pressure of this solvent movement will cause the sucrose solution to rise up the tube. At equilibrium, the gravitational pressure of the column of solution in the tube equals the osmotic pressure and prevents further net movement of water from the beaker. The height of the rise of the sucrose solution in the glass tube is a measure of the *osmotic pressure* of the sucrose solution. This is the pressure that would have to be exerted on the sucrose side of the membrane to prevent the flow of water across the membrane.

If the sucrose solution in the above example was replaced with a solution of NaCl at the same molarity, the solution in the glass tube would reach equilibrium at a point almost twice as high as that observed with sucrose. This is because NaCl dissociates into two ions per molecule and would thus have twice as many osmotically active solutes (osmoles) for the same molar concentration as the sucrose solution. In reality, the number of active particles in a solution of NaCl is less than this (~0.93 for NaCl), as explained later in this chapter. The total number of individual solutes present in a solution per given mass of solvent, regardless of their chemical nature (i.e., nonelectrolyte, ion, or colloid), determines the total osmotic pressure of the solution. In blood plasma, for example, nonelectrolytes such as glucose and urea and, to a much lesser extent, proteins contribute to the osmotic pressure.

Colligative Properties

In addition to the osmotic pressure being raised when a solute is dissolved in a solvent, the *vapor pressure* of the solution is *lowered* below that of the pure solvent. As a result of the change in vapor pressure, the *boiling point* of the solution is *raised* above and the *freezing point* of the solution is *lowered* below that of the pure solvent.

These four properties of solutions—(1) increased osmotic pressure, (2) lowered vapor pressure, (3) increased boiling point, and (4) decreased freezing point—are called **colligative properties**. All are directly related to the total number of dissolved solutes per mass of solvent. For example, a 1-molal (mol/kg) aqueous solution boils at a temperature 0.52°C higher than and freezes at a temperature 1.858°C lower than pure water. The vapor pressure of this solution is 0.3 mm Hg lower than the vapor pressure of pure water, which is 23.8 mm Hg at 25°C. The term *osmolality* expresses concentrations relative to *mass* of the solvent (1 osmolal solution is defined to contain 1 Osmol/kg H_2O), whereas the term *osmolarity* expresses concentrations per volume of solution (1 osmolar solution is defined to contain 1 Osmol/L). Osmolality (Osmol/kg H_2O) is a thermodynamically more exact expression because solution concentrations expressed on a weight basis are temperature independent, whereas those based on volume vary with temperature. Although the term *osmolarity* is often used in the medical literature, *osmolality* is what the clinical laboratory measures.

An electrolyte in solution dissociates into two (in the case of NaCl) or three (in the case of $CaCl_2$) dissolved ions; therefore the colligative effects of such solutions are multiplied by the number of dissociated ions formed per molecule. However, because of incomplete electrolyte dissociation and associations between solute and solvent molecules, many solutions do not behave as the ideal case, and a 1-molal solution may give an osmotic pressure lower than theoretically expected. The osmotic activity coefficient is a factor used to correct for deviation from the "ideal" behavior of the system:

$$\text{Osmolality} = \frac{\text{Osmol}}{\text{Kg}H_2O} = \Phi\, nC$$

where Φ = osmotic coefficient, n = number of solutes into which each molecule in the solution potentially dissociates, and C = molality in mol/kg H_2O.

Glucose has an osmotic coefficient of 1.00, whereas the Φ for sodium chloride is 0.93 at the concentrations found in serum—thus the derivation of 1.86 × Na^+ (mmol) in the formula to calculate plasma osmolality (NaCl potentially contributes two osmotically active ions times 0.93 = 1.86). Ethanol has an osmotic coefficient of 0.83. The total osmolality or osmotic pressure of a solution is equal to the sum of the osmotic pressures or osmolalities of all solute species present. The electrolytes Na^+, Cl^-, and HCO_3^-, which are present in relatively high concentrations, make the greatest contributions to serum osmolality. Nonelectrolytes such as glucose and urea, which are present normally at lower molal concentrations, contribute less, and serum proteins contribute less than 0.5% of the total serum osmolality because even the most abundant protein is present at millimolar concentrations.

Determination of Plasma and Urine Osmolality

Determination of plasma and urine osmolality can be useful in the assessment of electrolyte and acid-base disorders. Comparison of plasma and urine osmolalities can determine the status of water regulation by the kidneys in settings of severe electrolyte disturbances, as might occur in diabetes insipidus or the syndrome of inappropriate antidiuretic hormone (SIADH). The major osmotic substances in plasma are Na^+, Cl^-, glucose, and urea; thus expected plasma osmolality can be calculated from the following empirical equation:

$$\begin{aligned}\text{mOsmol/kg} = {} & 1.86\,[\text{Na}^+\ (\text{mmol/L})] \\ & + \text{glucose}\,[\text{mmol/L}] \\ & + \text{urea}\,[\text{mmol/L}] \\ & + 9\end{aligned}$$

or

$$\begin{aligned}\text{mOsmol/kg} = {} & 1.86\,[\text{Na}^+\ (\text{mmol/L})] \\ & + \text{glucose}\,[\text{mg/dL}]/18 \\ & + \text{urea N}\,[\text{mg/dL}]/2.8 \\ & + 9\end{aligned}$$

The 9 mOsmol/kg added to the previous equation represents the contributions of other osmotically active substances in plasma, such as K^+, Ca^{2+}, and proteins, and 1.86 is two times the osmotic coefficient of Na^+, reflecting the contributions of both Na^+ and Cl^-. Some versions of this equation do not include the plus 9 mOsmol/kg factor.

The reference interval for plasma osmolality is 275 to 300 mOsmol/kg. Comparison of measured osmolality versus calculated osmolality can reveal the presence of an **osmolal gap**, which can be important in determining the presence of exogenous osmotic substances. Comparison of calculated and measured osmolalities can also confirm or rule out suspected pseudohyponatremia caused by the electrolyte exclusion effect. Virtually all clinical laboratories use freezing point depression osmometers for the measurement of osmolality because of its simplicity and the fact that vapor pressure osmometers will not detect volatile substances such as methanol, ethanol, or isopropanol. Furthermore, freezing point depression, unlike vapor pressure, is independent of changes in ambient temperature.

Freezing Point Depression Osmometer

The components of a freezing point depression osmometer, often simply referred to as an osmometer (Fig. 24.2) are as follows:

1. A thermostatically controlled cooling bath or block maintained at −7°C
2. A rapid stir mechanism to initiate (or "seed") freezing of the sample
3. A thermistor probe connected to a circuit to measure the temperature of the sample (The thermistor is a glass bead attached to a metal stem whose resistance varies rapidly and predictably with temperature.)
4. A galvanometer that displays the current from the thermistor, and that is used as a guide when the measuring potentiometer is used
5. A measuring potentiometer

In most instruments, components 4 and 5 are replaced by a light-emitting diode (LED) display that indicates the time course of the freezing curve and the final result.

During analysis, the sample, in which the thermistor probe and the stirring wire are centered, is lowered into the bath and, with gentle stirring, is super-cooled to a temperature several degrees below its freezing point (−7°C). When the LED display (or galvanometer) indicates that sufficient super-cooling has occurred, the sample is raised to a point above the liquid in the cooling bath, and the wire stirrer is changed from a gentle rate of stir to a short but vigorous rate, which initiates freezing of the super-cooled solution. This freezing occurs only to the slush stage, with about 2% to 3% of the solvent solidifying. The released heat of fusion initially warms the solution, and then the temperature plateaus and remains stationary, indicating the equilibrium temperature at which both freezing and thawing of the solution are occurring. At the end of the equilibrium temperature plateau, the galvanometer again indicates decreasing temperature as the sample freezes further toward a complete solid. An example of the calculation to obtain osmolality is as follows: if the observed freezing point is −0.53°C, then

$$\text{mOsmol/kg H}_2\text{O} = \frac{-0.53}{-1.86} \times 1000 = 285$$

where −1.86°C is the molal freezing point depression of pure water.

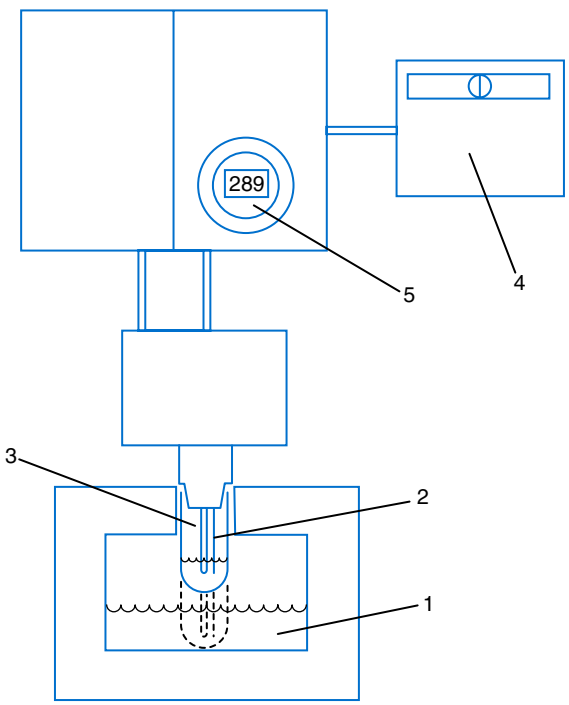


Fig. 24.2 Block diagram of a freezing point depression osmometer. *1,* Cooling fluid; *2,* stirring rod; *3,* thermistor; *4,* galvanometer; *5,* potentiometer with direct readout. The test tube is shown above the liquid in the cooling bath *(solid line)* and inside the cooling liquid *(dashed line).*

POINTS TO REMEMBER

- Dissolved solutes will result in a solution with an increased osmotic pressure, lowered vapor pressure, increased boiling point, and decreased freezing point.
- Freezing point depression osmometers are the most commonly used technique to measure osmolality in the clinical laboratory.
- Comparison of plasma and urine osmolality can be useful for the assessment of water balance and disorders such as diabetes insipidus and the syndrome of inappropriate antidiuretic hormone.
- A gap between the measured and calculated osmolality can reveal the presence of unmeasured osmolytes.

BLOOD GASES AND pH

The detection and clinical management of respiratory and metabolic disorders often depends on rapid and accurate measurements of the **partial pressures** of oxygen (pO_2) and carbon dioxide (pCO_2) in blood. Measures to support life in patients with cardiopulmonary impairment depend largely on assisted ventilation using mixtures of gases that are tailored in response to blood gas results. Modern instruments for blood gas determination are simple to operate and, with meticulous maintenance and quality control, are capable of rapid, reliable data. Details of the pathophysiology of blood gases in relation to respiration and acid-base disorders are discussed in detail in Chapter 36.

Nomenclature for this area of analysis has been recommended by the CLSI, but alternative nomenclatures exist and are in common use, as summarized in Box 24.1

BEHAVIOR OF GASES

Determination of gas pressures in blood depends on the application of certain physical principles (Table 24.1). The partial pressure (tension) of a gas dissolved in blood is by definition equal to the partial pressure of the gas in an imaginary ideal gas phase in equilibrium with the blood. The partial pressure of a gas is the same in erythrocytes and plasma; thus the partial pressure of a gas is the same in whole blood and plasma. The partial pressure of a gas in a gas mixture is defined as the substance fraction of gas (mole fraction) times the total pressure.

Various spaces where gases are present include the ambient environment (room air), the bronchial tree and alveoli of the patient, and the measuring chamber of a laboratory instrument. In all these spaces, atmospheric (barometric) pressure, *P*(Amb), is the prevailing pressure, and partial pressures of each of the gases present in these spaces must add up to the

BOX 24.1 Conversion Factors, Prefixes, Symbols, and Descriptors Used in Discussions of Gases Measured in Blood and Expired Air[a]

Conversion Factors

1 mm Hg = 0.133 kPa
1 kPa = 7.5 mm Hg
kPa: 1 kilopascal = 1000 pascal. The pascal is the SI derived unit of pressure; it equals 1 Newton/m^2.

General Prefixes

P (or p): partial pressure or tension
Usage: pO_2, pCO_2, pH_2O
Alternative: pO_2
S: saturation fraction
Usage: SO_2
Alternative: sO_2
c: substance concentration
Usage: ctO_2 for concentration of total O_2
Usage: $ctCO_2$ for concentration of total CO_2
Usage: $cHCO_3^-$ for concentration of bicarbonate
d: dissolved gas, used with substance concentration *(c)*
t: total, used with substance concentration *(c)*; thus

$$ctCO_2 = cHCO_3^- + cdCO_2$$

Specimen origin is indicated by lower case letters. Whole blood and plasma are distinguished by capitals.
a: arterial B: blood
v: venous P: plasma
c: capillary
Usage: pO_2(aB), for partial pressure of O_2 in arterial blood

Prefixes Associated With External Respiration

V: volume of air or blood (unit, L)
$\dot{V}$*:* volume rate (unit, L/min)
F: substance fraction, also called mole fraction
E: expired air
I: inspired air
A: alveolar air
Usage: $\dot{v}(A)$, alveolar ventilation; $\dot{v}(B)$, cardiac output; FO_2(I), fraction of O_2 in inspired air; pO_2(A), partial pressure of O_2 in alveolar air; and pCO_2(E), partial pressure of CO_2 in expired air.

Other Descriptors

BTPS: *B*ody *T*emperature (37°C or 310.16 K) and ambient *P*ressure, fully *S*aturated (pH_2O = 47 mm Hg or 6.25 kPa)
STPD: *S*tandard *T*emperature (0°C or 273.16 K) and standard *P*ressure (760 mm Hg or 101.08 kPa) of *D*ry gas
Amb: ambient atmosphere (unit is atm, atmosphere)
B: barometric (atmospheric)
BTPS: Usage: *P*(amb), *P(Amb)*
SVP: *S*aturated *V*apor *P*ressure, the vapor pressure of water. SVP_T means SVP at a specified temperature [e.g., $SVP_{37°C}$ = 47 mm Hg; pH_2O(saturated)]
ATPS: *A*mbient *T*emperature and *P*ressure, *S*aturated with water vapor

[a]This list is not complete but is presented to facilitate interpretation of terms used in the text and to illustrate various forms that may be encountered in the literature.
From Maas. A. H. (1991). IFCC reference methods for measurement of pH, gases and electrolytes in blood: reference materials. *European Journal of Clinical Chemistry and Clinical Biochemistry, 29*, 253–261.

TABLE 24.1 Physical Principles Applied in Blood Gas Measurements

Boyle's law: The volume of an ideal gas at a constant temperature varies inversely with the pressure exerted to contain it.	$V \propto 1/P$
Charles' (Gay-Lussac's) law: The volume of an ideal gas at a constant pressure varies directly with its absolute temperature.	$V \propto T$
Avogadro's hypothesis: Equal volumes of different ideal gases at the same temperature and pressure contain the same number of molecules.	$n_i/V_i = n_j/V_j$
Dalton's law: The total pressure exerted by a mixture of ideal gases is the sum of the partial pressures of each of the gases in the mixture.	$P = \Sigma P_i$
Henry's law: The amount of a sparingly soluble gas dissolved in a liquid is proportional to the partial pressure of the gas over the liquid.	$c = \alpha \times P$

value of P(Amb), which will vary with altitude and barometric pressure. Measurements of partial pressure used for clinical management are always made at *B*ody *T*emperature (37°C), at P(Amb), and in the presence of *S*aturated water vapor (pH_2O = 47 mm Hg). Use of this *BTPS* convention (see Box 24.1) has the following practical effects:

1. It relates laboratory data for blood gases strictly to the geographic location of the patient, so that reference intervals become altitude dependent.
2. It assumes a standard body temperature of 37°C and that the measuring device also holds the sample of blood at exactly 37°C. It also implies that in circumstances such as *imposed* hypothermia, when a patient's temperature is not 37°C, blood gas values determined at 37°C might need to be corrected to the actual body temperature to obtain an estimate of blood gas partial pressures in the patient.
3. It recognizes that partial pressures of measured gases in the blood coexist with a constant and standard saturated vapor pressure (SVP), which is identical for both the calibration conditions of the instrument and the measurement conditions of the blood sample.

Boyle's and Charles' laws and Avogadro's hypothesis are combined in what is called the *general gas equation*:

$$P = (nRT)/V$$

where P = pressure in units of millimeters of mercury (mm Hg) or kilopascals (kPa); V = volume in liters in which an ideal gas is contained; T = temperature in degrees kelvin (0°C = 273.16 K); n = number of moles of gas; and R = gas constant.

The SI unit of P is the pascal (Pa). However, millimeters of mercury (also called *torr*) have continued to remain popular in some countries (see Box 24.1 for conversion factors). Use of SI units does have a practical advantage in that 1 atm almost equals 100 kPa (1 atm = 101.325 kPa). Partial pressures expressed in kilopascals therefore are very close estimates of percentages of the gases in the mixture at 1 atm. Pressure, P (or p), may mean total pressure, as in the expression P(Amb) for the mixture of gases in ambient air, or partial pressure in arterial blood, as in pO_2(aB). Numeric value and units of R differ depending on the units used for P, so that

$$R = 6236 \text{ mm Hg} \times \text{L} \times \text{K}^{-1} \times \text{mol}^{-1}$$

or

$$R = 8.31 \text{ kPa} \times \text{L} \times \text{K}^{-1} \times \text{mol}^{-1}$$

It is important to note that pO_2 is related only to the concentration of dissolved O_2 (cdO_2) in the blood and pCO_2 to the concentration of dissolved CO_2 ($cdCO_2$) in the blood (see Henry's law; Table 24.1). The total concentration of O_2 in blood (ctO_2) is the sum of concentrations of dissolved O_2 and of O_2 bound to hemoglobin, with cdO_2 being a small component of tO_2 (see the later section on hemoglobin saturation). The total concentration of CO_2 ($ctCO_2$) is defined operationally as the sum of concentrations of dissolved CO_2, carbonic acid, HCO_3^-, undissociated bicarbonate, and carbonate ion.

Dalton's law of partial pressures governs the calibration and control of blood gas instruments. *Dalton's law* (see Table 24.1) may be written for room air as follows:

$$P(\text{Amb}) = pO_2 + pCO_2 + pN_2 + pH_2O + pX$$

where pX is the partial pressure of any other gas in the air sample.

Consider a calibrator gas certified to contain 15% O_2 (L/L or mol/mol) and 5% CO_2, the remainder being N_2. The mole fractions (or F) of the gases in the dry mixture are 0.15, 0.05, and 0.80, respectively. This mixture, after saturation with water vapor at 37°C (to mimic a patient's blood or alveolar air), is introduced into a blood gas instrument's measuring chamber (held at 37°C to mimic a patient's body temperature) for the purpose of calibrating the instrument for subsequent measurements of gases in patients' samples. If the local barometric pressure, P(Amb), on this occasion is 747 mm Hg, then humidified calibrator gas is present in the chamber at ambient, barometric pressure, such that

$$P(\text{Amb}) = 747 \text{ mm Hg} = pO_2 + pCO_2 + pN_2 + pH_2O$$

To set the instrument to the pO_2 and pCO_2 of the calibrator gas, the pH_2O at 37°C, which is equal to the SVP of water, 47 mm Hg must be included. Therefore,

$$\begin{aligned} 747 \text{ mm Hg} - pH_2O &= pO_2 + pCO_2 + pN_2 \\ &= 747 - 47 \\ &= 700 \text{ mm Hg} \end{aligned}$$

If P(Amb) corrected for pH_2O represents the sum of partial pressures for the dry gases whose mole fractions we know, we can calculate the exact pO_2 and pCO_2 values for the calibrator gas, under circumstances of measurement, and then enter these calibrator values into the instrument.

The law of partial pressure is also applied in defining gas mixtures used to determine $pO_2(0.5)$ or P_{50} and other derived quantities, and to control instrumentation with tonometered samples. *Henry's law* predicts the amount of dissolved gas in a liquid in contact with a gaseous phase (see Table 24.1).

The coefficient for O_2 in blood, αO_2, is 0.00140 (mmol/L)/mm Hg. Therefore, when arterial pO_2 is normal (≈100 mm Hg), the concentration of dissolved O_2 in arterial blood, cdO_2, is 0.140 mmol/L, which is a very small proportion of the ctO_2

content in blood (≈9 mmol/L), the bulk of which is O_2 bound by hemoglobin. Increasing the O_2 fraction of inspired air to 100% or increasing the pressure of inspired air, as in a hyperbaric chamber, forces more O_2 into a solution. The $cdCO_2$ can be calculated in the same way: αCO_2 at 37°C in plasma = 0.0306 mmol/L/mm Hg. Thus at a pCO_2 of 40 mm Hg, the $cdCO_2$ = 40 × 0.0306 = 1.224 mmol/L. In the determination of blood gases, pCO_2 is determined along with blood pH. As will be subsequently explained, these two parameters in conjunction with the Henderson-Hasselbalch equation permit the calculation of HCO_3^-, as follows:

$$\log cHCO_3^- = pH - pK' + \log [pCO_2 \times \alpha CO_2 (P)]$$

The antilog is then taken to derive $cHCO_3^-$.

APPLICATION OF THE HENDERSON-HASSELBALCH EQUATION IN BLOOD GAS MEASUREMENTS

Carbon dioxide and water react to form carbonic acid, which in turn dissociates to hydrogen ions and HCO_3^-.

$$CO_2 + H_2O \xrightleftharpoons{K_{hydration}} H_2CO_3 \xrightleftharpoons{K_{dissociation}} H^{\oplus} + HCO_3^{\ominus}$$

Thus the total concentration of CO_2 ($ctCO_2$), the concentration of bicarbonate ($cHCO_3^-$), the concentration of dissolved CO_2 ($cdCO_2$), and the H^+ ion concentration (cH^+) are interrelated.

This relationship and its effect on pH are described by the **Henderson-Hasselbalch equation** below. In plasma a small proportion of the $cdCO_2$ includes undissociated (dissolved) carbonic acid. The amount in plasma can be expressed as $cdCO_2 = \alpha \times pCO_2$, where α represents the solubility coefficient of CO_2.

$$pH = pK' + \log \frac{cHCO_3^-}{\alpha \times pCO_2}$$

pK' in the equation above is the negative log of the apparent overall (combined) dissociation constant for carbonic acid. It is *apparent* because concentrations are used rather than activities and *overall* because both the $cdCO_2$ and the concentration of carbonic acid are used. K' (and thus pK') depends not only on the temperature but also on the ionic strength of the solution.

For blood at 37°C, the normal mean value is p$K'(P)$ = 6.103, and the normal mean solubility coefficient for CO_2 gas is (0.0306 mmol) × (L^{-1}) × (mm Hg^{-1}). When pK' and α for normal plasma at 37°C are inserted, the Henderson-Hasselbalch equation takes the following form:

$$pH = 6.103 + \log \frac{cHCO_3^-}{0.0306 \times pCO_2}$$

where pCO_2 is reported in millimeters of mercury and $cHCO_3^-$ is reported in millimoles per liter.

By measuring any two of the three parameters, pCO_2, pH, or $cHCO_3^-$, and by using the Henderson-Hasselbalch equation with the above values for pK' and α, the other parameter may be calculated. Although used as constants, pK' and α must be recognized as means and are susceptible to biological variation. Changes in ionic strength of ±20% cause changes in pK' between 6.08 and 6.12. Variations in pK' of plasma also occur with temperature (pK will decrease by 0.0026 per 1°C increase and will decrease slightly with increasing pH). For most clinical purposes, these variations of pK' can be ignored. However, in pathologic cases with markedly deviant ionic strength, the change in pK' may be significant. The value of α is affected by the presence of increased ions or proteins in solution (value decreases) or lipids (value increases); for instance, in lipemic plasma, the value of α may be 0.033 or higher. Thus, parameters calculated on the assumption that pK' and α are invariant may have error under certain pathologic circumstances. Several authors suggest caution in using calculated HCO_3^- values from blood gas analyzers in extremely ill patients, as the calculated and measured tCO_2 values can differ by greater than 10% in adult and greater than 20% in pediatric ICU patients, respectively.

OXYGEN IN BLOOD

The total O_2 content (ctO_2) of a blood sample is the sum of the concentrations of hemoglobin-bound O_2 and of dissolved O_2. At a blood ctO_2 of 9 mmol/L, the O_2 associated with hemoglobin as **oxyhemoglobin** (O_2Hb) is 8.86 mmol/L. The O_2Hb is defined as erythrocyte hemoglobin with O_2 reversibly bound to Fe^{2+} of its heme group.

Hemoglobin A (Hb A, the normal adult gene product) reversibly binds O_2 via the heme-Fe^{2+} moiety and binds biological effectors at other allosteric sites. Methemoglobin (MetHb), carboxyhemoglobin (COHb), sulfhemoglobin (SulfHb), and cyanmethemoglobin, collectively termed dyshemoglobins, are forms of hemoglobin that are not capable of reversible O_2 binding because of chemical alterations of the heme moiety. Another group of abnormal hemoglobins, collectively referred to as *hemoglobin variants* or *hemoglobinopathies*, have genetically determined changes in their amino acid sequence that can alter their O_2 affinity. More than 900 hemoglobin variants have been described, of which only a small fraction are clinically significant (for details see Chapter 28).

Uptake of O_2 by blood in the lungs is governed primarily by the pO_2 of alveolar air and by the ability of O_2 to diffuse freely across the alveolar membrane into blood. At the pO_2 normally present in alveolar air (≈102 mm Hg) and with a normal membrane and normal hemoglobin A, more than 95% of hemoglobin will bind O_2. At a pO_2 > 110 mm Hg, more than 98% of normal hemoglobin A binds O_2. When all hemoglobin is saturated with O_2, further increases in the pO_2 of alveolar air simply increase the concentration of cdO_2 in the arterial blood. Delivery of O_2 by blood to tissues is governed by the large gradient between pO_2 of arterial blood and that of the tissue cells, and by the dissociation of O_2Hb in the erythrocytes at the lower pO_2 of the blood-tissue cell interface.

Three properties of arterial blood are essential to ensure adequate O_2 delivery to tissues:

1. Arterial pO_2 must be sufficiently high (≈90 mm Hg) to create a diffusion gradient from arterial blood to the tissue cells. The pO_2 at the venous end of the capillaries is typically ~30 to 45 mm Hg; thus the normal arteriovenous difference in pO_2 is 50 to 60 mm Hg. Low arterial pO_2 *(hypoxemia)* results in tissue O_2 starvation *(hypoxia)*.

2. The O_2-binding capacity of blood must be normal (i.e., the concentration of hemoglobin capable of binding and releasing O_2 must be normal). Decreased Hb concentration will cause so-called **anemic hypoxia**.
3. The hemoglobin must be able to bind O_2 in the lungs yet release it at the tissue. In other words, the affinity of hemoglobin for O_2 must be normal. Too great an affinity of hemoglobin for O_2 may cause "affinity-based" tissue hypoxia, in which O_2 is not released at the capillary-tissue interface. There are approximately 100 high-affinity hemoglobin variants described to date; however, individually they are all rare. Hemoglobin F, while not a hemoglobin variant, has a higher affinity for oxygen than hemoglobin A, the normal adult hemoglobin.

HEMOGLOBIN OXYGEN SATURATION

Before the factors that affect Hb affinity for O_2 are discussed, it is important to define the concept of hemoglobin **oxygen saturation (SO_2)**:

$$SO_2 = \frac{\text{Oxygen Content}}{\text{Oxygen Capacity}}$$

This is the fraction (percentage) of functional hemoglobin that is saturated with oxygen and is essentially an indirect means of estimating the pO_2. However, at least three different approaches exist for determining oxygen "saturation," and although each is distinct, they are often used interchangeably to determine "oxygen saturation." These three terms—hemoglobin oxygen saturation (SO_2), fractional oxyhemoglobin (FO_2Hb), and estimated oxygen saturation (O_2Sat)—have distinct definitions described in CLSI C46-A2. Ambiguous use of these terms occurs because in healthy subjects with normal amounts of normal hemoglobin, the values for all three entities are very similar. However, these assumptions can lead to erroneous conclusions in seriously ill patients and those with dyshemoglobins or hemoglobin variants.

Spectrophotometric methods are used to determine O_2Hb and HHb, and SO_2 is calculated according to

$$SO_2 = \frac{cO_2Hb}{cO_2Hb + cHHb}$$

where cO_2Hb is the concentration of oxyhemoglobin, cHHb is the concentration of deoxyhemoglobin, and the sum of oxyhemoglobin and deoxyhemoglobin represents all hemoglobin capable of reversibly binding O_2. SO_2 is usually expressed as a percent in the United States, but it may also be expressed as a decimal fraction of 1.00.

SO_2 most often is determined by simple pulse **oximetry** and cannot determine COHb, MetHb, or SulfHb concentrations. These devices measure absorbance at 660 and 940 nm for which O_2Hb and HHb have unique absorbance patterns. Pulse oximeters are appropriate for monitoring HbO_2 saturation, and they serve this purpose extremely well. However, use of SO_2 in the initial evaluation of a patient with dyshemoglobins can be very misleading. For instance, in a comatose patient with 15% COHb, the SO_2 by simple pulse oximetry might read 0.95, whereas the fraction of oxyhemoglobin in reality would be only 0.80. Thus it seems reasonable to assess for the presence of dyshemoglobins before SO_2 is used for clinical purposes. The reference interval for SO_2 from healthy adults is 0.94 to 0.98 (94% to 98%).

Another expression of oxygen "saturation" is fractional oxyhemoglobin (FO_2Hb), which is calculated as

$$FO_2Hb = (cO_2Hb/cHHb + cCOHb + cMethHb + cSulfHb)$$

This value requires determination of all hemoglobin species and can be performed on a co-oximeter present in modern blood gas analyzers. These instruments prepare a hemolysate from whole blood by sonication, and spectrophotometrically determine the total amount of hemoglobin and the percent of each of the aforementioned species. This is accomplished by using monochromatic light at 6 to 128 fixed wavelengths between 535 and 670 nm and measuring absorbance at each of the wavelengths. Newer co-oximeters use a diode array and 128 wavelengths. Because each species of hemoglobin has its own absorbance pattern, an integrated software algorithm calculates the percent of each one. The reference interval for FO_2Hb is 0.90 to 0.95 (90% to 95%).

Finally, using empirical equations, the integrated software in blood gas instruments can estimate the oxygen saturation from measured pH, pO_2, and hemoglobin. This value should be clearly referred to as *estimated oxygen saturation*, but it frequently is reported and referred to as "O_2Sat." If calculated values such as "O_2Sat" are reported, they should be interpreted with reservation because the algorithm assumes normal hemoglobin affinity for O_2, normal 2,3-diphosphoglycerol (2,3-DPG) concentrations, and the absence of dyshemoglobins. Such calculated estimates have been found to vary by as much as 6% saturation from measured values.

Decreases in arterial FO_2Hb may indicate a low arterial pO_2 or an impaired ability of hemoglobin to bind O_2. The amount of O_2 that the blood can carry is determined by three major factors: (1) the pO_2, which reflects how much O_2 is dissolved in the blood; (2) the amount of normal hemoglobin available in erythrocytes; and (3) the affinity of available hemoglobin for O_2. Decreases in pO_2 indicate a reduced ability of O_2 to diffuse from alveolar air into the blood. This may be due to hypoventilation or to increased venoarterial shunting that is secondary to cardiac or pulmonary insufficiency, resulting in an increased fraction of blood that has not reached equilibrium with alveolar air. Hypoventilation and increased shunting are also characterized by an increased pCO_2. Decreases in total hemoglobin concentration can result from a decreased number of erythrocytes that contain a normal concentration of hemoglobin (normochromic anemia) or a decreased mean cell concentration of hemoglobin in the erythrocytes (hypochromic anemia). Decreased FO_2Hb hemoglobin can also occur as a result of poisonings that convert part of the hemoglobin into the species COHb, MetHb, or SulfHb, which cannot properly bind or exchange O_2. Clinically it is important to distinguish between arterial hypoxemia (decreased arterial pO_2 and decreased FO_2Hb caused by decreased availability of O_2) and cyanosis (decreased FO_2Hb caused by abnormally high concentrations of reduced hemoglobin or chemically

altered hemoglobin incapable of carrying O_2). Note that in the setting of cyanosis, measurement of SO_2 or an estimated SO_2 ("O_2Sat") could be normal if cyanosis is due to the presence of MetHb or COHb.

Hemoglobin-Oxygen Dissociation

The degree of association or dissociation of O_2 with hemoglobin is determined by pO_2 and the affinity of hemoglobin for O_2. When the SO_2 of blood is determined over a range of pO_2 and is plotted against pO_2, a sigmoidal curve called the *O_2 dissociation curve* is obtained. The *shape* of the curve is affected by the increasing efficiency with which HHb molecules bind more O_2 once some O_2 has been bound (*cooperativity*; see also Chapter 28). The *location* of the curve relative to the pO_2 required to achieve a particular concentration of SO_2 in the blood is a function of the affinity of hemoglobin for O_2.

The affinity of hemoglobin for O_2 depends on five factors: temperature, pH, pCO_2, 2,3-DPG concentration, and the presence of dyshemoglobins such as COHb and MetHb. Dissociation of O_2 in relation to 2,3-DPG concentration and abnormal hemoglobin proteins, such as fetal hemoglobin (Hb F), thalassemias, and other hemoglobinopathies, are discussed in Chapter 28.

The graph in Fig. 24.3 shows the effect of plasma pH on the O_2 dissociation curve (the Bohr effect). Similar graphs can be made for variations of pCO_2, temperature and 2,3-DPG concentration.

Determination of P_{50}

P_{50} is defined as the pO_2 at which the hemoglobin of the blood is half saturated with O_2. The measured value of P_{50} differs from the standard value of P_{50} by some amount determined by the extent that pH differs from 7.40, pCO_2 differs from 40 mm Hg, temperature differs from 37°C, and 2,3-DPG differs from 5.0 mmol/L. The value of P_{50} therefore becomes a measure of change of hemoglobin affinity because of the factors that affect it.

Clinical Significance

Increased values for P_{50} indicate displacement of the O_2 dissociation curve to the right (i.e., decreased affinity of the hemoglobin for O_2). The chief causes are hyperthermia, acidemia, hypercapnia, high concentrations of 2,3-DPG, and the presence of a hemoglobin variant with decreased O_2 affinity. The physiologic effects of decreased affinity of hemoglobin are small. The affinity is still sufficient to bind adequate amounts of O_2 in the lungs, and the low affinity facilitates dissociation of O_2Hb at the peripheral tissues. Indeed, in anemia, low affinity (as a result of increases in 2,3-DPG) is a desirable compensatory mechanism. The compensatory responses in healthy individuals to alter the P_{50} can be remarkable. For instance, in fully acclimatized (70 days) high-altitude climbers (Mt. Everest), full cognition was maintained without supplemental oxygen at an altitude of 8400 m and an average $P(a)O_2$ of only 25 mm Hg.

Low values for P_{50} signify displacement of the O_2 dissociation curve to the left (i.e., increased affinity of hemoglobin). The main causes are hypothermia, acute alkalemia, hypocapnia, low 2,3-DPG, and variant hemoglobin. The physiologic consequence of increased affinity of hemoglobin for O_2 is less efficient dissociation of O_2Hb at the peripheral tissue and lower tissue pO_2.

DETERMINATION OF pCO_2, pO_2, AND pH

The instruments used for determination of pCO_2, pO_2, and pH are highly automated. Proper specimen collection and handling are critical for accurate determinations.

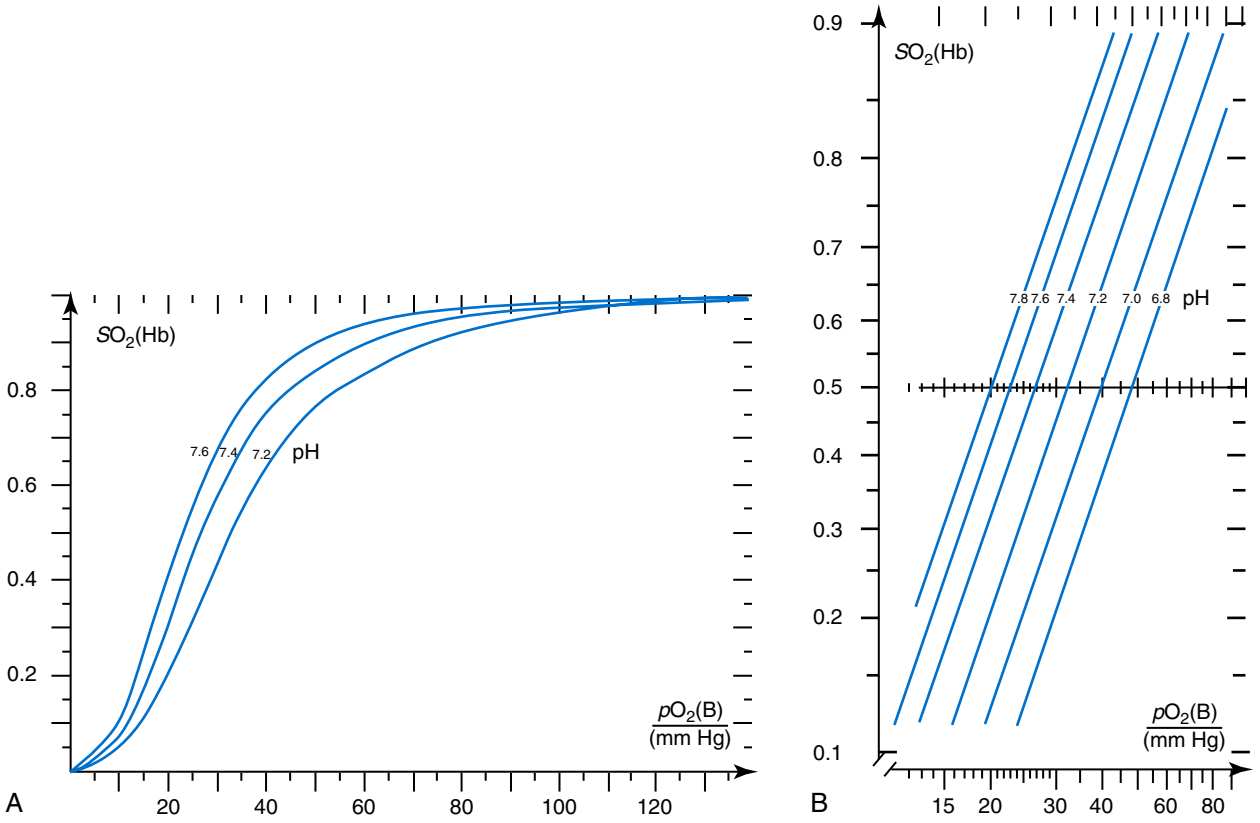
Specimens

Whole blood is the most likely specimen for a clinical laboratory to receive for gas analysis. Differences in measured blood gas values between arterial and venous blood are most pronounced for pO_2. pO_2 is generally ≈60 mm Hg lower in venous blood after O_2 is released in the capillaries, whereas pCO_2 is 2 to 8 mm Hg higher in venous blood. pH generally is only 0.02 to 0.05 pH units lower in a venous sample.

Proper specimen collection is critical to obtaining accurate results for blood analysis for gases and pH. Placement of indwelling catheters with heparin locks for short- and long-term intravenous therapies is common. Failure to flush the catheter properly has unpredictable effects on measured quantities and is often indicated by bizarre, nonphysiologic results.

Arterial or venous specimens are collected anaerobically with lyophilized (balanced) heparin anticoagulant in 1 to 3 mL sterile syringes. Lyophilized heparin is preferable to liquid heparin because liquid heparin, which has atmospheric pO_2 and pCO_2 values, dilutes the sample, with the effect being greatest when the syringe is not completely filled. An increasing ratio of liquid heparin to blood can have an increasingly marked effect on measured pCO_2 and the parameters calculated from it. Sample dilution with 10% volume of liquid heparin having an atmospheric pCO_2 of 0.25 mm Hg has been shown to lead to a 10 to 20 mm Hg decrease in measured blood pCO_2.

Anaerobic technique for collection means no exposure of blood to atmospheric air. The pCO_2 of air is about 0.25 mm Hg; this is much less than that of blood (≈40 mm Hg). Thus the CO_2 content and pCO_2 of blood exposed to air will decrease, and blood pH, which is a function of pCO_2, will rise. The PO_2 of atmospheric air (≈155 mm Hg) is ≈60 mm Hg higher than that of arterial blood and ≈100 mm Hg higher than that of venous blood. Therefore blood from patients' breathing room air that is exposed to atmospheric air gains O_2, and blood with pO_2 >150 mm Hg, as occurs in patients undergoing O_2 therapy, loses O_2. Blood can also be exposed to air simply from the air in the needle and the syringe hub dead space. Error will be minimal if the resulting bubble is ejected immediately after drawing by holding the syringe tip up and expelling the bubble(s). The potential effect of small bubbles on blood gas results was clearly demonstrated in



C

Deviation from standard conditions	Shift in dissociation curve	Affinity of hemoglobin for O_2	Coefficient of change*
pH(P) >7.4	←	↑	$\frac{\Delta \log pO_2}{\Delta pH(P)} = -0.46$
pH(P) <7.4	→	↓	
Temperature			
>37 °C	→	↓	$\frac{\Delta \log pO_2}{\Delta T} = +0.024\ K^{-1}$
<37 °C	←	↑	
pCO_2 >40 mm Hg	→	↓	$\frac{\Delta \log pO_2}{\Delta \log pCO_2} = +0.02$
pCO_2 <40 mm Hg	←	↑	
cDPG(E)			
>normal	→	↓	$\frac{\Delta \log pO_2}{\Delta(cDPG(E)/c^*)} = +0.04$
<normal	←	↑	

Fig. 24.3 (A) Oxygen dissociation curves for human blood with different plasma pH but constant pCO_2 of 40 mm Hg, a 2,3-diphosphoglycerol (2,3-DPG) concentration in erythrocytes of 5.0 mmol/L, and temperature at 37°C. (B) A Hill plot. Conditions are the same as in A. The coefficients given in this chart form the basis for the correction of measured pO_2. (C) The effect of pH(P) to shift the dissociation curve is called the *Bohr effect*; the coefficient above for $\Delta \log pO_2/\Delta pH(P)$ applies to conditions when the pCO_2 is 40 mm Hg, and changes in pH(P) are due to changes in concentrations of noncarbonic acids and bases. If, however, the changes in pH(P) are being caused by changes in pCO_2, then the absolute value of the coefficient is greater (i.e., $\Delta \log pO_2/\Delta pH[P] = -0.49$). The coefficients for the Bohr effect are specified for pO_2 of whole blood but use the pH of plasma, pH(P). The coefficient for DPG effect is based on the DPG concentration in the erythrocytes, cDPG(E). $c^* = 1$ mmol/L.

one study in which a 100 µL bubble of room air was added to ten 2 mL blood samples with pO_2 values between 25 and 40 mm Hg. In these samples, pO_2 increased an average of 4 mm Hg in only 2 minutes, whereas pCO_2 decreased 4 mm Hg. Before analysis, mixing of the sample by vigorously rolling the syringe between the palms should be done to establish a homogeneous sample.

Arterialized capillary blood is sometimes an acceptable alternative to arterial blood when an arterial cannula is not available, or when repeated arterial puncture must be avoided. Freely flowing cutaneous blood originates in the arterioles and corresponds closely to arterial blood in composition. However, arterialized capillary blood is not acceptable when systolic blood pressure is less than 95 mm Hg. Capillary

puncture should be preceded by warming of the selected skin puncture site for 10 minutes to achieve vasodilation and adequate blood flow. For collection from the finger of a child or an adult or from an infant's heel, warming may be accomplished by immersing the arm or leg in water warmed to 42°C. The first blood drop to appear should be wiped away, and subsequent free-forming drops should be taken up in a capillary collection tube containing lyophilized heparin. Only free-flowing blood provides a satisfactory sample, and collecting the drops as soon as they form minimizes aerobic exposure. *Transport and analysis* of specimens should be prompt. However, delayed analysis of up to one hour will have a minimal effect on reported values from most samples. The pH of freshly drawn blood decreases on standing at a rate of 0.04 to 0.08 pH unit/h at 37°C, 0.02 to 0.03/h at 22°C, and less than 0.01/h at 4°C. The decrease in pH is accompanied by a corresponding decrease in glucose and an equivalent increase in lactate. pCO_2 increases by ≈5 mm Hg/h at 37°C, by 1 mm Hg/h at 22°C, and by only ≈0.5 mm Hg/h at 2 to 4°C. The primary cause of these changes is glycolysis by leukocytes, platelets, and reticulocytes. In freshly drawn blood with a normal pO_2 that is maintained anaerobically, cell respiration causes PO_2 to decrease at a rate of ≈2 mm Hg/h at room temperature. Adverse effects of glycolysis and respiration on pH, $ctCO_2$, pO_2, and pCO_2 of blood can best be prevented by analysis within 30 to 60 minutes after collection. If analysis must be delayed longer than 30 minutes, the syringe should be immersed in a mixture of ice and water until analysis. Under these conditions, changes are negligible because glycolysis is inhibited. With today's blood gas instrumentation, introduction of a chilled sample carries little risk of low-temperature effect on measurements, because thermal equilibration to 37°C is rapid and complete.

The aforementioned small changes in values that can be expected with delays in analysis are true *only* when the white blood cell (WBC) count is normal or only slightly elevated. Glycolysis and the resulting effects on pH, pO_2, and pCO_2 increase dramatically with markedly elevated WBC, such as occurs in leukemia. Experiments have shown that pO_2 decreases by 20 mm Hg in just two minutes and by 40 mm Hg in only 5 minutes with WBC values greater than 100,000/μL as observed in patients with "blast crisis." Indeed, for samples with these types of WBC values, it is very difficult to overcome this effect. Even after the sample is immersed in ice, thermal equilibrium takes several minutes, allowing significant pO_2 loss before the contents reach 4°C. The only alternative to obtaining accurate blood gas values on such patients is immediate analysis performed at the point of care.

Instrumentation

A schematic diagram characteristic of a typical instrument is shown in Fig. 24.4. Electrochemical principles and structural features of electrodes are discussed in Chapter 10.

Operation of a traditional blood gas instrument begins with the operator presenting a blood specimen at the sample probe. The sample is taken through the probe by a peristaltic pump that loads the chamber with 60 to 150 μL of sample. The sample then resides in the chamber long enough to allow thermal equilibration and completion of measurements. On completion of measurement, the pump pushes the sample to waste.

Because electrodes are not stable over long periods of time, frequent calibration of pH, pCO_2, and pO_2 is required. The instrument (see Fig. 24.4) is designed so that a manually actuated or a microprocessor actuated valve (self-calibrating) admits calibrator gases, standard buffers, or a sample to a small chamber (C) maintained at 37°C ± 0.1°C. Most instruments contain a barometer so that barometric pressure P(Amb) is always known to the microprocessor during calibration. Almost all manufacturers now produce small, portable, stand-alone, and easy-to-operate instruments designed for "satellite lab" operations; several handheld devices that use disposable electrodes are also available. Readers are referred to Chapter 17 for a discussion of **point-of-care testing**.

Maintenance of Instrumentation

Sophistication of contemporary equipment and availability of high-quality calibrator materials have made reliable and accurate determination of blood pH and gases primarily a matter of meticulous maintenance, adherence to the manufacturer's recommended procedures, control of the equipment, and proper collection and handling of specimens. Software programs of the instrument's microprocessor often provide display warnings and diagnostic routines that alert the operator and assist in troubleshooting. The manufacturer's suggested maintenance schedule should be considered a minimum guideline, with reliance on experience to indicate maintenance frequency.

Cleanliness of the sample chamber and path is especially important. Automatic flushing to cleanse the sample chamber and path after each blood sample measurement is a feature of most instruments without disposable electrodes. Despite proper flushing, however, complete or partial clogging of the chamber, the path, or both may occur. Fibrin threads and small clots may be present in the specimen or may form while the sample resides in the warm chamber. If allowed to remain, they can affect subsequent measurements or calibrations by interfering with contact of blood, buffers, or gases with electrode membranes. Visibility of the path through the heat sink is helpful for detecting clogs, dirt, and bubbles. Bubbles that fail to rinse out can be a particular problem if they settle on an electrode.

Quality Assurance and Quality Control

Elements of good quality assurance of blood gas and pH measurements include (1) proper maintenance of the instrument, (2) use of control materials, (3) verification of electrode linearity, (4) checking of barometer accuracy, and (5) accurate measurement of temperature.

Minimum calibration and quality control (QC) frequency is mandated by federal law in the United States (Clinical Laboratory Improvement Amendments [CLIA] '88). CLIA regulations require a one-point calibration every 30 minutes (or within 30 minutes of a patient sample) and a two-point

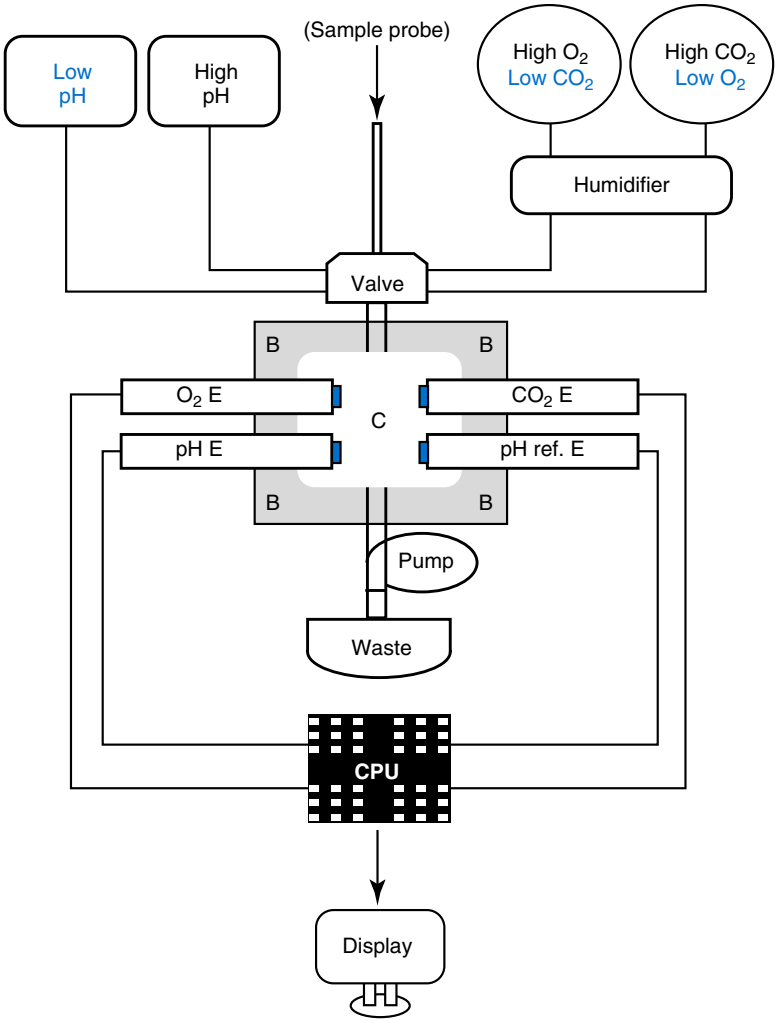


Fig. 24.4 Diagram of blood gas instrumentation. *B,* Constant temperature bath at 37°C; *C,* chamber; *CPU,* computer; *E (electrodes),* pH and gas standards are shown at top of diagram.

calibration every 8 hours. For QC, the minimum CLIA required frequency is one concentration of QC material every 8 hours, with the entire range of control concentrations covered in every 24-hour period. In many laboratories, however, the practical answer is to run three concentrations of QC for pH, pO_2, and pCO_2 on every instrument in use, at least once per shift following completion of maintenance and troubleshooting procedures. Newer analyzers, particularly the smaller satellite laboratory and point-of-care instruments, frequently have an "auto quality control (QC)" feature or use "electronic QCs." Auto QC consists of onboard QC material that is automatically analyzed by the instrument at designated intervals that fulfill regulatory requirements. Electronic QC, which is most common in devices with disposable electrode cartridges, consists of cartridges that verify the electronic specification of the instruments. For further discussion of these issues, see Chapter 17 on point-of-care testing.

CLIA similarly mandated external quality assurance (proficiency testing) and set criteria for satisfactory interlaboratory performance as follows: pH, target value ± 0.04; pO_2, target value ± 3 SD; and pCO_2, target value ± 8% or ± 5 mm Hg, whichever is greater.

Blood-Based and Fluorocarbon-Based Control Materials

Commercial blood-based control material usually consists of tanned human erythrocytes suspended in buffered medium and sealed in vials with a gas mixture of known O_2 and CO_2 content. Nonblood fluorocarbon materials with O_2-carrying properties similar to those of blood are also available. These products usually are made at three concentrations of pH, pCO_2, and pO_2. Unopened, these types of control materials have the advantages of a long refrigerated shelf life.

Aqueous Fluid Control Materials

These materials consist of a buffered medium with gas mixtures in sealed glass ampules. Immediately prior to opening the vial and introducing it into the instrument, the fluid is equilibrated with the gas by vigorous shaking by hand. The disadvantages of aqueous controls stem from their dissimilar matrix to blood. Lower viscosity and surface tension confer different washout characteristics and impair their ability to detect clogging. Greater electrical conductivity reduces their effectiveness in detecting inadequate grounding, and lower thermal coefficients make them slower to detect failures of

temperature control. Nevertheless, aqueous commercial controls are the most commonly used materials for quality control purposes.

Sources of Analytical Error

General causes of analytical error include calibration of the instrument with incorrect set points for pH buffers or calibrator gases, degraded calibration materials, failure of temperature control of the measurement chamber, and a dirty sample chamber or path. Incorrect calibration may arise from wrong entries made for buffer or gas values into the microprocessor, from incorrect manual calculations of pCO_2 and pO_2 values for calibrator gases, or from using gases that are dry because the humidification device is not working properly. Measurements of pO_2 are particularly sensitive to temperature error. To keep systematic error to 1% to 2%, the temperature control at 37°C must be within ±0.1°C. Built-in barometers can be checked by contacting nearby meteorological stations or by verifying the pressure against a NIST-certified barometer placed in the laboratory. Gases other than O_2 present in a blood sample may affect performance of the pO_2 electrode. The anesthetic gases halothane and nitrous oxide have a direct effect, because both can be reduced at the polarized cathode in competition with O_2. Under most circumstances, however, these effects are small and can be ignored.

Reference Intervals

Reference intervals for arterial blood pO_2, SO_2, pCO_2, and pH are extensively described in Chapter 36. Values for pO_2 and pCO_2 will decrease with increasing altitude, but compensatory mechanisms keep pH values the same.

pCO_2 values decrease with altitude above sea level at a rate of 3 mm Hg/km (5 mm Hg/mile). A physiologic change occurs with change in posture; pCO_2 is 2 to 4 mm Hg higher for a sitting or standing person than for one in the supine position. During pregnancy, pCO_2 falls gradually to a mean of about 28 mm Hg just before term.

TEMPERATURE CORRECTION OF MEASURED pH, pCO_2, AND pO_2

In the Henderson-Hasselbalch equation, pK' and α are used as constants for a temperature of 37°C. The temperature-controlled sample chamber of an instrument is specified to be 37°C ± 0.1°C, and it is at that temperature that all measurements of pH and partial pressure of gases are made. The body temperature of a febrile patient may be elevated to 40°C to 41°C, or a patient may be made hypothermic for certain surgeries and have a temperature as low as 23°C. Most blood gas instruments on keyboard entry of a patient's actual temperature can calculate and present temperature-corrected pH and pCO_2. Correction of pH and pCO_2 to the actual temperature of the patient is usually omitted in states of hyperthermia. However, significant disagreement exists with respect to hypothermic states.

The equation shown in Table 24.2 illustrates the complexity of the calculation used to correct pO_2 to the patient's body temperature. Complexity is unavoidable because at pO_2 <100 mm Hg (SO_2 ≤0.95), the hemoglobin-oxygen dissociation curve is shifted to the left by the decrease in temperature and by the concomitant rise in pH (see Fig. 24.3). For temperature corrections of pO_2 between 100 and 400 mm Hg, accurate formulas become even more complicated.

TABLE 24.2 Temperature Correction Formulas Recommended by the CLSI23

pH	$pH(T) = pH(37°C) - [0.147 + 0.0065 \times (pH[37°C] - 7.40) \times (T - 37°C)]$
pCO_2	$pCO_2(T) = pCO_2(37°C) \times 10^{[0.019 \times (T - 37°C)]}$
pO_2	$pO_2(T) = pO_2(37°C) \times 10^{\left(\frac{5.49 \times 10^{-11} \times pO_2^{3.88} + 0.071}{9.72 \times 10^{-9} \times pO_2^{3.88} + 2.30}\right) \times (T - 37°C)}$

CONTINUOUS AND NONINVASIVE MONITORING OF BLOOD GASES

Obtaining arterial, venous, or capillary blood is an invasive procedure, and test results reflect conditions pertaining only to a single point in time. Repetitive sampling carries risks, including infection and vascular complications. In premature infants particularly, repeated sampling imposes an undesirable blood loss.

Extensive discussion of noninvasive and continuous modes of monitoring is beyond the purview of this text. However, laboratorians must be aware of them, because blood samples and standard analytical equipment remain the reference for monitoring the effectiveness of such devices.

The accuracy of transcutaneous pO_2 monitoring can vary widely, depending on whether the site of application reflects arterial, capillary, or venous blood flow. Because there is little difference between arterial and venous pCO_2, transcutaneous monitoring of pCO_2 is less problematic, and pulse oximeters can often serve as a surrogate for pO_2. Transcutaneous monitoring works best in areas of thin skin, hence its popularity in neonates and in settings of euvolemia. In addition to monitoring pO_2 and pCO_2, new in-line devices that also monitor pH, electrolytes, glucose, and hematocrit have been introduced to the market.

POINTS TO REMEMBER

Blood gas instruments use the Henderson-Hasselbalch equation to calculate bicarbonate concentration and extremes of sample ionic strength; protein and lipid concentration may alter the accuracy of this calculated value.

The presence of bubbles in blood gas syringes has the potential to dramatically change the pO_2 and pCO_2 of blood gas samples.

Hemoglobin oxygen saturation (SO_2), fractional oxyhemoglobin (FO_2Hb), and estimated oxygen saturation may produce very dissimilar values in the presence of dyshemoglobins and hemoglobin variants.

Blood gas analyzers have more stringent requirements for the frequency of calibration and quality control assessment.

REVIEW QUESTIONS

1. The most common method for measuring sodium and potassium is
   a. flame emission spectrophotometry.
   b. enzymatic methods.
   c. atomic absorption spectrophotometry.
   d. ion-selective electrode.
2. Which of the following will shift the O_2 dissociation curve to the right?
   a. Hyperthermia
   b. Acute alkalemia
   c. Hypocapnia
   d. Low 2,3-DPG
3. Which of the following present the least risk of erroneous blood gas values?
   a. 30 minutes at ambient temperature
   b. WBC >100,000/μL
   c. Air bubbles in the sample
   d. First drop from a capillary puncture
4. The vapor pressure osmometer has fallen out of popular use due to several technical drawbacks. Which of the following are false regarding this instrument?
   a. It does not include any volatile solutes in its measurement of serum osmolality.
   b. It is sensitive to changes in ambient temperature.
   c. It contains a thermistor, a galvanometer, and a measuring potentiometer.
   d. It would be able to detect an increase in osmolality attributable to increased blood glucose.
5. Recent developments in genetic testing, paired with advances in understanding genotype-phenotype correlations in cystic fibrosis, have broadened the testing options for this serious condition. Currently the gold standard test for the diagnosis of cystic fibrosis is:
   a. sweat conductance.
   b. sweat chloride.
   c. immunoreactive trypsinogen.
   d. assay to detect common mutations.
   e. gene sequencing to detect common and novel mutations.
6. Solutes change the behavior of the solvent to which they are added in predictable ways, conferring colligative properties to the solution, including
   a. decreased osmotic pressure.
   b. increased vapor pressure.
   c. decreased boiling point.
   d. decreased freezing point.
   e. increased transmittance.
7. The most common cause of falsely elevated sweat chloride testing results is
   a. eczema.
   b. improper gauze placement.
   c. hypothyroidism.
   d. postcollection evaporation.
   e. analytic error.
8. What is the expected effect on plasma K^+ concentration if a freshly drawn sample is stored on the bench for 1.5 hours prior to being centrifuged?
   a. An increase in plasma K^+ of ~2 mmol/L
   b. A decrease in plasma K^+ of ~0.2 mmol/L due to transcellular shift into RBCs
   c. An insignificant increase in plasma K^+ of ~0.2 mmol/L
   d. A decrease in plasma K^+ due to increased hemolysis
   e. A decrease in plasma K^+ ~2 mmol/L due to transcellular shift into RBCs
9. Hemolysis of a plasma sample will most likely affect which of the following, and why?
   a. Sodium; it is present in equal amounts both intracellularly and extracellularly
   b. Glucose; it moves out of cells into plasma
   c. Insulin; it is the major intracellular anion and is present in high amounts in RBCs
   d. Potassium; it is localized mainly within cells, particularly RBCs
10. Which of the following equations is the correct Henderson-Hasselbalch equation?
   a. $pK' + pH + \log [cHCO_3]/[cH_2CO_3]$
   b. $pK' + pH + \log [cH_2CO_3]/[cHCO_3]$
   c. $pH + pK' + \log [cHCO_3]/[c\ H_2CO_3]$
   d. $pH + pK' + \log [c\ H_2CO_3]/[cHCO_3]$

SUGGESTED READINGS

Abraham, B., Fakhar, I., Tikaria, A., Hocutt, L., Marshall, J., Swaminathan, S., et al. (2008). Reverse pseudohyperkalemia in a leukemic patient. *Clinical Chemistry*, *54*, 449–451.

Berry, M. N., Mazzachi, R. D., Pejakovic, M., & Peake, M. J. (1989). Enzymatic determination of potassium in serum. *Clinical Chemistry*, *35*, 817–820.

Billman, G. F., Hughes, A. B., Dudell, G. G., Waldman, E., Adcock, L. M., Hall, D. M., et al. (2002). Clinical performance of an in-line, ex vivo point-of-care monitor: a multicenter study. *Clinical Chemistry*, *48*, 2030–2043.

Bisson, J., & Younker, J. (2006). Correcting arterial blood gases for temperature: (when) is it clinically significant? *Nursing in Critical Care*, *11*, 232–238.

Brydon, W. G., & Roberts, L. B. (1972). The effect of haemolysis on the determination of plasma constituents. *Clinica Chimica Acta*, *41*, 435–438.

CLSI (Clinical and Laboratory Standards Institute). (2009). *C46–A2. Blood gas and pH analysis and related measurements*. Wayne, PA: CLSI.

CLSI (Clinical and Laboratory Standards Institute). (2009). *C34–A3. Sweat testing: Sample collection and quantitative chloride analysis: Approved guideline* (3rd ed.). Wayne, PA: CLSI.

College of American Pathologists. (2017). *Chemistry/therapeutic drug monitoring C-A: CAP surveys*. Chicago: College of American Pathologists.

Farrell, P. M., White, T. B., Ren, C. L., Hempstead, S. E., Accurso, F., Derichs, N., et al. (2017). Diagnosis of cystic fibrosis: consensus guidelines from the cystic fibrosis foundation. *Journal of Pediatrics*, *181S*, S4–S15.

Graber, M., Subramani, K., Corish, D., & Schwab, A. (1988). Thrombocytosis elevates serum potassium. *American Journal of Kidney Diseases*, *12*, 116–120.

Grocott, M. P., Martin, D. S., Levett, D. Z., McMorrow, R., Windsor, J., Montgomery, H. E., et al. (2009). Arterial blood gases and oxygen content in climbers on Mount Everest. *New England Journal of Medicine*, *360*, 140–149.

Hill, C. E., Burd, E. M., Kraft, C. S., Ryan, E. L., Duncan, A., Winkler, A. M., et al. (2014). Laboratory test support for Ebola patients within a high-containment facility. *Laboratory Medicine*, *45*, e109–e111.

Lacher, D. A., Hughes, J. P., & Carroll, M. D. (2005). Estimate of biological variation of laboratory analytes based on the Third National Health and Nutrition Examination Survey. *Clinical Chemistry*, *51*, 450–452.

Mansour, M. M., Azzazy, H. M., & Kazmierczak, S. C. (2009). Correction factors for estimating potassium concentrations in samples with in vitro hemolysis: a detriment to patient safety. *Archives of Pathology & Laboratory Medicine*, *133*, 960–966.

Nijsten, M. W., de Smet, B. J., & Dofferhoff, A. S. (1991). Pseudohyperkalemia and platelet counts. *New England Journal of Medicine*, *325*, 1107.

Oliver, T. K., Jr., Young, G. A., Bates, G. D., & Adamo, J. S. (1966). Factitial hyperkalemia due to icing before analysis. *Pediatrics*, *1966*(38), 900–902.

Quiles, R., Fernandez-Romero, J. M., Fernandez, E., Luque de Castro, M. D., & Valcarcel, M. (1993). Automated enzymatic determination of sodium in serum. *Clinical Chemistry*, *39*, 500–503.

Zou, J., Nolan, D. K., LaFiore, A. R., & Scott, M. G. (2013). Estimating the effects of hemolysis on potassium and ldh laboratory results. *Clinica Chimica Acta*, *421*, 60–61.

25

Hormones

Timothy James Cole

OBJECTIVES

1. Define the following:
   - Amino acid–derived hormone
   - Autocrine system
   - Endocrine system
   - Endocrinology
   - Hormone
   - Hormone receptor
   - Incretin hormone
   - Paracrine system
   - Protein or polypeptide hormone
   - Steroid hormone
2. Describe the three classifications of hormones; include examples of each, chemical structures, characteristics, half-life, and tissues of origin.
3. List and describe three physiological functions of hormones; include the role of hormones in each category and give examples of hormones in each category.
4. Compare the two types of receptor-hormone interactions; include examples of each type, postreceptor actions, and the specific effects that each type of interaction produces in a cell.
5. Provide examples of feedback mechanisms in the endocrine system.
6. List and describe four techniques used to assess hormones; include basic principles, specimen requirements, and an advantage of each technique.

KEY WORDS AND DEFINITIONS

Autocrine A mode of hormone action in which a cell secretes a hormone that binds to autocrine receptors on that same cell, leading to changes in the cell.

Circadian rhythms Rhythmic repetition of certain phenomena in living organisms at about the same time each day.

Cyclic adenosine monophosphate (cAMP) A cyclic nucleotide that serves as an intracellular and, in some cases, extracellular "second messenger" mediating the action of many peptide or amine hormones.

Endocrine system The system of glands that release their secretions (hormones) directly into the circulatory system. In addition to the endocrine glands, included are the chromaffin and the neurosecretory systems.

Endocrinology The scientific study of the function and pathology of the endocrine glands.

Gonadotropin Any hormone that stimulates the gonads.

G-proteins Guanine nucleotide-binding proteins (G-proteins) are a family of intracellular proteins involved in transmitting chemical signals from outside the cell, and causing changes inside the cell. They communicate signals from many hormones, neurotransmitters, and other signaling factors.

G-protein–coupled receptors (GPCRs) A large superfamily of membrane receptors whose intracellular effects are mediated by G proteins.

Half-life In endocrinology, the time required for a hormone to fall to half its original concentration in the specified fluid or blood.

Homeostasis The maintenance of equilibrium of internal body functions in response to external changes.

Hormone A chemical substance that has a specific regulatory effect on the activity of a certain organ or organs on cell types.

Paracrine A type of hormone function in which hormone synthesized in and released from endocrine cells binds to its receptor in nearby cells of a different type and affects their function.

Phospholipase C-γ Any esterase that catalyzes the hydrolysis of the phosphoric ester bond of a membrane phospholipid, generating a phosphorylated alcohol and diacylglycerol.

Pituitary gland A small oval gland attached to the base of the vertebrate brain and consisting of an anterior and a posterior lobe (also known as the hypophysis and the pituitary body).

Receptor A molecular structure, normally a protein, within a cell or on the surface characterized by (1) selective binding of a specific substance and (2) a specific physiological effect that accompanies the binding; examples include (1) cell surface receptors for peptide hormones, neurotransmitters, antigens, complement fragments and immunoglobulins, and (2) cytoplasmic receptors for steroid and other lipid-derived hormones.

Second messenger Any of several classes of intracellular signaling molecules that translate electrical or chemical messages from the environment (first messengers) into cellular responses.

Zinc finger motif Found in nucleic acid–binding proteins that contain one or more zinc-binding domains (tandemly repeated, highly conserved stretches of 28 amino acids).

BACKGROUND

Hormones are chemical messengers synthesized and secreted by endocrine glands, organs, or isolated cells that have specific regulatory effects on the activity of target cells. As a component of the **endocrine system**, hormones are produced at one site in the body and, in general, exert their action(s) at distant sites. Some hormones exert actions locally (**paracrine**), and other hormones exert their action on their cell of origin, regulating their own synthesis and secretion (**autocrine**). Classic endocrine hormones include insulin, thyroxine, and cortisol; neurotransmitters and neurohormones are examples of paracrine hormones, and certain growth factors that stimulate synthesis and secretion of other hormones from the same cell are examples of autocrine hormones. Table 25.1 lists hormones that are commonly measured in the clinical laboratory. Hormones are classified as polypeptides or proteins, derivatives of amino acids, and steroids or other lipid derivatives.

Polypeptide or Protein Hormones

Adrenocorticotropic hormone (ACTH), insulin, and parathyroid hormone (PTH) are examples of polypeptide or protein hormones. They are water soluble and circulate freely in plasma as the secreted complete molecule or as active or inactive fragments. The **half-life** of these hormones in plasma is relatively short (≤10 to 30 minutes), and their concentrations may vary dramatically, depending on specific physiological and pathological states. These hormones initiate cellular responses by binding to a vast array of cell surface membrane **receptors** on target cells and activate intracellular signal-transduction pathways that lead to a specific action or changes within the target cell. For example, insulin, when released from pancreatic β-cells, binds to cell-surface insulin receptors at target tissues, such as the adipose and muscle, to promote the uptake of glucose from the bloodstream.

Amino Acid–Derived Hormones

Thyroid hormone such as thyroxine and the catecholamines are examples of hormones that are derived from amino acids; in both of these cases they are derived from tyrosine. They are water soluble and circulate in plasma bound to specific transport proteins (thyroxine) or circulate freely (catecholamines). Thyroxine binds avidly to three specific binding proteins—transthyretin, thyroid-binding globulin, and albumin—and has a half-life of approximately 7 to 10 days; free-circulating catecholamines such as epinephrine have a short half-life of a minute or less. The catecholamines interact with a family of nine closely related cell surface adrenergic receptors and initiate intracellular **second messenger** signal-transduction systems, whereas thyroid hormones move freely across the cell membrane to activate two specific intracellular nuclear receptors (NRs): thyroid hormone receptor-α and -β.

Steroid and Other Lipid-Derived Hormones

Endogenous steroid hormones such as cortisol, estrogen, and testosterone are synthesized from cholesterol, are hydrophobic, and are insoluble in water. These hormones circulate in plasma, reversibly bound to specific transport plasma proteins (e.g., cortisol-binding globulin, sex hormone-binding globulin) with only a small fraction unbound that is freely circulating to exert physiological action. The half-life of circulating steroid hormones is 30 to 90 minutes. Free steroid hormones at their sites of action, because of their hydrophobicity, enter the cell by passive diffusion and bind with intracellular NRs in the cytoplasm or the nucleus.

SYNTHESIS, RELEASE, AND GENERAL ACTION OF HORMONES

The physiological functions of hormones can be broadly categorized as those (1) affecting growth, development, and maturation; (2) controlling systemic homeostasis, energy balance, and integrated metabolism; and (3) regulating reproduction.

Synthesis and Release of Hormones

Many hormones are synthesized in specialized endocrine cells within specific endocrine glands. These include the polypeptide hormone insulin, which is synthesized in the β-cells of the pancreas; estradiol, which is synthesized in the granulosa cells of the ovary; and epinephrine (or adrenalin), which is synthesized by the medullary chromaffin cells of the adrenal gland. Some hormones require biosynthesis via complex enzymatic pathways, such as the cardiovascular hormone aldosterone, which is synthesized from precursor cholesterol in the outer zona glomerulosa cell layer of the adrenal cortex. Other hormones require simpler modification of a precursor compound such as vitamin D. All hormones attain a structural specificity that allows them to bind strongly to and activate specific receptors on target cells. In general, hormones are released from endocrine cells when they are required, either from intracellular stores or by induction

TABLE 25.1 **Frequently Measured Endocrine Hormones**

Endocrine Organ and Hormone	Hormone Type	Major Sites of Action	Principal Actions
Brain—Hypothalamus			
Thyrotropin-releasing hormone (TRH)	Peptide (3aa, Glu-His-Pro)[a]	Anterior pituitary	Release of thyroid-stimulating hormone (TSH) and prolactin (PRL)
Gonadotropin-releasing hormone (Gn-RH)	Peptide (10aa)	Anterior pituitary	Release of luteinizing hormone (LH) and follicle-stimulating hormone (FSH)
Corticotropin-releasing hormone (CRH)	Peptide (41aa)	Anterior pituitary	Release of adrenocorticotrophic hormone (ACTH) and β-lipotropic hormone (LPH)
Growth hormone-releasing hormone (GH-RH)	Peptides (40, 44aa)	Anterior pituitary	Release of growth hormone (GH)
Neuropeptide Y	36aa peptide released mainly from paraventricular neurons	Hypothalamus, thalamus, hippocampus, and gastrointestinal tract	Increased food intake, energy storage, reduced anxiety, and stress
Prolactin-releasing peptide	Peptide (20aa)	Anterior pituitary	Release of PRL
Kisspeptin	Peptide (54aa precursor cleaved to 13 and 14aa active peptides)	Gn-RH hypothalamic neurons	Stimulation of Gn-RH release via the G protein-coupled receptor GPR54
Anterior Pituitary Lobe			
Thyrotropin or TSH	Glycoprotein, heterodimer[b] (α, 92aa; β, 112aa)	Thyroid gland	Stimulation of thyroid hormone formation and secretion
Follicle-stimulating hormone (FSH)	Glycoprotein, heterodimer[b] (α, 92aa; β, 117aa)	Ovary	Growth of follicles with LH, secretion of estrogens, and ovulation
		Testis	Development of seminiferous tubules; spermatogenesis
Luteinizing hormone (LH)	Glycoprotein, heterodimer[b] (α, 92aa; β, 121aa)	Ovary Testis	Ovulation; formation of corpora lutea; secretion of progesterone Secretion of androgens
Prolactin (PRL)	Peptide (199aa)	Mammary gland	Proliferation of mammary gland; initiation of milk secretion; antagonist of insulin action
Growth hormone or somatotropin	Peptide (191aa)	Liver	Production of insulin-like growth factor (IGF-1) (promoting growth)
		Liver and peripheral tissues	Anti-insulin and anabolic effects
Corticotropin hormone or ACTH	Peptide (39aa)	Adrenal cortex	Stimulation of adrenocortical steroid biosynthesis and secretion
α-Melanocyte-stimulating hormone (α-MSH)	Peptide (13aa)	Skin	Dispersion of pigment granules, darkening of skin
Posterior Pituitary Lobe			
Vasopressin or antidiuretic hormone	Peptide (9aa)	Arterioles Renal tubules	Increase of blood pressure; water reabsorption
Oxytocin	Peptide (9aa)	Smooth muscles (uterus, mammary gland)	Contraction; action in parturition and in sperm transport; ejection of milk
Pineal Gland			
Melatonin	Indoleamine	Hypothalamus	Suppression of gonadotropin and GH secretion; induction of sleep
Thyroid Gland			
Thyroxine (T_4) and triiodothyronine (T_3)	Iodoamino acids	General body tissue	Stimulation of oxygen consumption and metabolic rate of tissue
Calcitonin or thyrocalcitonin	Peptide (32aa)	Skeleton	Uncertain in humans
Parathyroid Gland			
Parathyroid hormone (PTH) or parathyrin	Peptide (84aa)	Kidney	Increased calcium reabsorption, inhibited phosphate reabsorption; increased production of 1,25-dihydroxycholecalciferol
		Skeleton	Increased bone resorption

TABLE 25.1 Frequently Measured Endocrine Hormones—cont'd

Endocrine Organ and Hormone	Hormone Type	Major Sites of Action	Principal Actions
Adrenal Cortex			
Aldosterone	Steroid	Kidney	Salt and water balance
Cortisol	Steroid	Many	Metabolism of carbohydrates, proteins, and fats; anti-inflammatory effects; others
Dehydroepiandrosterone (DHEA) and dehydroepiandrostenedione sulfate (DHEAS)	Steroids	Hormone precursors	Converted to estrogens and testosterone
Adrenal Medulla			
Norepinephrine and epinephrine	Aromatic amines	Sympathetic receptors	Stimulation of sympathetic nervous system
Epinephrine	Aromatic amines	Liver and muscle, adipose tissue	Glycogenolysis Lipolysis
Ovary			
Activin A and B	Peptides[c] 2 β_A subunits 2 β_B subunits	Pituitary, ovarian follicle	Stimulates release of FSH; enhances FSH action; inhibits androgen production by theca cells
Estrogens	Phenolic steroids	Female accessory sex organs Bone	Development of secondary sex characteristics Control of skeletal maturation
Follistatin	Peptides (288aa, 315aa)	Pituitary, ovarian follicles	Inhibits FSH synthesis and secretion by binding activin
Inhibin A and B	Peptide (α subunit and β_A or β_B subunit)	Hypothalamus, ovarian follicle	Inhibits FSH secretion; stimulates theca cell androgen production
Progesterone	Steroid	Female accessory reproductive structure	Preparation of the uterus for ovum implantation, maintenance of pregnancy
Relaxin	Peptide[d]	Uterus	Inhibition of myometrial contraction
Testis			
Inhibin B	See above	Anterior pituitary, hypothalamus	Control of LH and FSH secretion
Testosterone	Steroid	Male accessory sex organs	Development of secondary sex characteristics, maturation, and normal function
Pancreas			
Amylin	Peptide (37aa)	Pancreas	Inhibits glucagon and insulin secretion
Glucagon	Peptide (29aa)	Liver	Glycogenolysis
Insulin	Peptide[e]	Liver, fat, muscle	Regulation of carbohydrate metabolism; lipogenesis
Gastrointestinal Tract			
Gastrin[f]	Peptide (17aa)	Stomach	Secretion of gastric acid, gastric mucosal growth
Ghrelin[f]	Peptide (28aa)	Anterior pituitary	Secretion of GH
Secretin	Peptide (27aa)	Pancreas	Secretion of pancreatic bicarbonate and digestive enzymes
GIP	Peptide (42aa)	Gastrointestinal tract	Inhibition of gastric secretion and motility; increase in insulin secretion
Kidney			
1,25-$(OH)_2$ cholecalciferol	Sterol	Intestine Bone	Facilitation of absorption of calcium and phosphorus; increase in bone resorption in conjunction with PTH
		Kidney	Increase in reabsorption of filtered calcium
Erythropoietin	Peptide (165aa)	Bone marrow	Stimulation of red cell formation

Continued

TABLE 25.1 Frequently Measured Endocrine Hormones—cont'd

Endocrine Organ and Hormone	Hormone Type	Major Sites of Action	Principal Actions
Renin-angiotensin-aldosterone system	Peptides (renin, 297aa; Ang I, 10aa; Ang II, 8aa, produced from Ang I by angiotensin-converting enzyme)	Renin (from kidney) catalyzes hydrolysis of angiotensinogen (from liver, 485aa) to Ang I in the intravascular space	Ang II increases blood pressure and stimulates secretion of aldosterone (see adrenal)
Liver			
IGF-1, formerly called somatomedin	Peptide (70aa)	Most cells	Stimulation of cellular and linear growth
IGF-2	Peptide (67aa)	Most cells	Insulin-like activity
Heart			
Atrial natriuretic peptide (atriopeptin)	Peptide with an intrachain disulfide bond (28aa)	Vascular, renal, and adrenal tissues	Regulation of blood volume and blood pressure
B-type natriuretic peptide	Peptide with an intrachain disulfide bond (32aa)	Vascular, renal, and adrenal tissues	Regulation of blood volume and blood pressure
Adipose Tissue			
Adiponectin	Peptide oligomers of 30 kDa subunits	Muscle Liver	Increases fatty acid oxidation Suppresses glucose formation
Leptin	Peptide (167aa)	Hypothalamus	Inhibition of appetite, stimulation of metabolism
Resistin	Peptide (94aa)	Liver	Insulin resistance
Tissue necrosis factor-α	157aa secreted protein	Adipose in an autocrine or paracrine fashion	Pluripotent cytokine with a wide-range of actions
Bone			
Osteocalcin	49aa	Bone, pancreas adipose tissue	Bone mineralization and calcium homeostasis, systemic metabolism
Osteopontin	314aa	Bone, pancreas, neutrophils	Bone remodeling, immune functions, chemotaxis
Osteonectin	285aa	Bone, paracrine action	Bone mineralization
Skeletal Muscle			
Myostatin	109aa monomer, active as a dimer	Skeletal muscle; paracrine actions	Inhibition of muscle differentiation and growth

A more complete list endocrine hormones can be found in Chapter 36: Hormones. In *Tietz textbook of clinical chemistry and molecular diagnostics* (6th ed.).
[a]*aa,* Amino acid residues.
[b]Glycoprotein hormones composed of two dissimilar peptides. The α-chains are similar in structure or identical; the β-chains differ among hormones and confer specificity.
[c]Each activin and inhibin is found in multiple forms.
[d]Two chains linked by two disulfide bonds: α, 24aa; β, 29aa.
[e]Two chains linked by two disulfide bonds: α, 21aa; β, 30aa.
[f]Also produced in the brain.

of rapid biosynthesis. The biosynthetic pathways of hormones are critical to maintain **homeostasis**. Defects in those pathway enzymes can lead to debilitating endocrine disease, such as the genetic condition of congenital adrenal hyperplasia, which is a disease caused by mutations in the steroid biosynthetic enzymes of the adrenal cortex, which are required for production of cortisol, aldosterone, and androgen precursors from cholesterol. The loss of end-product hormone production or the inappropriate buildup of steroid precursors can lead to complex cardiovascular and reproductive development dysfunction and gender ambiguity. Examples of the three broad functional categories of hormone action are presented as follows and illustrate specific aspects on the synthesis, release, turnover, and action of endocrine hormones.

Growth, Development, and Maturation

Normal embryonic development and postnatal growth of the whole human organism is dependent on the complex integrative function of many hormones, including growth hormone, insulin-like growth factor-1 (IGF-1), thyroxine, the gonadal steroids (estrogen and androgen), and cortisol. Several pituitary (also called the hypophysis) hormones, such as growth

hormone, ACTH, and thyroid-stimulating hormone (TSH), are responsible specifically for the growth and development of other endocrine glands, such as the adrenal, gonads, and thyroid. For example, growth hormone is synthesized and secreted by pituitary somatotroph cells and stimulates the production and release of hepatic IGF-1. Both factors then have broad somatic growth actions, particularly in bone and cartilage. Therefore these pituitary hormones are responsible for the control of synthesis and the secretion of many secondary hormones. These hormones provide important negative feedback control on the secretion of the pituitary hormones. Other regulators of secretion of the pituitary hormones include **circadian rhythms** and the hypothalamic pulse generator that controls the pulsatile secretion of **gonadotropins**.

Energy Homeostasis and Integrated Metabolism

Energy homeostasis and the activity of many metabolic pathways in cells are tightly regulated by a diverse number of systemic and locally released hormones. There is a continual change in demand for energy use and storage during states such as feeding, exercise, starvation, infection, injury and/or trauma, or emotional stress. These states initiate the release and action of many hormones to regulate energy use and nutrient metabolism, modulate energy storage, and regulate physical and behavioral responses to stress. These hormonal activated pathways in cells are complex and may involve hormones from different organs. In general, systemic nutrient uptake, appetite, energy storage, and release are under neurological control via actions of specific hypothalamic and neuroendocrine hormones (e.g., leptin and neuropeptide Y).

The following are examples illustrating the feedback control of hormone secretion and action, which is critical for homeostasis:

- Regulation of blood glucose concentrations: In response to a glucose load, insulin is promptly released from the pancreatic β-cells, which regulate the uptake of glucose into peripheral metabolic tissues (fat, muscle, liver, and brain) for the metabolism necessary to produce energy from glucose. As circulating glucose concentrations return to preload concentrations, insulin secretion slows. Several counterregulatory hormones come into play to further regulate this process to ensure that blood glucose concentrations do not become too low. These include glucagon, cortisol, epinephrine, and growth hormone.
- Water and electrolyte homeostasis is regulated by the steroid hormone aldosterone released from the adrenal gland, renin from the kidney, and vasopressin (antidiuretic hormone) from the posterior **pituitary gland**.

Endocrine Control of Reproduction

The gonadotropin hormones luteinizing hormone (LH) and follicle-stimulating hormone (FSH) are released from pituitary gonadotroph cells in response to hypothalamic gonadotropin-releasing hormone to regulate the development, growth, and functions of the ovary and testis. In turn, the ovarian and testicular hormones estradiol and testosterone, respectively, initiate and regulate the following: pubertal growth; development and maintenance of secondary sex characteristics; the growth, development, and maintenance of the skeleton and muscles; and in part, the distribution of body fat.

CELL SIGNALING MECHANISMS OF HORMONE RECEPTORS

The specific action of an endocrine hormone on its target cell or tissue is a function of the interaction between the hormone and its cognate receptor. Several types of hormone-receptor interactions occur with cells, and an individual cell may express and contain hundreds of different receptors that may respond to multiple specific hormones at any given time. The hormone-receptor complex provides the strong specificity of the hormone's action, eliciting a target tissue response. Many endocrine hormones circulate in picomolar or nanomolar concentrations (10^{-9} to 10^{-12} mol/L) but elicit specific and strong responses in specific target cells. Hormone receptors may be present on the cell surface imbedded in the plasma membrane or reside intracellularly within the cytoplasm or nucleus. Receptors at the cell surface and intracellular NRs initiate a variety of different postreceptor cellular responses by activating complex intracellular signal transduction pathways. This is shown schematically in Fig. 25.1 for three common classes of hormone receptors. Binding of the hormone or ligand to the receptor initiates activation of the receptor and initiation of downstream signaling that cause cellular changes in both the cytoplasm and the nucleus.

Cell Surface Hormone Receptors

Polypeptide or hydrophilic hormones bind to cell surface receptors, and the conformational change or the protein–protein interactions initiated resulting from this binding activates an effector system within the cells, which is responsible for the downstream actions of the hormone (see Fig. 25.1). For many peptide and polypeptide hormones, the intracellular effector that is activated by the hormone–receptor interaction is a specific intracellular **G-protein** (guanyl-nucleotide–binding protein), and the cell surface receptors are called **G-protein–coupled receptors (GPCRs)** (see Fig. 25.1). GPCRs are the largest family of cell surface receptors, with approximately 830 genes coding for human GPCRs. GPCRs are large proteins with seven membrane-spanning domains, an amino-terminal, a hormone-binding domain that is extracellular, and an intracellular carboxy-terminal tail of loop domains. The second major class of cell surface hormone receptors is the enzyme-coupled receptors (see Fig. 25.1). These receptors have been divided into six subfamilies of receptors, of which the most well-characterized family is the receptor tyrosine-kinase (RTK) group of receptors, which mediate the actions of many common hormones and growth factors, such as insulin, IGF-1, epidermal growth factors, and fibroblast growth factors. These receptors are characterized by initiating intracellular kinase cascades and regulating cell growth and metabolism, and are common targets for mutation in tumor cells.

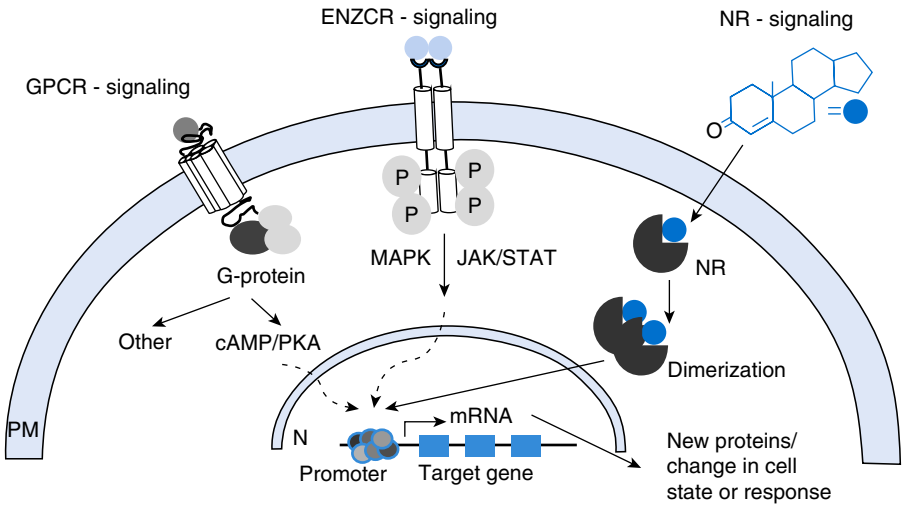


Fig. 25.1 Endocrine cell signaling is mediated by both cell surface and intracellular hormone receptors. Cell surface receptors comprise two main classes of proteins called G-protein–coupled receptors *(GPCR)* and enzyme-coupled receptors *(ENZCR)*, whereas intracellular hormone receptors form the nuclear receptor *(NR)* superfamily of proteins. Cell surface receptors bind a circulating hormone using a specific extracellular protein domain and transduce a signal to the interior of the cell. In contrast, receptors for lipophilic hormones (e.g., a steroid hormone) reside within the cell, in many cases in the cytoplasm in an inactive complex of proteins. The activated NR is then targeted to the nucleus. *cAMP*, Cyclic adenosine monophosphate; *JAK/STAT*, Jak kinase/STAT pathway; *MAPK*, MAP kinase pathway; *N*, nucleus; *P*, phosphate; *PKA*, protein kinase A; *PM*, plasma membrane.

SIGNAL TRANSDUCTION FROM CELL SURFACE G-PROTEIN–COUPLED RECEPTORS

GPCRs are grouped under four main classes, each of which contains receptors for specific types of hormones. The class A GPCRs contain the rhodopsin-like receptors that include receptors for olfactory compounds and melatonin; class B represents the secretin-like GPCRs (receptors for secretins, glucagon, and vasoactive intestinal polypeptide); class C, the largest by number, contains the metabotropic glutamate and/or pheromone receptors; and class D contains the fungal mating pheromone receptors. GPCRs are seven transmembrane proteins that on hormone binding undergo a conformational change allowing direct interaction of the intracellular C-terminal tail of the activated GPCR with specific G-proteins that are organized in a trimeric complex of an α, β, and γ G-protein subunit. Activated G-protein subunits then initiate signaling cascades via activation of specific enzymes or generation of small molecules that serve as potent second messengers to mediate the hormone response within the cell. The best known effector enzyme is adenylyl cyclase, which generates the important second messenger **cyclic adenosine monophosphate (cAMP)** (see Fig. 25.1). The activated G-protein α binds a molecule of GTP and is inactivated by an intrinsic GTPase activity that converts the bound GTP to GDP, thereby turning off the downstream signaling pathway.

SIGNAL TRANSDUCTION FROM CELL SURFACE ENZYME–COUPLED RECEPTORS

The insulin receptor is a member of the second major class of cell surface hormone receptors, called the enzyme-coupled receptors. These have been classified into six subfamilies of enzyme-coupled receptors, with the best characterized being the large family of RTKs, a family of 90 characterized receptors. These receptors mediate the actions of many common growth factor hormones such as the epidermal, fibroblast, platelet-derived, and vascular endothelial growth factors. RTKs are single transmembrane spanning proteins that dimerize upon hormone binding, an event that initiates intracellular signaling responses (see Fig. 25.1). The receptor dimer is then auto-phosphorylated by intrinsic C-terminal tyrosine kinase catalytic domains. The phosphorylated Tyr residues act as specific binding sites for effector signaling proteins, such as phosphoinositol-3-kinase (PI-3-kinase) and **phospholipase C-γ,** which then transmit the signal within the cell. These effector proteins contain SH2 domains that specifically recognize phosphotyrosine residues on the activated receptor. MAP–kinase pathway activation results in subsequent downstream target gene activation, as well as cell proliferation and differentiation. These pathways do not use second messengers but involve activated enzyme cascades. RTKs and downstream signaling components are often mutated in cancer cells that then acquire a high proliferative phenotype. Gene mutations often produce constitutively active receptors or signaling proteins such as RAS and PI-3-kinase.

Intracellular Nuclear Receptors and Signal Transduction

Lipid-soluble hormones (e.g., estradiol) are primarily transported bound to plasma proteins such as albumin, with only a small fraction of the hormone circulating free or unbound. Free hormone enters the cell via passive diffusion and in most cases binds to intracellular NRs in the cytoplasm, or in some

cases the nucleus (see Fig. 25.1). NRs are characterized by a hormone or ligand-binding domain, a central DNA-binding domain, and an amino-terminal variable domain that contributes to recruitment of nuclear proteins such as transcriptional coactivators. In general, lipid-soluble hormones bind and activate the ligand-binding domain of intracellular NRs that then signal specific changes to the nucleus (see Fig. 25.1). The human genome contains a group of 48 NR genes that encode the NR superfamily of receptors and are activated by a range of lipophilic steroid, dietary lipids, vitamins, fatty acid, and special small molecule ligands (e.g., thyroid hormone). A large number of these receptors are still classified as orphan receptors with no known hormone ligand, and little is understood of their functional role in human biology and in clinical disease. Hormone binding initiates a conformational change that enables the hormone-receptor complex to localize to the nucleus and bind specific regulatory DNA hormone response elements (HREs) near specific target genes. Hormone translocation of the cytoplasmic NRs to the nucleus, once activated, is mediated by a conformational change in the receptor that exposes a nuclear localization sequence normally located within the DNA-binding domain. The binding specificity of the activated NR for specific HREs near the target gene is determined by the first **zinc finger motif** in the receptor's DNA-binding domain. Binding of the hormone-receptor complex to HREs promotes recruitment of a co-activator and sometimes co-repressor proteins that ultimately mediate activation or repression of gene transcription. Messenger RNA levels of target genes can be enhanced up to 1000-fold, leading to large changes in the synthesis of specific proteins that then drive cellular and physiological change. In addition to these nuclear gene expression changes, many NRs can be activated to produce rapid so-called nongenomic effects in the cytoplasm or at the cell membrane.

CLINICAL DISORDERS OF HORMONE ACTION

Although several chapters of this textbook detail a variety of endocrine disorders, a brief introduction to endocrine disorders is given here. In general, endocrine diseases may result from a deficiency or an excess of a single hormone or several hormones, or from resistance to the action of hormones in target cells. Hormone deficiency can be congenital (genetically inherited) or acquired, and hormone excess can result from endogenous overproduction (from within the body) or from exogenous overmedication. Hormone resistance can occur at several levels but can most simply be characterized as receptor-mediated, post-receptor-mediated, or at the level of the target tissue. The clinical manifestations will depend on the hormone system affected and the type of abnormality. Diabetes is an example of an endocrine metabolic disorder and is the most common endocrine disorder in Western countries. It is classified as either type 1 or type 2. Type 1 diabetes results from failure of the pancreas to secrete insulin although the pancreas is otherwise normal. Type 2 diabetes results from end-organ resistance to the action of insulin, which, in this case, is secreted from the pancreas in abundant amounts and circulates at high concentrations. Secondary diabetes occurs when a nonendocrine disease, such as pancreatitis, destroys the pancreas, including the insulin-secreting cells. The biochemical hallmark of diabetes is hyperglycemia, which can be present when there is insulin deficiency or insulin excess, and insulin excess can accompany both hyperglycemia and hypoglycemia.

TECHNIQUES FOR MEASUREMENT OF HORMONES AND RELATED ANALYTES

Hormones are measured with a variety of analytical techniques, including bioassay, receptor assay, immunoassay, and instrumental techniques such as mass spectrometry interfaced with liquid or gas chromatography. Bioassays are based on observations of physiological responses specific for the hormone being measured. *In vivo* bioassays usually involve the injection of test materials (such as blood or urine from a patient) into suitably prepared animals; target gland or organ responses such as growth or steroidogenesis are then measured. *In vitro* bioassays involve the incubation of tissue samples, membranes, dispersed cells, or permanent cell lines in a defined culture medium, with subsequent measurement of an appropriate hormone response. Bioassays tend to be imprecise and are now rarely used in clinical medicine.

Receptor-based binding assays depend on the *in vitro* interaction of a hormone with its biological receptor. In this type of assay, unlabeled hormone displaces trace amounts of radioactively labeled hormone from specific receptor sites. In general, receptor-binding assays are simpler to perform and have greater sensitivity than bioassays, and also have an advantage over immunoassays in that they reflect the biological function of a hormone—namely, the capacity to combine with specific receptors. By contrast, immunoassays may measure active hormone and inactive prohormone, hormone polymer, and metabolites when all share the common antigenic determinant or set of determinants of the assay antibody. In general, receptor-binding assays are not as sensitive as immunoassays, and proteolytic enzymes in the biological specimen may sometimes degrade the receptor or destroy the labeled tracer.

Enzyme-linked immunosorbent assays or standard immunoassays that use specific antibodies are widely used to quantify hormones. Currently labeled antibody (immunometric) assays with nonisotopic labels are the method of choice for measuring concentrations of most hormones, especially peptides and proteins. Immunometric assays use saturating concentrations of two or more antibodies (often monoclonal) that are prepared against different epitopes of the protein molecule. One of the two antibodies is usually attached to a solid-phase separation system and extracts the hormone from the serum specimen. The second antibody is linked to a signal molecule, which is then measured (see Chapter 15). Mass spectrometry coupled with gas and liquid chromatographs are powerful qualitative and quantitative analytical tools that are widely used to measure hormone concentrations. Technical advancements in mass spectrometry have resulted in the development

of matrix-assisted laser desorption and/or ionization and electrospray ionization techniques that allow the sequencing of peptides and mass determination of picomole quantities of analytes. Tandem mass spectrometry offers greater analytical sensitivity, accuracy, and speed, and may allow simultaneous determination of multiple hormones related to a clinical condition (see Chapter 13).

Specimen Requirements

To prevent preanalytical variation, appropriate protocols must be followed when sending samples to the laboratory for hormone measurement. Some hormones are directly affected by food (e.g., insulin) or by circadian variability (e.g., cortisol; see Chapter 6). In many clinical circumstances, the metabolic environment plays a crucial role in hormone production, and it is essential to obtain a simultaneous sample for measurement of both the hormone and the molecule(s) regulated by that hormone. An isolated measurement of plasma insulin without concurrent knowledge of the plasma glucose, or measurement of PTH independent of serum calcium, is of little, if any, value. When a patient is evaluated for possible hormone deficiency or hormone excess, it is often necessary to perform a stimulation or suppression test, such as the Synacthen or ACTH test that is used to assess adrenal insufficiency. Most hormone assays can be performed on plasma or serum, and many can be performed on urine samples, usually a 24-hour collection. Increasingly, saliva has become a convenient body fluid for hormone analysis, particularly for hormones secreted in a diurnal rhythm (e.g., cortisol). Unlike blood sampling, which requires the patient to present to a blood drawing facility, patients can be provided with salivary collection material so they can provide specimens to the laboratory collected at multiple times during the day.

POINTS TO REMEMBER 1

- Hormones provide communication and regulation of activities between cells and tissues of the body.
- Hormones are a diverse group of lipid-based, amino acid–derived and polypeptide molecules.
- Hormones circulate in body fluids at variable levels and are predictive of health and disease.

POINTS TO REMEMBER 2

- Hormones bind and activate specific protein receptors at target cells.
- Hormone receptors are located both at the plasma membrane and in the cytoplasm of the cells.
- Hormones trigger complex intracellular signaling cascades to bring about cellular changes.

REVIEW QUESTIONS

1. Steroid hormones, by binding specific nuclear receptors, have a direct effect on:
   a. enzyme phosphorylation.
   b. gene expression.
   c. cAMP formation.
   d. DNA replication.
2. When a protein hormone binds a cell membrane receptor, the classical result is:
   a. phosphorylation of intracellular enzymes.
   b. inhibition of mRNA within the cell nucleus.
   c. replication of DNA and division of the cell.
   d. glycosylation of intracellular proteins.
3. An example of an amino acid–related hormone would be:
   a. testosterone.
   b. thyroid hormone.
   c. growth hormone.
   d. prostaglandin.
4. Which of the following hormones is classified as a polypeptide?
   a. Cortisol
   b. Estrogen
   c. Adrenocorticotropic hormone
   d. Thyroid hormone
5. The type of hormone assay that best reflects hormone functionality as opposed to quantification of the hormone itself is the:
   a. receptor-based assay.
   b. mass spectrometry–based assays.
   c. bioassay technique.
   d. liquid chromatography.
6. A hormone released from the pancreas in response to a glucose load that is involved in cellular energy production is considered to be involved in:
   a. modification of the response to stress.
   b. synthesis of milk proteins.
   c. promotion of bone development and growth
   d. homeostatic control of metabolism.
7. Which of the following is a characteristic of steroid hormones?
   a. Account for the majority of hormones
   b. Circulate freely
   c. Water soluble
   d. Long half-life in the circulation

SUGGESTED READINGS

Brivanlou, A. H., & Darnell, J. E., Jr. (2002). Signal transduction and the control of gene expression. *Science*, *295*, 813–818.

Farfel, Z., Bourne, H. R., & Iiri, T. (1999). The expanding spectrum of G protein disease. *The New England Journal of Medicine*, *340*, 1012–1020.

Evans, R. M. (1988). The steroid and thyroid hormone receptor superfamily. *Science*, *240*, 889–895.

Hubbard, S. R., & Miller, W. T. (2007). Receptor tyrosine kinases: mechanisms of activation and signaling. *Current Opinion in Cell Biology*, *19*, 117–123.

Isberg, V., Mordalski, S., Munk, C., Rataj, K., Harpsøe, K., Hauser, A. S., et al. (2016). GPCRdb: an information system for G protein-coupled receptors. *Nucleic Acids Research*, *44*, D356–D364.

Katritch, V., Cherezov, V., & Stevens, R. C. (2013). Structure-function of the G protein-coupled receptor superfamily. *Annual Review of Pharmacology and Toxicology*, *53*, 531–556.

Kelly, D. P. (2010). Introduction to the nuclear receptor review series. *Circulation Research*, *106*, 1557–1558.

Lefkowitz, R. J., Rockman, H. A., & Koch, W. J. (2000). Catecholamines, cardiac beta-adrenergic receptors, and heart failure. *Circulation*, *101*, 1634–1637.

Lemmon, M. A., & Schlessinger, J. (2010). Cell signaling by receptor tyrosine kinases. *Cell*, *141*, 1117–1134.

McKenna, N. J., & O'Malley, B. W. (2002). Minireview: nuclear receptor coactivators: an update. *Endocrinology*, *143*, 2461–2465.

New, M. I., Lekarev, O., & Mancenido, D. (2013). Congenital adrenal hyperplasia owing to 21-hydroxylase deficiency. In *Genetic Steroid Disorders* (pp. 29–51). London: Elsevier.

Norman, A. W., & Litwack, G. (1997). *Hormones* (2nd ed.). San Diego: Academic Press.

Oldham, W. M., & Hamm, H. E. (2008). Heterotrimeric G protein activation by G-protein coupled receptors. *Nature Reviews Molecular Cell Biology*, *9*, 60–71.

Pawson, T. (2004). Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems. *Cell*, *116*, 191–203.

Qi, M., & Elion, E. A. (2005). MAP kinase pathways. *Journal of Cell Science*, *118*, 3569–3572.

Shaw, R. J., Cantley, L. C. Ras, PI(3)K and mTOR signaling controls tumour cell growth. *Nature*, 44:424–430.

Yen, F. M., Ando, S., & Feng, X. (2006). Thyroid hormone action at the cellular, genomic and target gene levels. *Molecular and Cellular Endocrinology*, *246*, 121–127.

26

Catecholamines and Serotonin

Graeme Eisenhofer, Thomas Rosano, Ronald J. Whitley

OBJECTIVES

1. Define the following terms:
   - 5-HIAA
   - Carcinoid
   - Carcinoid syndrome
   - Catecholamine
   - Gastroenteropancreatic neuroendocrine tumor
   - Homovanillic acid
   - Metanephrine
   - Neuroblastoma
   - Normetanephrine
   - Paraganglioma
   - Pheochromocytoma
   - Serotonin
2. Discuss the following biogenic amines, including function and clinical relevance, synthesis, metabolism, and storage and release:
   - Norepinephrine
   - Epinephrine
   - Dopamine
   - Serotonin
   - Melatonin
3. List the clinically relevant metabolites of the biogenic amines as detailed in Objective 2 above.
4. State the clinical significance and function of the catecholamines and serotonin in each of the following systems:
   - Central nervous system
   - Sympathetic nervous system
   - Adrenal medullary system
   - Peripheral dopaminergic system
   - Enteric nervous system
5. Compare pheochromocytoma and neuroblastoma, including causes, symptoms, organ(s) and/or cell type involved, and laboratory evaluation of each; describe the clonidine suppression test and the results obtained when pheochromocytoma is present.
6. Explain the laboratory values obtained for serotonin and 5-HIAA in the assessment of a gastroenteropancreatic tumor and the carcinoid syndrome.
7. Describe the collection techniques used to obtain specimens for catecholamine, metanephrine, and serotonin analysis.
8. Describe how diet interferes with serotonin analysis; state how tricyclic antidepressants interfere in the measurement of catecholamines.
9. List the laboratory methods used to analyze metanephrines, catecholamines, serotonin, and 5-HIAA.
10. Analyze and solve case studies related to catecholamine and serotonin disorders using descriptions of symptoms and results of laboratory analyses.

KEY WORDS AND DEFINITIONS

Carcinoid syndrome A syndrome due to carcinoid tumors and characterized by attacks of severe cyanotic flushing of the skin—lasting from minutes to days—and by diarrheal watery stools, bronchoconstrictive attacks, sudden drops in blood pressure, edema, and ascites. Symptoms are caused by secretion from the tumor of serotonin, prostaglandins, and other biologically active substances.

Carcinoid tumor A yellow circumscribed tumor arising from enterochromaffin cells, usually in the small intestine, appendix, stomach, or colon, and less commonly in the bronchus; sometimes used alone to refer to the gastrointestinal tumor (also called argentaffinoma).

Catecholamine One of a group of biogenic amines having a sympathomimetic action; the aromatic portion of the molecule is the catechol, and the aliphatic portion is an amine; examples include dopamine, norepinephrine, and epinephrine.

Catecholamine metabolites Products of catecholamine metabolism, such as dihydroxyphenylacetic acid, methoxytyramine, homovanillic acid, dihydroxyphenylglycol, methoxyhydroxyphenylglycol, normetanephrine, metanephrine, and vanillylmandelic acid.

Chromaffin cell Neuroendocrine cell derived from embryonic neural crest found in the medulla of the adrenal gland and in other ganglia of the sympathetic nervous system; so-named because of the presence of cytoplasmic granules that give a brownish reaction with chromium salts.

3,4-Dihydroxyphenylglycol (DHPG) The metabolite produced within the peripheral sympathetic or central nervous system noradrenergic nerves by deamination of norepinephrine (can also be formed from epinephrine); O-methylated to methoxyhydroxyphenylglycol in extraneuronal tissues.

L-Dopa An amino acid—3,4-dihydroxyphenylalanine—produced by oxidation of tyrosine by tyrosine hydroxylase; the precursor of dopamine and an intermediate product in the biosynthesis of norepinephrine, epinephrine. Can also be formed in melanocytes by the actions of tyrosinase as part of the production of melanin, the pigment in skin.

Dopamine A catecholamine formed in the body by the decarboxylation of dopa; an intermediate product in the synthesis of norepinephrine; acts as a neurotransmitter in the central nervous system and is produced peripherally in large amounts with autocrine/paracrine actions.

Epinephrine (adrenaline) A catecholamine hormone secreted by the adrenal medulla.

Gastroenteropancreatic neuroendocrine tumor (GEP-NET) Encompasses neuroendocrine tumors of the digestive system and pancreas but also covers neuroendocrine pulmonary tumors and includes carcinoids.

Homovanillic acid (HVA) A product of dopamine metabolism; elevated urinary concentrations are used to diagnose neuroblastoma.

5-Hydroxyindoleacetic acid (5-HIAA) A metabolite of serotonin (5-hydroxytryptamine) excreted in large amounts by patients with carcinoid tumors.

Metanephrine A catecholamine metabolite resulting from O-methylation of epinephrine; formed mainly within adrenal chromaffin cells; excreted mainly in the urine as a sulfate-conjugated metabolite; measurements of the free and conjugated metabolite with normetanephrine provide useful tests for the diagnosis of pheochromocytoma.

3-Methoxy-4-hydroxyphenylglycol (MHPG) A metabolite of epinephrine and norepinephrine formed primarily from the O-methylation of dihydroxyphenylglycol and in smaller amounts from the deamination of normetanephrine and metanephrine; found in brain, blood, cerebrospinal fluid, and urine, where its concentrations can be used to measure catecholamine turnover.

Neuroblastoma A type of neural crest-derived tumor usually arising in the autonomic nervous system (sympathicoblastoma) or in the adrenal medulla that affects mostly infants and children up to 10 years of age.

Neuroendocrine tumors (NET) Neoplasms that arise from cells of the endocrine and peripheral nervous systems, most commonly in the digestive tract, but also found in the lung, pancreas, pituitary, thyroid, and other tissues. The tumors include gastroenteropancreatic neuroendocrine tumors, as well as paragangliomas, neuroblastomas, neuroectodermal tumors, medulloblastomas, retinoblastomas, and pineoblastomas. They usually contain secretory granules and often produce biogenic amines and polypeptides.

Norepinephrine (noradrenaline) A major neurotransmitter produced by some brain neurons and peripheral sympathetic nerves, which acts on α- and β_1-adrenergic receptors; produced in the adrenal chromaffin cells as a precursor for epinephrine.

Normetanephrine An O-methylated metabolite of norepinephrine produced in extraneuronal cells and the adrenal medulla; excreted in the urine largely as a sulfate-conjugated metabolite; measurements of free and conjugated metabolites provide useful tests for diagnosis of pheochromocytoma.

Pheochromocytoma A usually benign, well-encapsulated, lobular, vascular tumor of chromaffin tissue of the adrenal medulla or sympathetic paraganglia.

Serotonin (5-hydroxytryptamine) A monoamine vasoconstrictor synthesized in the intestinal enterochromaffin cells or in central or peripheral neurons; found in high concentrations in many body tissues, including intestinal mucosa, pineal body, and central nervous system.

Vanillylmandelic acid (VMA) The main end-product of norepinephrine and epinephrine metabolism excreted in the urine; formed primarily in the liver from the oxidation of methoxyhydroxyphenylglycol.

Catecholamines and serotonin are biogenic amines that facilitate neuronal or hormonal signaling for a wide range of physiological processes. The naturally occurring catecholamines, **dopamine** and **norepinephrine (noradrenaline)**, function as neuromodulators in the brain. Norepinephrine also functions as a neuromodulator in peripheral sympathetic nerves that innervate tissues throughout the body. **Epinephrine (adrenaline)** functions as a hormone released by the adrenal medulla. Catecholamines are critical in maintaining the body's homeostasis and in responding to acute and chronic stress through the orchestration of cardiovascular, metabolic, glandular, and visceral organ activities. **Serotonin (5-hydroxytryptamine)** serves as a neuromodulator in the brain and as a modulator of vascular and gastrointestinal functions in the periphery. Abnormal production of catecholamines or serotonin occurs in a number of neuroendocrine tumors, where clinical signs and symptoms reflect the pharmacological properties of the secreted amines. Clinical measurement of the biogenic amines or of their metabolites aids in tumor detection and monitoring, and analytical advances have produced sensitive and specific laboratory methods for clinical practice.

CHEMISTRY, BIOSYNTHESIS, RELEASE, AND METABOLISM

Chemistry

The **catecholamines**—dopamine, norepinephrine, and epinephrine—are phenylethylamines with hydroxyl groups on positions three and four of the benzene ring and an ethylamine sidechain on position one (Fig. 26.1). Hydroxyl and methyl substitutions on the ethylamine sidechain distinguish the individual catecholamines in terms of both structure and function. The catecholamines demonstrate varying degrees of alkaline instability in biological fluids, and their dihydroxybenzene or catechol structure is sensitive to oxidative formation of quinones in the presence of air and light. Serotonin with its indoleamine structure is distinct from the catecholamines but is an important naturally occurring biogenic amine. Melatonin is the principal indoleamine produced from serotonin by the pineal gland.

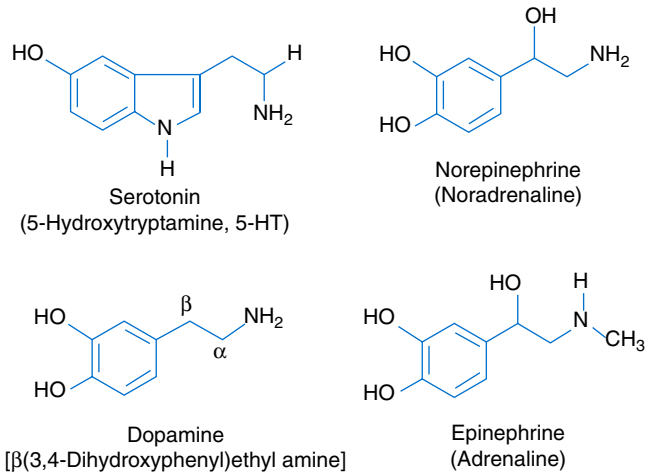


Fig. 26.1 Chemical structures of the catecholamines and serotonin.

Biosynthesis

The rate-limiting step in catecholamine biosynthesis involves conversion of tyrosine to 3,4-dihydroxyphenylalanine (**L-dopa**) by the enzyme, tyrosine hydroxylase (Fig. 26.2). A related enzyme, tryptophan hydroxylase, catalyzes conversion of tryptophan to 5-hydroxytryptophan (5-HTP) in the first step of serotonin synthesis. Tissue sources of catecholamines are principally dependent on the presence of tyrosine hydroxylase, which is largely confined to dopaminergic and noradrenergic neurons of the central nervous system and to the sympathetic and adrenal medullary systems in peripheral tissues Similarly, sources of serotonin are largely dependent on the presence of tryptophan hydroxylase in central nervous system serotonergic neurons, the

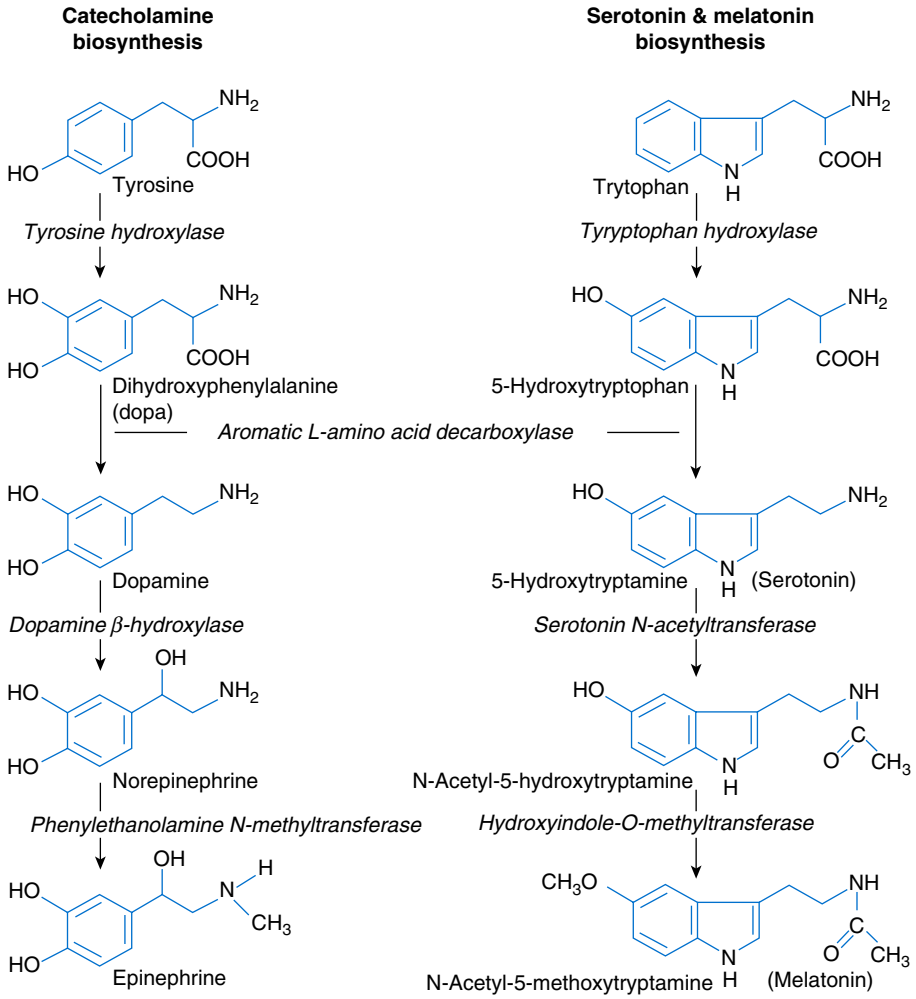


Fig. 26.2 Biosynthesis of catecholamines and serotonin, and metabolism of serotonin to melatonin.

pineal gland, and some peripheral endocrine tissue, particularly enterochromaffin cells of the digestive tract. Platelets also contain large amounts of serotonin derived from serotonin synthesized in enterochromaffin cells of the gastrointestinal tract.

Conversion of L-dopa to dopamine and 5-HTP to serotonin is catalyzed by aromatic-L–amino acid decarboxylase (see Fig. 26.2), an enzyme with a wide tissue distribution and broad substrate specificity for aromatic amino acids. The dopamine and serotonin formed in the cytoplasm by aromatic-L–amino acid decarboxylase are transported into vesicular secretory granules, where the amines are concentrated and stored ready for exocytotic release as the principal neuromodulators of central nervous system dopaminergic and serotonergic neurons. The dopamine formed in noradrenergic neurons and chromaffin cells is further converted to norepinephrine by dopamine β-hydroxylase, an enzyme specifically located in secretory granules. The additional presence of phenylethanolamine *N*-methyltransferase in adrenal medullary chromaffin cells leads to further conversion of norepinephrine to epinephrine.

Melatonin is synthesized from serotonin in the pineal gland by two highly specific enzymes; the first step is catalyzed by serotonin-*N*-acetyltransferase and the second by hydroxyindole-*O*-methyltransferase.

Storage and Release

Monoamines, stored in secretory granules, exist in a highly dynamic equilibrium with the surrounding cytoplasm, with passive outward leakage of monoamines into the cytoplasm counterbalanced by inward active transport under the control of vesicular monoamine transporters (Fig. 26.3). Monoamines share the acid environment of the secretory granule matrix with ATP, peptides, and proteins, the best known of which are the chromogranins. The chromogranins are common components of monoamine-containing secretory granules. Their widespread presence among endocrine tissues has led to their measurement in plasma as useful, albeit relatively nonspecific markers of neuroendocrine tumors.

Monoamines are released from secretory vesicles into the extracellular space through the process of exocytosis. Neuronal release may also occur by calcium-independent nonexocytotic processes involving an increased loss of monoamines from storage vesicles into the cytoplasm and the reversal of normal inward carrier-mediated transport to outward transport of monoamines into the extracellular environment. Examples of this process include the release of catecholamines induced by tyramine and amphetamine.

Uptake and Metabolism

Because the enzymes responsible for metabolism of catecholamines have intracellular locations, the primary mechanism limiting the duration of action of catecholamines in the extracellular space is uptake by active transport (see Fig. 26.3). Uptake is facilitated by neuronal or extraneuronal transporters. Neuronal monoamine transporters provide the principal mechanism for rapid termination of the signal in neuronal transmission, whereas transporters at extraneuronal locations are more important for limiting the spread of the signal and for clearing catecholamines from the bloodstream. Of the norepinephrine released by sympathetic nerves, about 90% is removed back into nerves by neuronal uptake, 5% is removed by extraneuronal uptake, and 5% escapes these processes to enter the bloodstream. In contrast, of the epinephrine released directly into the bloodstream from the adrenals, about 90% is removed by extraneuronal monoamine uptake.

In addition to terminating the actions of released monoamines, plasma membrane monoamine transporters function in sequence with vesicular monoamine transporters to recycle catecholamines for rerelease (see Fig. 26.3). Thus, most of the norepinephrine released and recaptured by sympathetic nerves is sequestered back into secretory granules by vesicular monoamine transporters, thereby reducing the requirements for synthesis of new transmitter. Plasma membrane monoamine transporters also function as part of metabolizing systems, requiring the additional actions of enzymes for irreversible inactivation of the released amines.

Catecholamines undergo metabolism by multiple pathways (Fig. 26.4). Most metabolism occurs within the same cells in which catecholamines are synthesized and is dependent on the leakage of amines from secretory granules into the cytoplasm. In sympathetic nerves, the presence of monoamine oxidase (MAO) leads to conversion of norepinephrine to **3,4-dihydroxyphenylglycol (DHPG)**. This is then largely

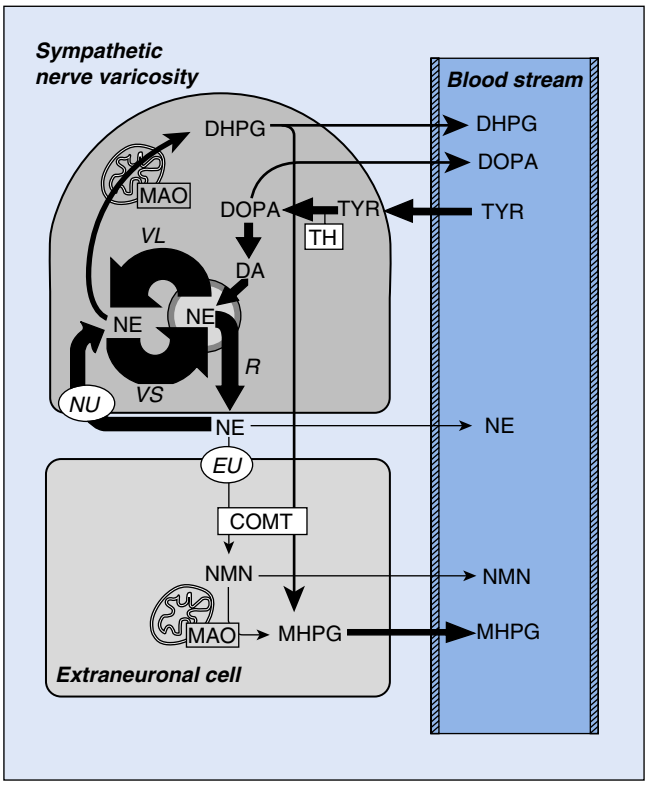


Fig. 26.3 Schematic diagram illustrating the dynamics of synthesis, exocytotic release *(R)*, neuronal reuptake *(NU)*, extraneuronal uptake *(EU)*, vesicular leakage *(VL)*, vesicular sequestration *(VS)*, and metabolism of norepinephrine *(NE)* in sympathetic nerve endings in relation to extraneuronal tissue and the bloodstream. Relative magnitudes of the various processes are reflected by the relative sizes of arrows. *COMT*, Catechol-*O*-methyltransferase; *DA*, dopamine; *DHPG*, 3,4-dihydroxyphenylglycol; *L-dopa*, 3,4-dihydroxyphenylalanine; *MAO*, monoamine oxidase; *MHPG*, 3-methoxy-4-hydroxyphenylglycol; *NMN*, normetanephrine; *TH*, tyrosine hydroxylase; *TYR*, tyrosine.

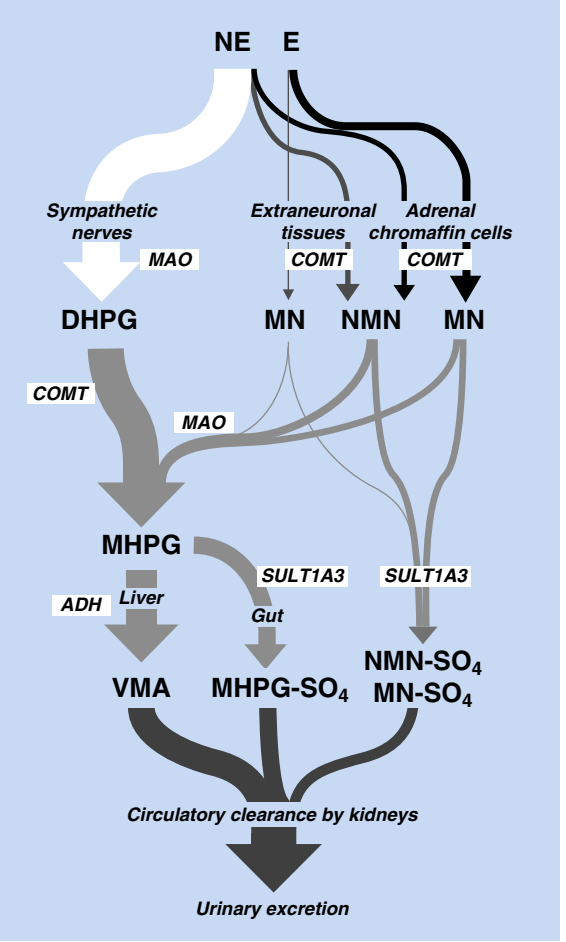


Fig. 26.4 Schematic diagram showing the main pathways for metabolism of the norepinephrine and epinephrine derived from sympathoneuronal or adrenal medullary sources. Deamination in sympathetic nerves *(white)* is the major pathway of catecholamine metabolism and involves intraneuronal deamination of norepinephrine leaking from storage granules, or of norepinephrine recaptured after release by sympathetic nerves. Metabolism in adrenal chromaffin cells *(black)* involves *O*-methylation of catecholamines leaking from storage granules into the cytoplasm of adrenal medullary cells. The extraneuronal pathway *(gray)* is a relatively minor pathway of metabolism of catecholamines released from sympathetic nerves or the adrenal medulla, but it is important for further processing of metabolites produced in sympathetic nerves and adrenal chromaffin cells. The free metanephrines produced in extraneuronal tissues or adrenal chromaffin cells are further metabolized by deamination or sulfate conjugation. *ADH,* Alcohol dehydrogenase; *COMT,* catechol-*O*-methyltransferase; *DHPG,* 3,4-dihydroxyphenylglycol; *E,* epinephrine; *MAO,* monoamine oxidase; *MHPG,* 3-methoxy-4-hydroxyphenylglycol; *MHPG-SO$_4$,* 3-methoxy-4-hydroxyphenylglycol sulfate; *MN,* metanephrine; *MN-SO$_4$,* metanephrine-sulfate; *NE,* norepinephrine; *NMN,* normetanephrine; *NMN-SO$_4$,* normetanephrine-sulfate; *SULT1A3,* phenolsulfotransferase type 1A3; *VMA,* vanillylmandelic acid.

metabolized by catechol-*O*-methyltransferase (COMT) in extraneuronal tissues to **3-methoxy-4-hydroxyphenylglycol (MHPG)**. **Vanillylmandelic acid (VMA)**, the primary end-product of norepinephrine and epinephrine metabolism, is produced almost exclusively in the liver. This process is dependent on the hepatic localization of alcohol dehydrogenase, an enzyme required for the conversion of MHPG to VMA.

In adrenal chromaffin cells, the additional presence of COMT leads to the metabolism of norepinephrine to **normetanephrine** and of epinephrine to **metanephrine** (see Fig. 26.4). Because the intraneuronal deamination pathway far predominates over the extraneuronal *O*-methylation pathway, normetanephrine and metanephrine represent relatively minor products of catecholamine metabolism. As a consequence, the adrenal medulla represents the single largest tissue source of normetanephrine and metanephrine, accounting for 24% to 40% of the former and more than 90% of the latter. The metanephrine and normetanephrine produced in the adrenal medulla or in extraneuronal tissues—the former from catecholamines leaking from secretory granules and the latter from catecholamines released from sympathoadrenal sources—may be deaminated to MHPG and then converted to VMA in the liver or may be sulfate-conjugated by a sulfotransferase enzyme expressed mainly in the gastrointestinal tissues.

Serotonin is not a substrate for COMT; it follows simpler pathways of metabolism than those for catecholamines. Serotonin is deaminated to **5-hydroxyindoleacetic acid (5-HIAA)**, the major urinary excretion product of serotonin metabolism.

PHYSIOLOGY OF CATECHOLAMINE AND SEROTONIN SYSTEMS

Catecholamines and serotonin regulate physiological events at the cellular level through interaction with α- and β- adrenergic receptors. Dopamine transmits signals primarily by interaction with a large family of dopamine receptors. The major physiological effects of catecholamines and serotonin are explained in part by the diverse receptor interactions that occur in function-specific locations throughout the vasculature and organ systems of the body.

Central Nervous System

Norepinephrine, dopamine, and serotonin are produced primarily in regions of the brainstem by neurons with projections to other areas of the brain or spinal cord. The norepinephrine-producing neurons of the brainstem participate in regulating the activity of the sympathetic nervous system and the overall state of attention and vigilance.

Dopamine produced in dopaminergic neurons has functions, and shows distributions that are notably different from those of norepinephrine. Dopamine in the brain influences reward-seeking behavior and is important for the initiation and maintenance of movement. Disturbances in dopamine production and release in the brain are therefore involved in drug-dependency states and are central to the movement disorder that characterizes Parkinson disease. Dopamine neurotransmission is also involved in processing sensory signals and in regulating hormonal release. Dopaminergic neurons in the retina and olfactory bulb have ultrashort projections that transmit signals within these neuronal centers for vision and smell. Dopamine neurons in the hypothalamus have regulatory influences on the release of several hormones of the pituitary gland.

Serotonin, like norepinephrine, is produced by small clusters of neurons in the brainstem regions. It serves a diverse range of behavioral and physiological functions including memory, learning, feeding behavior, sleep patterns, thermoregulation, pain modulation, cardiovascular function, and hypothalamic regulation of pituitary hormones.

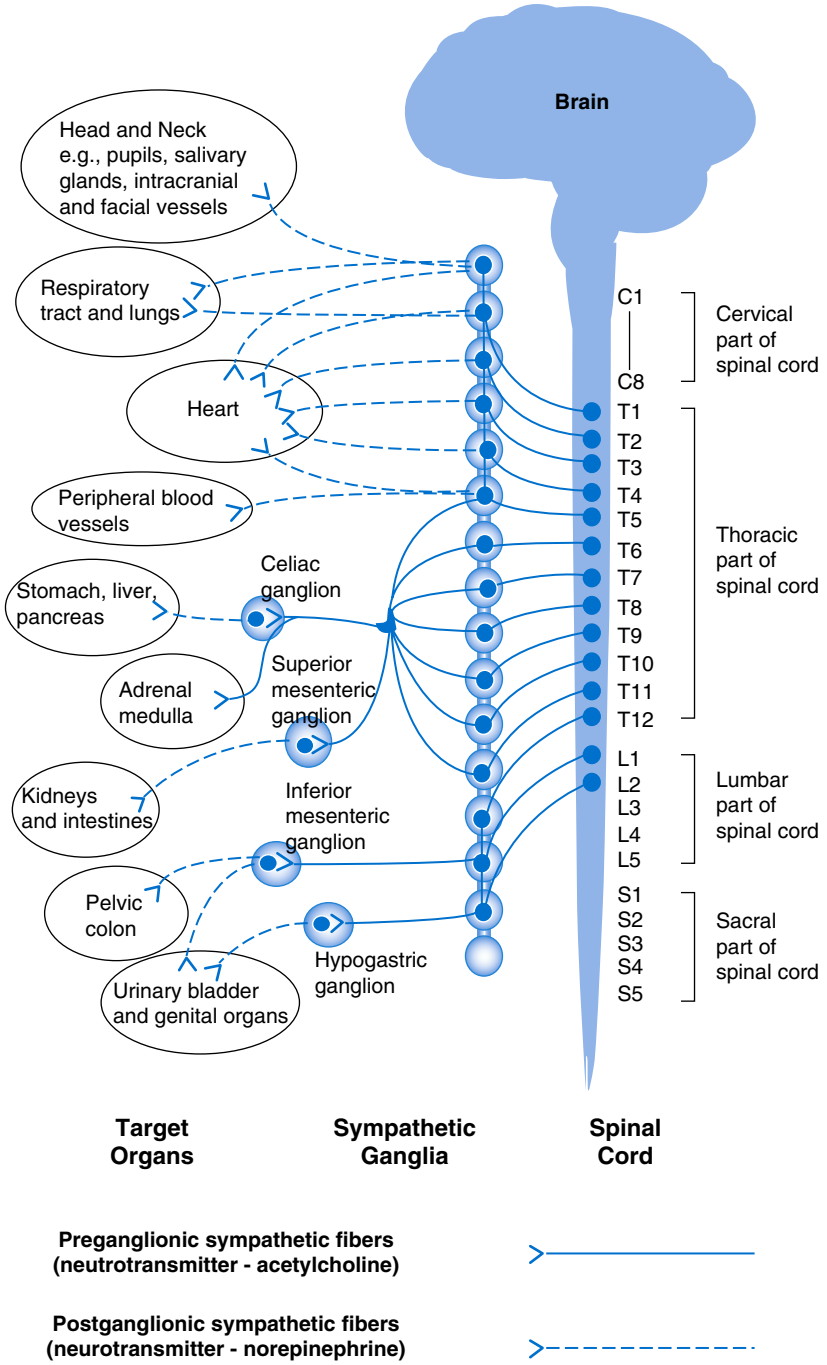


Fig. 26.5 Schematic diagram of sympathetic division of the autonomic nervous system. Preganglionic cholinergic fibers *(solid lines)* from the spinal cord project to the paravertebral sympathetic chain, visceral peripheral ganglia, and the adrenal medulla, whereas postganglionic noradrenergic fibers *(dashed lines)* project from sympathetic ganglia to sympathetically innervated target organs. About 5% to 10% of the norepinephrine released from sympathetic nerves innervating target organs escapes neuronal and extraneuronal uptake to enter the bloodstream. This contrasts with epinephrine, which functions as a circulating hormone and is released directly from the adrenal medulla into the bloodstream. Most of the catecholamines entering the bloodstream are removed by extraneuronal uptake processes and are metabolized (e.g., in the liver), so that only small proportions are excreted in the urine.

Sympathetic Nervous System

Sympathetic nerve transmission operates below the level of consciousness in controlling the physiological function of many organs and tissues of the body (Fig. 26.5). The sympathetic nervous system plays a particularly important role in regulating cardiovascular function in response to postural, exertional, thermal, and mental stress. With sympathetic activation, heart rate is increased, peripheral arterioles are constricted, skeletal arterioles are dilated, and blood pressure is elevated. Sympathetic signals work in balance with the parasympathetic portion of the autonomic nervous system to maintain a stable internal environment.

Efferent or outgoing signals from the central nervous system are transmitted by preganglionic cholinergic sympathetic neurons that exit the spinal cord and converge on sympathetic ganglia along the spinal column or in visceral ganglia (see Fig. 26.5). The terminal branches of the postganglionic fibers that project from these ganglia into target organs have varicosities that form a rich ground plexus for synaptic contact with a large number of effector cells in target tissues. Most sympathetic postganglionic nerves liberate norepinephrine as their neurotransmitter. In limited locations, sympathetic nerve endings release acetylcholine. Sympathetic fibers that release acetylcholine innervate sweat glands and the adrenal medulla, the latter stimulating the release of epinephrine from chromaffin cells.

Plasma and urinary norepinephrine are derived primarily from postganglionic sympathetic neurons with little contribution from the central nervous system or from hormonal release by the adrenal medulla. Because of intervening neuronal and extraneuronal removal processes, the amount of norepinephrine that escapes into the bloodstream represents less than 10% of the total norepinephrine released by sympathetic nerves. Alterations in release and plasma concentrations of norepinephrine occur in response to stimuli such as exercise, overeating, low salt intake, upright position, mental stress all of which increase sympathetic outflow. Increased plasma concentrations of norepinephrine also occur with aging and are found in disorders such as cardiac failure, hypertension, and depression.

Adrenal Medullary System

The human adrenal glands overlie the superior poles of the kidneys. Each gland consists of an outer cortex and a thin inner central medulla containing chromaffin cells. A characteristic feature of adrenal medullary chromaffin cells is the presence of numerous catecholamine storage granules. These granules turn brown when exposed to potassium bichromate solutions, ammoniacal silver nitrate, or osmium tetroxide because of the oxidation and polymerization of epinephrine and norepinephrine. This process is known as the *chromaffin reaction,* hence the terms *chromaffin cells* and *chromaffin granules.*

Although often considered a part of the sympathetic nervous system, the adrenal medulla produces and secretes epinephrine with different functions from the norepinephrine secreted by sympathetic nerves. The adrenal medulla and the sympathetic nerves are regulated separately, often in divergent directions in response to different forms of stress. Epinephrine is secreted from adrenal chromaffin cells directly into the bloodstream to act on cells distant from sites of release. Epinephrine and norepinephrine have overlapping but different potencies of effect on α- and β-adrenoceptors. The proximity of sites of norepinephrine and epinephrine release to adrenoceptors determines differences in adrenoceptor-mediated responses to the two catecholamines. Because of the above factors, epinephrine exerts its effects on different populations of adrenoceptors than those affected by norepinephrine. Epinephrine released from the adrenal glands is more important as a metabolic hormone than as a hemodynamic regulatory hormone. In particular, epinephrine stimulates lipolysis, ketogenesis, thermogenesis, and glycolysis, and raises plasma glucose by stimulating glycogenolysis and gluconeogenesis. Epinephrine also has potent effects on pulmonary function, causing dilation of airways.

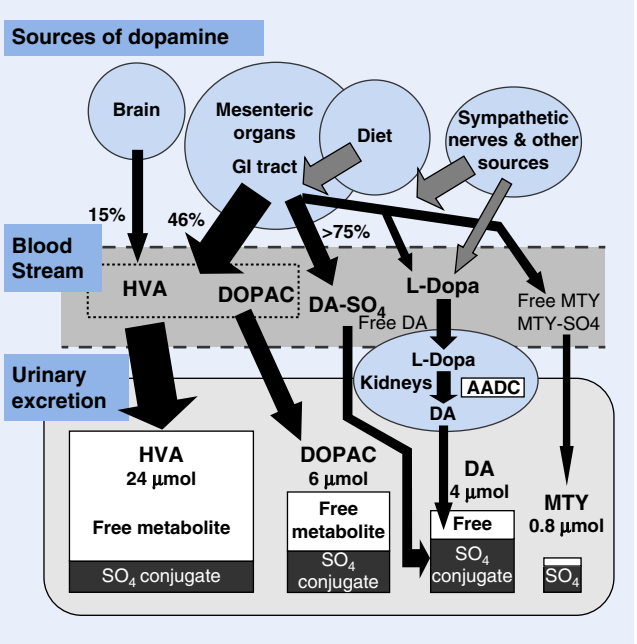


Fig. 26.6 Schematic representation of the main sources of dopamine and the principal metabolites of dopamine in plasma and urine. The brain makes a relatively minor contribution, whereas dopamine synthesized in the gastrointestinal tract or derived from the diet contributes substantially to dopamine metabolites in the bloodstream and urine. This contrasts with the free dopamine excreted in urine, which is derived almost entirely from renal extraction of circulating L-dihydroxyphenylalanine and local decarboxylation to dopamine by L-aromatic amino acid decarboxylase. *AADC*, Aromatic acid decarboxylase; *DA*, dopamine; *DA-SO$_4$*, dopamine-sulfate; *DOPAC*, dihydroxyphenylacetic acid; *GI*, gastrointestinal; *HVA*, homovanillic acid; *L-dopa*, L-dihydroxyphenylalanine; *MTY*; 3-methoxytyramine; *MTY-SO$_4$*, 3-methoxytyramine-sulfate.

Peripheral Dopaminergic System

Dopamine is usually thought of as a neuromodulator in the brain or as an intermediate in the production of norepinephrine and epinephrine in peripheral tissues. However, the contribution of the brain to plasma concentrations and urinary excretion of dopamine metabolites is relatively minor, and dopamine produced in the sympathetic nerves and the adrenal medulla is mainly converted to norepinephrine. Most circulating and urinary dopamine and dopamine metabolites are therefore derived from other sources (Fig. 26.6).

In the kidneys, dopamine functions as an autocrine or paracrine effector substance contributing to the regulation of sodium excretion. Unlike neuronal catecholamine systems, the production of dopamine in the kidneys is largely independent of local synthesis of L-dopa by tyrosine hydroxylase. Production of dopamine in the kidneys depends mainly on uptake of L-dopa from the circulation and conversion to dopamine by aromatic amino acid decarboxylase (see Fig. 26.6). Thus, most of the free dopamine excreted in urine derives from renal uptake and decarboxylation of circulating L-dopa.

Although the kidneys represent the major source of urinary free dopamine, this source does not account for the much larger quantities of excreted dopamine metabolites, such as *homovanillic acid (HVA)* and dopamine sulfate. Substantial proportions of these metabolites are produced in the gastrointestinal tract, where dopamine appears to function as an enteric neuromodulator or as a paracrine or autocrine substance. However, unlike in the kidneys, where dopamine is produced mainly from circulating L-dopa, in the gastrointestinal tract, production of dopamine requires the presence of tyrosine hydroxylase or other sources of L-dopa.

Consumption of food increases plasma concentrations of L-dopa, dopamine, and dopamine metabolites, particularly dopamine sulfate, indicating that dietary constituents may represent an important source of peripheral dopamine (see Fig. 26.6). It is now clear that dopamine sulfate is produced mainly in the gastrointestinal tract from both dietary and locally synthesized dopamine. This is consistent with findings that the gastrointestinal tract contains high concentrations of the specific sulfotransferase enzyme responsible for sulfate conjugation of catecholamines and **catecholamine metabolites**.

Enteric Nervous System

The enteric nervous system (ENS) is an independent and integrated system of neurons and supporting cells located in the gastrointestinal tract, gallbladder, and pancreas. The ENS is composed of the myenteric plexus and the submucous plexus. Both are embedded in the wall of the gut and extend from the esophagus to the anus, and contain more than 100 million sensory neurons, interneurons, and motor neurons. This network senses the environment within the lumen, regulates local blood flow, and controls epithelial cell secretion.

The ENS is connected to the central nervous system by extrinsic parasympathetic and sympathetic motor neurons, and by spinal and vagal sensory neurons. Through these bidirectional connections, the ENS is monitored and modified. Despite the presence of these extrinsic nerve connections, the ENS functions autonomously in some intestinal regions. Neural transmission within the ENS is controlled by a large variety of neuromodulatory amines and peptides, such as serotonin, norepinephrine, acetylcholine, ATP, and nitric oxide. Serotonin acts as a local paracrine molecule, participating in mucosal sensory transduction. More than 95% of the body's serotonin is produced within the gastrointestinal tract, and most is synthesized and stored in enterochromaffin cells in the gut mucosa. Intrinsic sensory neurons activated by serotonin stimulate peristaltic reflex and secretion, whereas extrinsic sensory neurons initiate bowel sensations such as nausea, vomiting, abdominal pain, and bloating. The paracrine actions of serotonin are terminated by uptake into epithelial cells by the same serotonin transporter used in serotonergic neurons.

Clinical Applications

Measurements of catecholamines, serotonin, and their metabolites are commonly used in studies of physiological function and disease processes, but clinical laboratory measurements of the amines and their metabolites are directed primarily at the diagnosis of neuroendocrine tumors. Catecholamine-producing neuroendocrine tumors include pheochromocytomas, paragangliomas, and neuroblastomas, whereas serotonin-producing tumors are typically carcinoids.

Pheochromocytoma and Paraganglioma

Catecholamine-producing tumors that derive from chromaffin cells of the adrenal glands are referred to as **pheochromocytomas**, whereas extraadrenal tumors are termed *paragangliomas* or *extraadrenal pheochromocytomas*. Pheochromocytomas are rare, with an annual detection rate of 2 to 5 per million. Autopsy studies, indicating the prevalence of 0.05% to 0.1%, suggest that most of the tumors remain undetected and contribute to premature death.

The presence of a pheochromocytoma is usually suspected because of signs and symptoms that reflect the biological effects of catecholamines released by the tumor. Hypertension, the most common sign, can be sustained or paroxysmal. Symptoms include headache, palpitations, diaphoresis (excessive sweating), pallor, nausea, attacks of anxiety, and generalized weakness. Among patients tested because of such signs and symptoms, the pretest prevalence of tumors is low—usually less than 1%.

Patients with a higher risk for pheochromocytoma include those with a hereditary predisposition to the tumor, a previous history of the tumor, or the incidental finding of an adrenal mass during routine abdominal imaging. See the box "Points to Remember: Pheochromocytoma and Paraganglioma" for a summary of the key points of pheochromocytomas.

POINTS TO REMEMBER: PHEOCHROMOCYTOMA AND PARAGANGLIOMA

Biology

- Pheochromocytomas are tumors of mainly adrenal chromaffin cells that produce, store, metabolize, and secrete catecholamines, usually with a predominance of norepinephrine over epinephrine.
- Paragangliomas, developing from extra-adrenal sympathetic chromaffin tissue, usually produce near exclusively norepinephrine and very rarely predominantly dopamine.

Presentation

- Tumor is usually suspected because of signs and symptoms of catecholamine excess (e.g., hypertension, palpitations, headaches, excessive sweatiness) or the incidental finding of an adrenal mass during routine imaging procedures for unrelated medical conditions.
- Most pheochromocytomas are benign; 15% of adrenal tumors and 35% of extra-adrenal tumors are malignant.
- At least 30% of pheochromocytomas have a hereditary basis, resulting from mutations of at least 15 genes identified to date.

Continued

POINTS TO REMEMBER: PHEOCHROMOCYTOMA AND PARAGANGLIOMA—cont'd

- Less than 1% of patients tested because of signs and symptoms have the tumor (low pretest prevalence), but the prevalence is higher among patients with identified mutations of disease-causing genes or those with an incidental adrenal mass; thus testing in these patients is recommended independently of the presence of signs and symptoms.

Biochemical Diagnosis

- Biochemical diagnosis is based on evidence of excess production of catecholamines, which is best determined from measurements of metanephrines in urine or plasma.
- Because secretion of catecholamines by tumors is episodic, but metabolism to metanephrines within tumors is continuous, measurements of plasma or urinary fractionated metanephrines provide more reliable diagnostic tests for the tumor than measurements of catecholamines (metanephrines increased in >97% and catecholamines in 69% to 92% of patients with pheochromocytomas).

The biochemical diagnosis of pheochromocytoma focuses on measurements of metanephrines rather than their precursor catecholamines. Normally, most norepinephrine is metabolized by deamination within sympathetic nerves, so that *O*-methylation represents a minor pathway of catecholamine metabolism. In patients with pheochromocytoma, intratumoral *O*-methylation becomes a dominant pathway of catecholamine metabolism. Consequently, the presence of the tumor leads to relatively large increases in the production of *O*-methylated metabolites, compared with minor increases in deaminated metabolites. The metanephrines are produced continuously as a result of ongoing leakage of catecholamines from chromaffin granule stores into the cell cytoplasm; in contrast, catecholamines are often released episodically or in relatively low quantities compared with metanephrines.

The particularly high diagnostic sensitivity of plasma metanephrines has now been confirmed by numerous independent studies, leading to guidelines that recommend measurements of either plasma or urinary fractionated metanephrines during the initial work-up of a patient with a suspected pheochromocytoma. Negative results of these tests virtually exclude a pheochromocytoma. Exceptions include small (<1 cm) tumors encountered during routine screening and tumors that do not synthesize norepinephrine and epinephrine. Among the latter, tumors that produce only dopamine may be detected by measurements of plasma methoxytyramine, the *O*-methylated metabolite of dopamine. Measurements of plasma dopamine are also useful, whereas urinary dopamine and methoxytyramine provide relatively insensitive and nonspecific tests for the detection of a dopamine-producing tumor (see Fig. 26.6).

Plasma or urinary fractionated metanephrines are usually elevated sufficiently to conclusively establish the presence of most cases of pheochromocytoma. However, in about 20% of patients with the tumor, increases are smaller. Such true-positive results are difficult to distinguish from the larger proportion of false-positive results due to inappropriate sampling conditions (e.g., blood sampling without a preceding 30-minute period of supine rest), medications such as tricyclic antidepressants, or clinical conditions that increase sympathetic outflow. Additional biochemical testing is then necessary.

When biochemical testing continues to yield equivocal results, the clonidine suppression test may be useful for further confirming or excluding a pheochromocytoma. Originally, this test was designed to distinguish increased plasma catecholamines due to pheochromocytoma or sympathetic activation. By activating α_2-adrenoceptors in the brain and on sympathetic nerve endings, clonidine suppresses norepinephrine release by sympathetic nerves. Decreases in elevated plasma norepinephrine after clonidine therefore suggest sympathetic activation, whereas a lack of decrease suggests a pheochromocytoma. The test has subsequently been shown to be useful for distinguishing increases in plasma normetanephrine due to a pheochromocytoma from those due to sympathetic activation.

Most pheochromocytomas are sporadic, but as many as 30% to 40% have a hereditary basis. Hereditary forms of the tumor result from mutations of at least 15 disease-causing genes identified to date. Pheochromocytomas occur at a relatively low prevalence in about 2% of patients with neurofibromatosis type 1 (NF1). Risk for most of the other known hereditary conditions is higher. Pheochromocytomas in multiple endocrine neoplasia type 2 (MEN 2) result from mutations of the *ret* proto-oncogene, which also predisposes to medullary thyroid cancer. In von Hippel-Lindau syndrome (VHL), family-specific mutations of the VHL tumor suppressor gene determine the varied clinical presentation of tumors, which, in addition to pheochromocytomas and paragangliomas, include retinal angiomas, central nervous system hemangioblastomas, and tumors in the kidneys, pancreas, and testis. Familial paragangliomas and pheochromocytomas may occur secondary to mutations of genes for four subunits of succinate dehydrogenase (*SDHA, SDHB, SDHC,* and *SDHD*). Mutations of genes encoding the SDH complex assembly factor 2 *(SDHAF2),* transmembrane protein 127 *(TMEM127),* MYC-associated factor X *(MAX)* fumarate hydratase and malate dehydrogenase 2 represent other tumor-susceptibility genes.

Mutations of the different tumor susceptibility genes have been found to give rise to distinct catecholamine metabolite profiles (Fig. 26.7). Tumors in patients with MEN 2 and NF1 are almost always confined to the adrenals and are characterized by increases in plasma-free metanephrine, indicating epinephrine production. In contrast, tumors in patients with *VHL, SDHD,* and *SDHB* mutations, including those at adrenal locations, are characterized by increases in plasma or urinary normetanephrine without increases in metanephrine. Increases in methoxytyramine, the metabolite of dopamine, are additionally present in up to 70% of patients with mutations of *SDHD* and *SDHB* genes. In some patients, methoxytyramine represents the only metabolite increased.

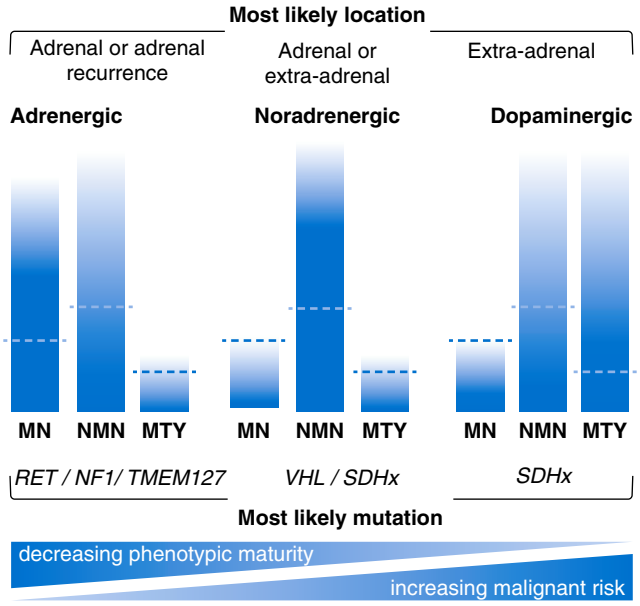


Fig. 26.7 Catecholamine biochemical phenotypes among patients with mutations of *RET, NF1, TMEM127, VHL* and *SDHB, SDHD, SDHC* and *SDHA (SDHX)* genes as reflected by patterns of increases in plasma normetanephrine *(NMN)*, metanephrine *(MN)* and methoxytyramine *(MTY)* and in relation to tumor location, phenotypic maturity, and malignant risk. (From Eisenhofer, G., & Peitzsch, M. [2014]. Laboratory evaluation of pheochromocytoma and paraganglioma. *Clinical Chemistry, 60*, 1486–1499. Reproduced with permission from the American Association for Clinical Chemistry.)

The distinct mutation-dependent biochemical profiles can be useful for stratifying patients for genetic testing. For mutations that confer a high risk of disease, such screening is generally recommended at yearly intervals and should include biochemical testing for evidence of excess catecholamine production, individualized according to the expected biochemical profile of the tumor.

Although most often benign, about 10% of pheochromocytomas and 35% of paragangliomas eventually metastasize. The risk of malignancy is particularly high in patients with *SDHB* mutations. Tumors in such patients typically produce low quantities of catecholamines and are much larger at diagnosis than tumors in other patients. This fact, along with their typically extra-adrenal location, explains their higher predisposition to malignancy. Measurements of plasma methoxytyramine appear to serve as a promising biomarker for identifying metastatic pheochromocytoma. Diagnosis of malignant pheochromocytoma on the basis of histopathological features is not possible and instead requires evidence of metastatic lesions in liver, lungs, bones, lymph nodes, or other sites where chromaffin cells normally do occur. All patients with a previous history of tumor are at risk of recurrent or metastatic disease and should undergo periodic tumor screening.

Neuroblastoma

Neuroblastomas are neoplasms that derive from primordial neural crest cells of the sympathetic nervous system. Neuroblastomas are almost exclusively a pediatric cancer, accounting for approximately 7% of all childhood neoplasms; they are the most common malignancies in the first year of life. The incidence of neuroblastoma is approximately 10 cases per million children. The vast majority of neuroblastomas develop sporadically. Activating mutations in the tyrosine kinase domain of the anaplastic lymphoma kinase oncogene account for most cases of hereditary neuroblastoma and may occur as somatic mutations in 5% to 15% of sporadic cases. Mutations of paired-like homeobox 2b *(PHOX2B)*, also known as *neuroblastoma Phox (NBPhox)*, account for other cases of hereditary neuroblastoma.

Most neuroblastomas are intra-abdominal, arising in the adrenal gland or in the upper abdomen. Less frequent locations include chest, neck, and pelvis. About 60% are extra-adrenal. Metastases in disseminated neuroblastomas may involve bone marrow, bone, lymph nodes, liver, and, less frequently, skin, testis, and intracranial structures.

The biological behavior of a neuroblastoma varies from regression or maturation to an aggressive course with an unfavorable outcome. Neuroblastomas are most notable for a subset of cases with complete regression or maturation to ganglioneuroma, a benign neoplasm. However, most clinically diagnosed tumors are aggressive and have an unfavorable outcome. This highly variable clinical outcome underlies the need for diagnostic markers to stratify patients for therapeutic intervention. The patient's age at diagnosis serves as a primary criterion for stratification; infants younger than 18 months have a more favorable outcome than older children. Unfortunately, the overall incidence of metastatic neuroblastoma at the time of diagnosis is approximately 60%, and the need for earlier detection of children with the progressive disseminating tumor remains a diagnostic challenge. See the box "Points to Remember: Neuroblastoma" for a summary of the key points of neuroblastomas.

POINTS TO REMEMBER: NEUROBLASTOMA

Biology

- Tumors that occur almost exclusively in children, which develop from primitive neural crest cells of the sympathoadrenal system, with about 60% derived from extra-adrenal sympathetic tissue and 40% from the adrenals.
- Biological behavior of tumors is highly variable, with some tumors spontaneously regressing but most (60%) having an aggressive course with disseminated malignant disease and an unfavorable outcome.

Presentation

- Neuroblastomas produce variable amounts of dopamine and norepinephrine but show a poor capacity for storage and release of catecholamines. Consequently, the tumors rarely produce signs and symptoms of catecholamine excess.
- Suspicion of disease is usually based on palpation of a mass, space-occupying complications of the tumor, or effects of bone marrow involvement.
- Hereditary causes of neuroblastomas are rare; most neuroblastomas occur sporadically.

Continued

POINTS TO REMEMBER: NEUROBLASTOMA—cont'd

Biochemical Diagnosis

- Biochemical diagnosis is based on overproduction of dopamine and norepinephrine, but because the tumors have poorly developed machinery for catecholamine storage and release, diagnosis depends mainly on measurements of catecholamine metabolites.
- Measurements of urinary vanillylmandelic acid and homovanillic acid represent the most widely used tests, but in a large prospective screening study they were shown to detect only 73% of tumors.

Hypertension and signs and symptoms of catecholamine excess are uncommon in neuroblastoma due to intracellular metabolism and release of mainly inactive metabolites. Patients commonly present with a tumor mass and clinical signs from compression effects on neighboring structures or hematological abnormalities from bone marrow involvement.

Laboratory evidence of a functional catecholamine-producing tumor is important in the clinical evaluation when a neuroblastoma is suspected. Neuroblastoma tumor cells have the capacity to synthesize dopamine and norepinephrine, but lack phenylethanolamine *N*-methyltransferase and consequently do not produce epinephrine. Because of the variability in catecholamine production and metabolism, no single reliable marker of catecholamine overproduction has been identified. Combinations of catecholamines and metabolites are therefore often measured in the diagnostic evaluation of neuroblastoma.

HVA and VMA, produced from dopamine and norepinephrine, respectively, are most widely used for the diagnosis of neuroblastoma. Results are expressed as the ratio of catecholamine metabolites to creatinine excretion in random urine specimens with a reported clinical sensitivity of around 90%. However, diagnostic sensitivity of these measurements in screening programs has been particularly limited (73%) and appears to miss patients with aggressive disease diagnosed at a later age. Consequently, screening programs have been generally unsuccessful in reducing the rate of metastatic neuroblastoma.

Additional markers, such as plasma dopamine and L-dopa, may have clinical value and allow the alternate use of plasma. The measurement of *O*-methylated metabolites, such as normetanephrine and methoxytyramine, may offer additional methods for the diagnosis of neuroblastoma.

Some evidence suggests that the pattern of catecholamine metabolism is associated with important biological and genetic prognostic factors in neuroblastoma. Immature metabolic patterns in neuroblastoma tumor tissue are based on higher excretion of dopamine or HVA relative to norepinephrine or VMA. The immature metabolic pattern is suggested to be associated with aggressive tumor behavior and other unfavorable prognostic factors, but the clinical application of such metabolic patterns has not been established.

Finally, clinical specificity in detecting neuroblastoma with catecholamine and metabolite measurements may be influenced by the choice of laboratory methods, dietary interference, and other catecholamine overproduction conditions. Significant advances in the analytical specificity of laboratory methods have reduced many of the exogenous sources of interference. Histopathology remains the ultimate diagnostic criterion for diagnosis and distinguishing neuroblastoma from pheochromocytoma or other catecholamine-producing neurogenic tumors, such as ganglioneuroma and ganglioneuroblastoma.

Gastroenteropancreatic Neuroendocrine and Carcinoid Tumors

Gastroenteropancreatic neuroendocrine tumors (GEP-NET), including **carcinoid tumors**, are derived from a variety of neuroendocrine cell types, in particular, enterochromaffin cells. These tumors are widely distributed in the body and are highly heterogeneous in their clinical presentation. The usual carcinoid tumor is solid and yellow-tan in appearance. Tumor cells exhibit a monotonous morphology, with pink granular cytoplasm and round nuclei with infrequent mitoses. Most of these neuroendocrine tumors are recognized by their reactions to silver stains and to neuroendocrine cell markers, such as chromogranin and neuron-specific enolase.

The tumors are traditionally classified according to their presumed origin from the embryonic foregut (bronchus, lung, stomach, duodenum, and pancreas), midgut (ileum, jejunum, appendix, and proximal colon), or hindgut (rectum and distal colon). The most common sites for these tumors are bronchus or lung—33%, ileum or jejunum—20%, rectum—10%, and appendix—8%. The annual detection rate of clinically significant GEP-NET has been estimated at 2.5 to 5 cases per 100,000 persons but has been steadily increasing, presumably as a result of improved recognition. These tumors may develop in all age groups but appear most frequently in adults; a mean age of 63 has been reported for tumors of the small intestine and respiratory tract. Clinically, most patients are asymptomatic until metastases are present. Bowel obstruction and abdominal pain are the most frequent presenting symptoms.

GEP-NET tumors show aggressive malignant behavior depending on the origin, the depth of penetration, and the size of the primary tumor. Most rectal carcinoids are found incidentally at endoscopy. They are often smaller than 1 cm and have a low rate of metastasis, even though they may show extensive local spread. Carcinoids of the appendix are seen in about 1 in every 300 appendectomies. Almost all measure less than 1 cm, and distant metastasis is rare. By contrast, 90% of intestinal neuroendocrine tumors that penetrate halfway through the muscle wall will have spread to lymph nodes and distant sites at the time of diagnosis. More than 70% of these tumors, 1 to 2 cm in diameter, metastasize to the liver. Fortunately, most grow slowly, and patients may live for many years. See the box "Points to Remember: Gastroenteropancreatic and Carcinoid Tumors" for a summary of the key points of carcinoid and GEP-NET.

POINTS TO REMEMBER: GASTROENTEROPANCREATIC AND CARCINOID TUMORS

Biology

- Neuroendocrine tumors derived from enterochromaffin cells of the gastrointestinal and respiratory tracts.
- Usually develop as small tumors (<2 cm) with slow growth but with the propensity to metastasize.
- Synthesize, store, and release a variety of peptide hormones and biogenic amines, including serotonin.

Presentation

- Bowel obstruction and abdominal pain.
- Symptoms related to secretion of vasoactive amines and peptides resulting in carcinoid syndrome (flushing, diarrhea, bronchoconstriction, and right-sided heart failure); relatively uncommon and usually occurring only after the development of metastases.

Biochemical Diagnosis

- Biochemical diagnosis of carcinoids depends mainly on measurements of serotonin, serotonin metabolites (5-hydroxyindoleacetic acid), and the serotonin precursor (5-hydroxytryptophan) in urine, plasma, whole blood, and platelets.
- False-positive results are a common problem resulting from dietary influences.

As with normal gut endocrine cells, GEP-NET synthesize, store, and release a variety of hormones and biogenic amines. One of the best characterized of these substances in **carcinoid syndrome** is serotonin. These tumors also produce and secrete other biologically active substances, including histamine, kallikrein, bradykinins, tachykinins, prostaglandins, dopamine, norepinephrine, and peptide hormones. Production of these substances varies in relation to the tissue origin of the tumor. Midgut neuroendocrine tumors release large quantities of serotonin, whereas tumors derived from the foregut secrete primarily 5-HTP (a serotonin precursor) and histamine, rather than serotonin. Primary hindgut NET usually show no secretory activity. Pancreatic NET are capable of producing gastrin, insulin, glucagon, and vasoactive intestinal peptide (VIP).

Secretion of vasoactive substances into the systemic circulation plays an important role in the development of carcinoid syndrome. The fully developed syndrome associated with humoral manifestations of these tumors is striking but uncommon, usually occurring only after metastasis to the liver and the release of these substances directly into the systemic circulation. The classic clinical presentation of the carcinoid syndrome includes pronounced flushing (especially on the face and neck), diarrhea, and bronchoconstriction, and eventual right-sided valvular heart failure. Overproduction of serotonin is found in 90% to 100% of patients with the carcinoid syndrome and is thought to be responsible for diarrhea through its known effects on gut motility and fluid secretion.

Clinical laboratory evaluation of the carcinoid syndrome relies on measurements of serotonin and its metabolites in body fluids and tissue. In patients with the typical carcinoid syndrome, 5-HTP is converted to serotonin and is stored in tumor secretory granules and in platelets. A small amount of serotonin remains in plasma, but most is converted to 5-HIAA, which is excreted in urine. These patients have increased blood and platelet serotonin concentrations and increased urinary 5-HIAA excretion. However, some foregut NET lack aromatic L-amino acid decarboxylase and secrete 5-HTP rather than serotonin into the bloodstream. Patients with these tumors have normal serotonin concentrations in blood and in platelets, but urinary amounts are increased because 5-HTP is converted to serotonin in the kidney; urinary 5-HIAA may be slightly elevated.

Patients with serotonin-producing carcinoid tumors usually have striking increases in urinary 5-HIAA excretion (at least tenfold), but occasionally elevations are smaller. False-positive elevations have been known to occur if the patient ingests a wide range of serotonin-rich foods or medications. Conversely, alcohol, aspirin, and other drugs suppress 5-HIAA concentrations. Patients should avoid these agents during 24-hour urine collections. Fasting plasma 5-HIAA has been proposed as a convenient replacement for urine collections.

Most physicians rely on the measurement of 5-HIAA to diagnose carcinoid syndrome. But when a patient who is strongly suspected for carcinoid syndrome shows normal or borderline increases in urinary 5-HIAA, the documentation of elevated serotonin concentrations in whole blood or platelets may help establish the diagnosis. Platelet serotonin has been reported to be more sensitive than urinary 5-HIAA for detecting carcinoids that produce small or moderate amounts of serotonin, such as foregut and hindgut carcinoids, and midgut carcinoids with low tumor volume. Also, platelet serotonin concentrations are not affected by the patient's diet. However, platelets can be saturated at high serotonin secretion rates, and 5-HIAA is often preferred for monitoring high serotonin production. Measurements of chromogranin A provide an important complement to measurements of serotonin and 5-HIAA and are particularly useful for many GEP-NET that do not produce appreciable amounts of serotonin.

Preanalytical and Analytical Considerations

In clinical practice, laboratory determinations of catecholamines, serotonin, and their metabolites in biological fluids are performed primarily for the diagnosis and follow-up of patients with catecholamine- or serotonin-producing tumors. In accordance with guidelines, most laboratories now offer measurements of urinary or plasma metanephrines (normetanephrine and metanephrine) as primary tests for the diagnosis of pheochromocytoma and paraganglioma. Liquid chromatography with tandem mass spectrometry (LC-MS/MS) has increasingly become the method of choice for these measurements (Fig. 26.8). Measurements of catecholamines, usually in urine and less frequently in plasma, are also employed at some medical centers as secondary tests. Measurements of urinary VMA have shown declining importance. For the detection of neuroblastoma, urinary HVA and VMA remain the most

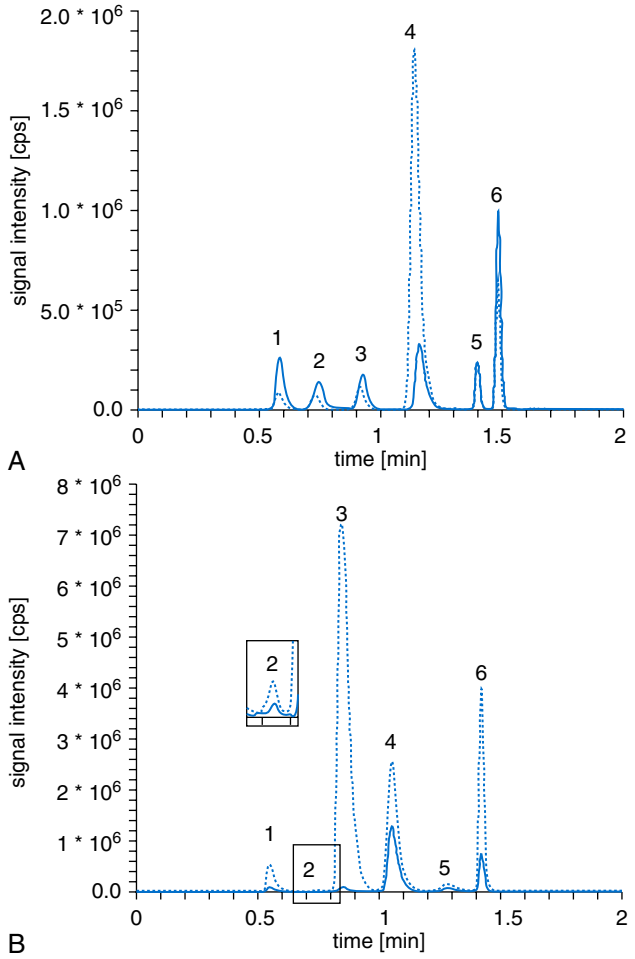


Fig. 26.8 Chromatograms obtained by liquid chromatography with tandem mass spectrometry for simultaneous measurement of urinary catecholamines and their free *O*-methylated derivatives for a calibrator (A) and extracted urine samples of two patients (B). Chromatographic peaks are shown for norepinephrine *(1)*, epinephrine *(2)*, normetanephrine *(3)*, dopamine *(4)*, metanephrine *(5)*, and 3-methoxytyramine *(6)*, with respective concentrations in the calibrator (A) of 59.1, 54.6, 54.6, 65.3, 50.7, and 59.8 nmol/L. For the calibrator, internal standards are shown by a *dotted line*. Patient sample-derived chromatograms (B) represent the urinary output of catecholamines and their *O*-methylated metabolites from a patient with confirmed pheochromocytoma *(dashed line)* and a patient without pheochromocytoma. Note the much larger signal intensity *(peak height)* for normetanephrine (3) in the patient with rather than without pheochromocytoma.

commonly ordered biochemical tests in clinical practice, but other catecholamine metabolites and dopamine are also measured. Diagnostic evaluation of patients with carcinoid tumor routinely involves measurement of 5-HIAA; the measurement of serotonin in platelets and urine has also been advocated.

Collection, Processing, and Storage of Samples

The conditions for plasma or urine collection are crucial to the reliability and interpretation of test results. Some clinicians prefer 24-hour urine collection to blood sampling because the former avoids the rigid sampling conditions associated with blood catecholamine collection (e.g., samples should be taken after supine rest) and is more convenient for clinical staff to implement. However, 24-hour urine collections are often difficult and inconvenient for patients. Spot or overnight urine collection with correction to urinary creatinine excretion provides an alternative to 24-hour urine collection and may overcome some of these problems.

For measurements of urinary metanephrines, no preservatives are required, whereas for catecholamines, samples are best preserved with hydrochloric acid. For storage over protracted periods, aliquots are best kept frozen at −80°C to minimize auto-oxidation and deconjugation. Blood samples for both metanephrines and catecholamines are best collected into tubes containing heparin or ethylenediaminetetraacetic acid (EDTA) as an anticoagulant and stored on ice before centrifugation at 4°C, with separation of plasma for further storage at −80°C.

Whole blood measurement of serotonin is popular because time-consuming isolation of platelets is not required. For whole blood serotonin, venous blood is drawn into a tube containing potassium EDTA as an anticoagulant, gently mixed and placed on ice, and transferred to a storage tube. An aliquot of blood is then removed for a platelet count. Blood serotonin samples are stored frozen at −20°C, preferably within 2 hours after collection.

Platelet-rich plasma samples are prepared from whole blood by low-speed centrifugation. To prevent lowering of serotonin concentration, platelet-rich plasma is prepared within 1 hour after the blood is collected and placed on ice. Plasma and pellets are stored frozen at −20°C and are analyzed within 1 to 2 weeks after collection.

Twenty-four-hour urine samples for serotonin and 5-HIAA are collected in 2-L brown polypropylene bottles, containing 250 mg of sodium metabisulfite and EDTA as preservatives. Samples are acidified to pH 4 with acetic acid before freezing and, importantly, the specimen should be refrigerated during collection.

Influences of Diet and Drugs

Dietary constituents or drugs may cause direct analytical interference in assays or may influence the physiological processes that determine plasma and urinary concentrations of monoamines and monoamine metabolites. In the former circumstances, interference is highly variable depending on the particular measurement method used. In the latter circumstances, interference is usually of a more general nature and is independent of the measurement method.

The development of new drugs, the variations in assay techniques, and the continuing improvements in analytical procedures make it difficult to identify which directly interfering medications should be avoided for a given analytical test. However, for modern mass spectrometric detection methods, such interferences are rarely a problem compared with electrochemical and spectrophotometric detection methods.

More readily identifiable sources of interference, which are independent of the particular assay method, tend to be associated with drugs that have primary actions on monoamine systems. Tricyclic antidepressants, which inhibit monoamine reuptake, are a major source of false-positive results for

norepinephrine and normetanephrine. Other medications that cause significant interference but are less commonly encountered during testing for pheochromocytoma include L-dopa, Sinemet, α-methyldopa (Aldomet), and MAO inhibitors.

Dietary interference is sometimes particularly troublesome for measurements of serotonin metabolites. Dietary sources of 5-hydroxyindoles, such as walnuts, bananas, avocados, eggplants, pineapples, plums, and tomatoes, should be restricted 3 to 4 days before and during urine collection. If possible, patients should abstain from all known medications that may cause an apparent increase, such as guaiacolate, mephenesin, phenacetin, and acetaminophen, or a decrease in 5-HIAA, including methenamine, phenothiazine tranquilizers, homogentisic acid, acetic acid, and levodopa. Due to influences of dietary catecholamines, plasma methoxytyramine should be measured after an overnight fast.

Analysis of Urinary and Plasma Fractionated Metanephrines

Normetanephrine, metanephrine, and methoxytyramine, are present in plasma and urine in free and sulfate-conjugated forms. Plasma concentrations of the conjugates are 20- to 30-fold higher than those of the free metabolites, reflecting more rapid circulatory clearance of free than conjugated metabolites. Clearance of free metabolites occurs by active uptake into tissues followed by deamination or conjugation. The conjugated metabolites are cleared relatively slowly by renal excretion. Metanephrines in urine are therefore routinely measured after acid hydrolysis and represent mainly conjugated metabolites, whereas metanephrines in plasma are usually measured in the free form.

Measurements of metanephrines in urine and plasma are now widely performed in many laboratories by LC-MS/MS (see Fig. 26.7). Such methods offer advantages of short chromatographic run times, high sample throughput, and elimination of drug interference that render earlier methods, using liquid chromatography with electrochemical detection (LC-EC), obsolete. A sample preparation step, usually employing ion exchange chromatography, remains a requirement; however, this is accomplished more simply for LC-MS/MS than for LC-EC methods, with some methods allowing on-line sample purification.

Immunoassay measurements of plasma and urine metanephrines suffer from relatively poor accuracy and precision and are not recommended for routine clinical use.

Analysis of Urinary and Plasma Catecholamines

Plasma catecholamines are still commonly measured by LC-EC methods, whereas urinary catecholamines are increasingly measured with LC-MS/MS methods. The most common extraction procedure employed for LC-EC measurements of catecholamines involves alumina extraction, sometimes combined with an additional cation-exchange step. For LC-MS/MS methods, simpler methods of extraction can be used, such as 96-well based ion-exchange chromatographic procedures, but with final elution of the analytes from extraction columns under acid conditions rather than basic conditions commonly used for metanephrines. Further advantage of the LC-MS/MS method is that both catecholamines and metanephrines can be purified and measured together

Analysis of Urinary Vanillylmandelic Acid and Homovanillic Acid

VMA, in contrast to HVA, is not significantly conjugated; both are excreted in urine in relatively high amounts, making their analysis relatively simple. Because large amounts of HVA are also present in urine in the free form, both metabolites are commonly measured without a hydrolysis step. Compared with metanephrines and catecholamines, measurements of VMA and HVA have limited value in diagnosing pheochromocytoma but are commonly used for the diagnosis of neuroblastoma.

Spectrophotometric methods for determination of VMA and HVA have been superseded by gas or liquid chromatography-based methods, employing today mainly mass spectrometric detection. Mass spectrometric methods are highly specific and, along with combined measurements of VMA and HVA, certain methods allow the additional measurement of 5-HIAA. Gas chromatography-based methods typically employ extraction of urinary analytes into an organic phase followed by derivatization of analytes in the dried down residue. Sample preparation is simpler still for LC-MS/MS methods, some of which involve only the dilution of urine in the mobile phase.

Analysis of Serotonin and 5-Hydroxyindoleacetic Acid

Analyses of serotonin, its deaminated acid metabolite, and 5-HIAA and its precursor, 5-HTP, provide important biomarkers for the diagnosis of carcinoid tumors. Typically, circulating serotonin is almost entirely confined to platelets, so that measurements are usually performed in whole blood, platelet-rich plasma, or isolated platelet pellets. With sensitive analytical instruments, measurements are possible in platelet-free plasma, which has been proposed as an alternative matrix, and which is useful for the detection of serotonin-producing neoplasms at locations where venous outflow does not empty into the portal circulation.

Measurements of urinary serotonin are of less importance because the serotonin in this matrix primarily reflects the clearance of plasma-free serotonin and renal decarboxylation of circulating 5-HTP. Nevertheless, urinary measurement is helpful in identifying tumors deficient in aromatic amino acid decarboxylase that produce substantial amounts of 5-HTP, which consequently shows up in urine as serotonin. Such tumors may be more directly detected by the measurement of circulating 5-HTP.

Measurements of 5-HIAA are most commonly carried out in urine. As an end-product of serotonin production, 5-HIAA better reflects tumor burden than serotonin. Measurements in plasma after an overnight fast have been advocated as more reliable because they avoid the confounding influence of dietary serotonin.

Liquid chromatography with fluorometric or electrochemical detection remains the most frequently used method for measurements of serotonin and 5-HIAA. Usually these techniques have been developed for measuring serotonin and 5-HIAA separately, but some methods allow for simultaneous measurement of catecholamine and 5-HIAA, VMA, and HVA. For detection of high concentrations of serotonin, as in platelet-rich plasma, simple methods of deproteinization of samples with perchloric acid or trichloroacetic acid before direct injection onto the high-performance liquid chromatography (HPLC) column are used.

Most HPLC assays for serotonin and 5-HIAA employ octadecylsilyl (C18) reversed-phase columns, although strong cation exchange columns have been used for serotonin. The chromatography is usually performed with an isocratic mobile phase at an acid pH that contains an organic modifier and perhaps an ion-pair reagent for serotonin measurements. As with catecholamines, serotonin is protonated in the pH range of 3 to 6, and the addition of an anionic ion-pair reagent creates an uncharged conjugate, which enhances the affinity of serotonin for the hydrophobic stationary phase.

For the measurement of very small amounts of serotonin or for specialized projects, HPLC with amperometric or coulometric detection is often favored over fluorometric detection. To enhance analytical sensitivity, some HPLC procedures incorporate precolumn derivatization with fluorescent and chemiluminescent reagents, thereby achieving detection limits in the femtomole range. Electrochemical detection using amperometric or coulometric measurement is preferred for the specific measurement of small quantities of 5-HIAA. Some HPLC systems use fluorometric detection, with or without derivatization, for a less demanding measurement of 5-HIAA.

As with measurements of metanephrines and catecholamines, the use of LC-MS/MS methods for the quantification of serotonin and 5-HIAA is increasing. With these methods, sample clean-up procedures are simpler, enabling online sample purification.

Postanalytical Considerations and Reference Intervals

The use of appropriately matched reference populations is important for effective screening for monoamine-producing tumors among different populations of patients. Plasma and urinary metanephrines show variable sex and age differences, and may have different ranges in hospitalized patients compared with normotensive healthy volunteers. The amounts of catecholamines and metanephrines in 24-hour urine specimens and plasma are non-normally distributed so that non-parametric methods are most appropriate for establishing reference intervals (see Chapter 5). For plasma metanephrines it is essential that reference intervals are established using blood samples collected supine, and for methoxytyramine after an overnight fast. Age can be particularly important to consider, especially for pediatric populations, but also in adults for plasma normetanephrine. Because of the variations in methods of analysis and subsequent analytical test results, any laboratory establishing measurements of monoamines or metabolites for diagnostic purposes should validate its own reference intervals in line with current standards in clinical practice (for details, see Chapter 5).

REVIEW QUESTIONS

1. The urinary metabolite measured as an indicator of norepinephrine synthesis is:
   a. dopamine.
   b. homovanillic acid (HVA).
   c. serotonin.
   d. vanillylmandelic acid (VMA).
2. False-positive elevations of 5-hydroxyindoleacetic acid (5-HIAA) in urine can occur if:
   a. an individual is a chronic alcoholic.
   b. a patient is receiving aspirin therapy.
   c. an individual has recently eaten fruit, such as bananas, kiwis, and plums.
   d. a patient is currently using tricyclic antidepressants.
3. The two urine metabolites that are most widely used for the diagnosis of neuroblastoma are:
   a. HVA and VMA.
   b. VMA and 5-HIAA.
   c. HVA and 5-HIAA.
   d. dopamine and epinephrine.
4. Midgut neuroendocrine tumors release which of the following substances in a large quantity when compared with tumors derived from other developmental gut components?
   a. 5-hydroxytryptophan
   b. Serotonin
   c. Epinephrine
   d. Dopamine
5. The method most frequently used for measuring 5-HIAA in urine to detect serotonin production is:
   a. nephelometry.
   b. liquid chromatography with UV detection.
   c. liquid chromatography with electrochemical, fluorimetric or mass spectrometric detection.
   d. ion exchange chromatography.
6. The recommended analytes for diagnosis of pheochromocytoma are:
   a. urinary VMA and HVA.
   b. plasma or urinary metanephrines.
   c. urinary or plasma catecholamines.
   d. urinary 5-HIAA and HVA.
7. A catecholamine-producing tumor derived from the chromaffin cells of the adrenal medulla is referred to as a:
   a. neuroblastoma.
   b. pheochromocytoma.
   c. paraganglioma.
   d. gastroenteropancreatic neuroendocrine tumor.

8. More than 95% of the serotonin in the body is produced within the gastrointestinal tract.
   a. True.
   b. False.
9. Chromaffin cells are located in the ___ and are responsible for synthesis of___:
   a. kidney; renin.
   b. brain; dopamine.
   c. adrenal gland; epinephrine.
   d. gut; serotonin.
10. Epinephrine is also called:
   a. noradrenaline.
   b. adrenaline.
   c. metanephrine.
   d. serotonin.

SUGGESTED READINGS

Beaulieu, J. M., Espinoza, S., & Gainetdinov, R. R. (2015). Dopamine receptors – IUPHAR review 13. *British Journal of Pharmacology*, *172*, 1–23.

Cheung, N. K., & Dyer, M. A. (2013). Neuroblastoma: Developmental biology, cancer genomics and immunotherapy. *Nature Review Cancer*, *13*, 397–411.

Dahia, P. L. (2014). Pheochromocytoma and paraganglioma pathogenesis: Learning from genetic heterogeneity. *Natural Review Cancer*, *14*, 108–119.

Eisenhofer, G. (2001). The role of neuronal and extraneuronal plasma membrane transporters in the inactivation of peripheral catecholamines. *Pharmacological Therapy*, *91*, 35–62.

Eisenhofer, G., Kopin, I. J., & Goldstein, D. S. (2004). Catecholamine metabolism: a contemporary view with implications for physiology and medicine. *Pharmacological Reviews*, *56*, 331–349.

Eisenhofer, G., & Peitzsch, M. (2014). Laboratory evaluation of pheochromocytoma and paraganglioma. *Clinical Chemistry*, *60*, 1486–1499.

Franscini, L. C., Vazquez-Montes, M., Buclin, T., Perera, R., Dunand, M., Grouzmann, E., et al. (2015). Pediatric reference intervals for plasma free and total metanephrines established with a parametric approach: relevance to the diagnosis of neuroblastoma. *Pediatrics Blood Cancer*, *62*, 587–593.

Furness, J. B. (2012). The enteric nervous system and neurogastroenterology. *Nature Reviews Gastroenterology & Hepatology*, *9*, 286–294.

Goldstein, D. S. (2010). Catecholamines 101. *Clinical Autonomic Research*, *20*, 331–352.

Kema, I. P., de Vries, E. G., & Muskiet, F. A. (2000). Clinical chemistry of serotonin and metabolites. *Journal of Chromatography B, Biomedical Sciences and Applications*, *747*, 33–48.

LaBrosse, E. H., Comoy, E., Bohuon, C., Zucker, J. M., & Schweisguth, O. (1976). Catecholamine metabolism in neuroblastoma. *Journal of the National Cancer Institute*, *57*, 633–638.

Lenders, J. W., Eisenhofer, G., Mannelli, M., & Pacak, K. (2005). Phaeochromocytoma. *Lancet*, *366*, 665–675.

Lenders, J. W. M., Duh, Q. Y., Eisenhofer, G., Gimenez-Roqueplo, A. P., Grebe, S. K., Murad, M. H., et al. (2014). Pheochromocytoma and paraganglioma: an Endocrine Society Clinical Practice guideline. *The Journal of Clinical Endocrinology and Metabolism*, *99*, 1915–1942.

Lionetto, L., Lostia, A. M., Stigliano, A., Cardelli, P., & Simmaco, M. (2008). Hplc-mass spectrometry method for quantitative detection of neuroendocrine tumor markers: vanillylmandelic acid, homovanillic acid and 5-hydroxyindoleacetic acid. *Clinical Chimica Acta*, *398*, 53–56.

Manger, W. M. (2009). The protean manifestations of pheochromocytoma. *Hormone and Metabolic Research*, *41*, 658–663.

Maris, J. M. (2010). Recent advances in neuroblastoma. *The New England Journal of Medicine*, *362*, 2202–2211.

McCorvy, J. D., & Roth, B. L. (2015). Structure and function of serotonin g protein-coupled receptors. *Pharmacological Therapy*, *150*, 129–142.

Oberg, K. (2011). Circulating biomarkers in gastroenteropancreatic neuroendocrine tumours. *Endocrine Related Cancer*, *18*(Suppl. 1), S17–S25.

Rindi, G., & Wiedenmann, B. (2011). Neuroendocrine neoplasms of the gut and pancreas: new insights. *Natural Review Endocrinology*, *8*, 54–64.

Schilling, F. H., Spix, C., Berthold, F., Erttmann, R., Fehse, N., Hero, B., et al. (2002). Neuroblastoma screening at one year of age. *The New England Journal of Medicine*, *346*, 1047–1053.

27

Vitamins and Trace Elements

Ravinder Sodi, Andrew Taylor

OBJECTIVES

1. Classify the vitamins discussed in this chapter according to their water or fat solubility.
2. State the chemical names of each vitamin discussed in this chapter.
3. Summarize the physiological roles of each vitamin discussed.
4. Describe the symptoms of toxicity and deficiency of each vitamin.
5. Describe methods used in the analysis of each vitamin and the factors which affect them.
6. State what is meant when an element is considered "essential."
7. Summarize the physiological functions and clinical significance of each trace element discussed.
8. Describe methods used in the analysis of trace elements and the factors which affect them.
9. Describe disorders associated with deficiency of the trace elements discussed.

KEY WORDS AND DEFINITIONS

Acrodermatitis enteropathica A hereditary disorder due to defective zinc uptake resulting in dermatitis, diarrhea, and alopecia.

Acute-phase response (APR) A response of the body to inflammation that results in an increase or decrease in the plasma concentrations of a class of proteins known as acute phase reactants.

Apoenzyme A protein moiety of an enzyme that requires a coenzyme to be fully functional.

Beriberi A disease caused by a deficiency of thiamine (vitamin B_1) and characterized by polyneuritis (dry form), cardiac pathology with edema (wet form).

Coenzyme An organic molecule (sometimes derived from a vitamin) of low molecular weight that, when combined with an inactive protein called an *apoenzyme*, forms an active compound or a complete enzyme called a *holoenzyme*, which functions catalytically in an enzyme system.

Cofactor A natural reactant, usually a metal ion or a coenzyme, that is required in an enzyme-catalyzed reaction.

Hartnup disease An inborn error of metabolism affecting the absorption of nonpolar amino acids (particularly tryptophan) characterized by a massive aminoaciduria.

Hemorrhagic disease of the newborn A self-limited hemorrhagic disorder of the first days of life, caused by a deficiency of the vitamin K–dependent blood coagulation factors II, VII, IX, and X.

Imerslund-Grasbeck syndrome A rare autosomal recessive, familial form of vitamin B_{12} deficiency caused by defects in the cubam receptor located in the terminal ileum, which is involved in vitamin B_{12} absorption.

Kashin–Beck disease A chronic, endemic disease of the bone resulting from mainly selenium and iodine deficiency.

Keshan disease A fatal, congestive cardiomyopathy caused by deficiency of essential trace elements in the diet particularly selenium.

Menkes syndrome An X-linked recessive disorder of copper metabolism caused by mutations in the *ATP7A* gene (locus: Xq12–q13), which encodes a copper transporter leading to copper deficiency. Signs and symptoms include kinky hair, hypotonia, and developmental delay.

Nyctalopia Visual disturbances that occur in dim light that may be caused as a result of vitamin A deficiency among other causes. Also known as night blindness.

Pellagra A clinical deficiency syndrome due to deficiency of niacin (or failure to convert tryptophan to niacin) and characterized by dermatitis, inflammation of mucous membranes, diarrhea, and dementia.

Pernicious anemia A deficiency in the production of red blood cells through a lack of vitamin B_{12}.

Recommended dietary allowance (RDA) The average daily level of intake of nutrients sufficient to meet the requirements of nearly all (97% to 98%) healthy people.

Scurvy A condition due to deficiency of ascorbic acid (vitamin C) in the diet.
Systemic inflammatory response syndrome (SIRS) The systemic inflammatory response to a wide variety of severe clinical insults.
Total parenteral nutrition (TPN) The practice of feeding a person intravenously, circumventing the gut.
Trace elements Inorganic molecules found in human and animal tissues in milligram per kilogram amounts or less.
Vitamer Term used to describe any of a number of compounds that possess a given vitamin activity.
Vitamin An essential organic micronutrient that must be supplied exogenously and in many cases is the precursor to a metabolically derived coenzyme.
Wilson disease A rare, progressive, autosomal recessive disease due to a defect in metabolism of copper resulting in excessive copper accumulation in the liver, brain, eye, and other organs.
Wernicke encephalopathy Condition arising as a result of thiamine deficiency, which presents with ophthalmoplegia, ataxia, and confusion usually in alcoholics.
Xerophthalmia Refers to the spectrum of ocular manifestations due to vitamin A deficiency, ranging from night blindness to conjunctival dryness to the appearance of small gray plaques with foamy surfaces (Bitot spots).

VITAMIN AND TRACE ELEMENT STATUS

An adequate supply of **vitamins** and trace elements is critical in maintaining health. There is increasing public awareness that good nutrition including supplementation with vitamins and trace elements is necessary for improving the quality of life but this has not necessarily been supported by randomized trials in selected groups. The general principle regarding assessment of nutritional status is to determine the extent to which the metabolic demand for nutrients has been or is currently being met by the supply. In clinical practice, this requires balancing supply and demand.

The requirements for most nutrients including vitamins and trace elements, to maintain health have been characterized and made available in reports from the Institute of Medicine (IOM) of the National Academies (US). However, the effects of disease may increase demands. For example, hypermetabolism, as a result of trauma or infection, increases the need for protein and energy and for the vitamin and trace element **cofactors** necessary for their metabolism. Increased losses from the gut, kidney, and skin, or through dialysis, may also increase the overall demand for these nutrients.

An estimate of supply also is obtained from a careful dietary history, especially if performed by a dietitian, together with knowledge of any artificial nutritional supplements or therapy that may have been provided enterally or intravenously. Table 27.1 summarizes the **Recommended Dietary Allowance (RDA)** used in the United States and the Population Reference Intakes from the European community for vitamins and trace elements in adults. Table 27.1 also summarizes the amounts present in 2000 kcal of most tube feeds used in nutritional support and common dietary sources. For details regarding children, see the suggested readings. Table 27.2 summarizes the functions of vitamins in man.

In an attempt to improve accuracy of assessment of nutritional status, clinicians often turn to the laboratory to obtain a result that may reflect the net balance of supply and demand. Clinical laboratorians need to be aware of when such tests are useful and how to place the results of laboratory tests into the context of the clinical situation of the patient. It is important to be aware of the limitations of laboratory tests, especially in acutely ill patients. Nutritional assessment in health and disease is considered in detail in Suggested Reading reference 7.

Effect of Inflammation on Vitamins and Trace Elements

The concentration in plasma of various vitamins and trace elements can alter significantly when a **systemic inflammatory response syndrome** (**SIRS**; previously known as the **acute-phase response [APR]**) results from trauma or infection. This usually occurs independently of tissue stores. The associated changes may be due to alterations in the binding proteins in plasma such as albumin or retinol-binding protein (RBP), which decrease as part of SIRS or due to protein malnutrition. Studies in patients with SIRS as categorized on the basis of elevated C-reactive protein (CRP) concentrations have documented decreases in vitamin A, E, B_2, B_6, C, D, and carotenoids. One study in 1303 patients examined the effect of the magnitude of SIR on plasma micronutrient concentrations with the aim of providing guidance on the interpretation of results. The findings are shown in Table 27.3. This study concluded that the degree of inflammatory response affected the interpretation of plasma micronutrient concentrations and showed that a reliable clinical assessment could only be made if the CRP was less than 10 mg/L for vitamins A and D or less than 5 mg/L for vitamins B_6 and C. There have been guidelines issued on nutrition support by national bodies such as the National Institute for Health and Care Excellence (NICE) in the United Kingdom that recommend the monitoring of micronutrient concentrations in patients receiving **total parenteral nutrition (TPN)**; therefore when implementing guidelines and deciding on nutritional support, it is important to take the effect of SIRS into account.

TABLE 27.1 Daily Oral and Intravenous Vitamin and Trace Element Intakes for Adults

	RDA (USA)	PRI (Europe)	Some Natural Food Sources	Amount in 2000 kcal Tube Feed	IV Intake
Vitamins					
A, μg	900 (men); 700 (women)	700	Liver, yellow/orange fruits, leafy vegetables, fish, milk	1000–2160	1000
E, mg	15 (men and women)	0.4/g PUFA	Fruits and vegetables, meats, vegetable oils, unprocessed cereals, nuts, and seeds	20–64	9.1
K, μg	120 (men); 90 (women)	100–200	Leafy green vegetables, especially spinach, eggs, liver	—	150
B_1, mg	1.2 (men); 1.1 (women)	1.1	Brown rice, vegetables, potatoes, liver, eggs	1.4–3.4	6
B_2, mg	1.3 (men); 1.1 (women)	1.6	Vegetables such as green beans and asparagus, bananas, dairy products	2–6	3.6
B_6, mg	1.3 (men and women <50 year); 1.7 (men >50 year); 1.5 (women >50 year)	1.5	Meat, bananas, vegetables, nuts	2–13.8	6
B_{12}, μg	2.4	1.4	Meat and animal products	3–15	5
Folate, μg	400	200	Leafy vegetables, cereal especially if fortified, bread, pasta, liver	340–880	600
C, mg	90 (men); 75 (women)	45	Fruits especially citrus fruits and vegetables	100–300	200
Biotin, μg	30[a]	15–100	Eggs, organ meats, yeast, milk	100–660	60
Niacin, mg	16 (men); 14 (women)	18	Meat, fish, eggs, vegetables, mushrooms, nuts	18–45	40
Pantothenic acid, mg	5[a]	3–12[b]	Meat, vegetables like broccoli, avocadoes, grains	7–20	15
Trace Elements					
Copper, mg	0.9	1.1	Organ meats, seafood, nuts, seeds, grains, cocoa products	2–3.4	0.3–1.3
Selenium, μg	55	55	Organ meats, seafood, nuts, vegetables (dependent of soil selenium content)	30–130	40–100
Zinc, mg	11 (men); 8 (women)	9.5	Red meats and some seafood	13–36	2.5–6.0
Iodine, μg	150	150	Milk, iodized salt	N/A	70–170
Manganese, mg	2.3[a] (men); 1.8[a] (women)	1–10[b]	Nuts, legumes, tea, grains	2.4–8	0.06–0.1
Chromium, μg	25[a] (men); 35[a] (women)	30–200	Cereals, meats, fish	10–20	10–15
Molybdenum, μg	45[a]	74–240	Legumes, grains, and nuts	19	10–25

[a]Adequate intake.
[b]Acceptable range.
IV, Intravenous; *PRI*, population reference intake (Europe); *PUFA*, polyunsaturated fatty acids; *RDA*, recommended dietary allowance (US).
Reference intakes for pregnant and lactating women, infants and children vary with age and weight (see Suggested Reading reference 7).

VITAMINS

Vitamin A

Vitamin A is the nutritional term for the group of compounds with a 20-carbon structure containing a methyl-substituted cyclohexenyl ring (β-ionone ring) and an isoprenoid side chain (Fig. 27.1), with a hydroxyl group (retinol), an aldehyde group (retinal), a carboxylic acid group (retinoic acid), or an ester group (retinyl ester) at the terminal C15. Included in the vitamin A family are some dietary carotenoids (C40 polyisoprenoid compounds) that are classified as provitamin A because they are cleaved biologically to yield retinol. The principal dietary carotenoids are β-carotene, α-carotene, and β-cryptoxanthin.

Absorption, Transport, Metabolism, and Excretion

Preformed vitamin A, most often in the form of retinyl ester, or carotenoids are subject to emulsification and mixed micelle formation by the action of bile salts before they are transported into the intestinal cell. Here the

TABLE 27.2 Vitamins Required by Humans

Common or Generic Name	Vitamer or Common Chemical Name	Function	Symptoms and Causes of Deficiency or Associated Diseases	Currently Used Methods	Reference Intervals
Fat-Soluble Vitamins					
Vitamin A	Retinol, retinal, retinoic acid, carotenoids	Antioxidant, role in vision, gene expression, embryonic development, immune and reproductive functions	Nyctalopia (night blindness), Bitot spot, xerophthalmia, keratomalacia; common in infants and children, especially in less developed countries; due to fat malabsorption, cystic fibrosis and may occur due to retinol-binding protein deficiency as a results of protein malnutrition, liver disease, and zinc deficiency	HPLC, LC-MS/MS	1–6 year: 20–40 μg/dL (0.70–1.40 μmol/L) 7–12 year: 26–49 μg/dL (0.91–1.71 μmol/L) 13–19 year: 26–72 μg/dL (0.91–2.51 μmol/L) >19 year: 30–80 μg/dL (1.05–2.80 μmol/L)
Vitamin E	Tocopherols, tocotrienols	Antioxidant (prevents the peroxidation of unsaturated fatty acids), role in gene transcription, immunity, inhibits platelet aggregation, recently been implicated in bone physiology	Lipid peroxidation, red blood cell fragility causing hemolytic anemia especially in premature infants; may occur due to fat malabsorption, cystic fibrosis	HPLC, LC-MS/MS	Premature neonates: 0.1–0.5 mg/dL (2.3–11.6 μmol/L) 1–12 year: 0.3–0.9 mg/dL (7–21 μmol/L) 13–19 year: 0.6–1.0 mg/dL (14–23 μmol/L) >19 year: 0.5–1.8 mg/dL (12–42 μmol/L) As a ratio of cholesterol: 3.5–9.5 μmol/mmol cholesterol
Vitamin K	Phylloquinones (K_1), menaquinones (K_2), menadiones (K_3)	Coagulation, bone metabolism	Increased clotting time, hemorrhagic disease of the newborn; also due to fat malabsorption, cystic fibrosis, liver disease	Prothrombin time, PIVKA, HPLC, LC-MS/MS	0.2–2.2 nmol/mmol triglyceride
Water-Soluble Vitamins					
Vitamin B_1	Thiamine, aneurin	Forms the coenzyme thiamine pyrophosphate (TPP) required for decarboxylation reactions involved in carbohydrate metabolism, and nerve function	Beriberi, Wernicke-Korsakoff syndrome in alcoholics, rare thiamine-responsive IOM	Erythrocyte transketolase, HPLC, LC-MS/MS	Erythrocyte transketolase activity: 0.75–1.30 U/g Hb (48.4–83.9 kU/mol Hb) Percent TPP effect (activation): Normal: 0%–15% Marginal: 16%–25% Deficient: >25% TPP concentration: 173–293 nmol/L or 280–590 ng/g Hb (erythrocytes); 90–140 nmol/L or 275–675 ng/g Hb (whole blood)

Continued

TABLE 27.2 Vitamins Required by Humans—cont'd

Common or Generic Name	Vitamer or Common Chemical Name	Function	Symptoms and Causes of Deficiency or Associated Diseases	Currently Used Methods	Reference Intervals
Vitamin B_2	Riboflavin	Essential component of coenzymes involved in reduction-oxidation (redox) reactions in the body	Angular stomatitis, dermatitis, photophobia, riboflavin-dependent IOM	Erythrocyte glutathione reductase, HPLC, LC-MS/MS	Erythrocyte glutathione reductase activation by FAD: Adequacy: 1.20 Marginal: 1.21–1.40 Deficiency: >1.41 Serum or plasma concentrations of riboflavin (median [range]): plasma FAD: 101 [57–170] nmol/L plasma FMN: 6.3 [3.3–14.1] nmol/L plasma riboflavin: 11 [4–34] nmol/L erythrocyte FAD: 1.9 [0.7–3.8] pmol/g Hb erythrocyte FMN: 0.11 [0.04–0.44] pmol/g Hb erythrocyte riboflavin: 0.02 [0.01–0.13] pmol/g Hb
Vitamin B_3	Niacin, nicotinic acid, nicotinamide	Coenzyme or cosubstrate in many biological redox reactions, and thus for energy metabolism	Pellagra (dermatitis, dementia, diarrhea) Seen in communities with corn-based staple diets, in carcinoid syndrome (precursor tryptophan diverted to serotonin formation), Hartnup disease (unable to absorb tryptophan), and medications such as isoniazid	HPLC & LC-MS/MS for urine metabolite, nicotinamide coenzymes	2.4–6.4 mg/day (17.5–46.7 μmol/day) or 1.6–4.3 mg/g creatinine (11.7–31.4 μmol/g creatinine)
Vitamin B_5	Pantothenic acid, panthenol, pantetheine	General metabolism, acetyl and acyl transfer	Burning feet syndrome	Microbiological, CPB, HPLC, LC-MS, GC-MS	Whole blood or serum: 344–583 μg/L (1.57–2.66 μmol/L) Urinary excretion: 1–15 mg/day (5–68 μmol/day)
Vitamin B_6	Pyridoxine, pyridoxal, pyridoxamine	The active form pyridoxal phosphate is required for synthesis, catabolism of various amino acids	Epileptiform convulsions, dermatitis, anemia, medications such as penicillamine and isoniazid decrease it, pyridoxine-responsive IOM notably homocystinuria, and hyperhomocysteinemia (together with vitamin B_{12} and folate deficiencies)	Aspartate transaminase, HPLC, LC-MS/MS	Plasma PLP: 9.5–24 ng/mL (39–98 nmol/L). Erythrocyte PLP: 250–680 pmol/g Hb
Vitamin B_7	Biotin, vitamin H	Coenzyme for carboxylation reactions involved in gluconeogenesis, lipogenesis and catabolism of branched-chain amino acids; roles in cell signaling, epigenetic regulation of genes and chromatin structure	Dermatitis, developmental delay Seen with excessive raw egg consumption, those on parenteral nutrition, and IOM notably biotinidase deficiency	Microbiological, CPB, carboxylases, avidin binding, urinary metabolites	0.5–2.20 nmol/L

Vitamin B_9	Pteroylglutamic acid, folic acid, folate	Required for the interconversions of amino acids such as homocysteine to methionine and the biosynthesis of purines and pyrimidines, required for DNA synthesis	Megaloblastic anemia, neural tube defects Caused by gut sterilization, malabsorption, decreased intake, increased requirements (e.g. pregnancy), medications (e.g., methotrexate), anticonvulsants Deficiency linked to hyperhomocysteinemia, cancer and stroke	Red blood cell and serum folate, CPB, microbiological, homocysteine	>3 μg/L (7.0 nmol/L) for serum folate and <150 μg/L (340 nmol/L) for RBC folate
Vitamin B_{12}	Cyanocobalamin, hydroxycobalamin, methylcobalamin	Required for erythropoiesis, methylation processes necessary for DNA and cell metabolism, and is cofactor for various enzymes notably those involved in the metabolism of methylmalonic acid and homocysteine	Pernicious and megaloblastic anemia, peripheral neuropathy Caused by decreased intake (vegetarians), short bowel syndrome (loss of distal ileum), malabsorption syndromes, medications (N_2O, phenytoin, methotrexate, and proton pump inhibitors), and Imerslund-Grasbeck syndrome Deficiency results in methylmalonic aciduria and homocysteinemia	CPB, immunometric, microbiological, methylmalonate, homocysteine, holotranscobalamin	WHO consultation defined a serum vitamin B_{12} concentration <203 ng/L (150 pmol/L) as deficient
Vitamin C	Ascorbic acid	Connective tissue formation, antioxidant	Scurvy, infantile scurvy (Barlow disease) Linked with osteoporosis, anemia, diabetes mellitus, cancer	Spectrophotometric-enzymatic methods, HPLC	0.4 and 1.5 mg/dL (23–85 μmol/L) or 20–53 μg/10^8 leukocytes (1.14–3.01 fmol/leukocyte)

Reference intervals are likely to vary between laboratories and are influenced by variables including the systemic inflammatory response.
CPB, Competitive protein binding; *HPLC*, high-performance liquid chromatography; *IOM*, inborn errors of metabolism; *LC-MS/MS*, liquid chromatography-tandem mass spectrometry; *PIVKA*, protein induced by vitamin K absence or antagonism; *RIA*, radioimmunoassay.

retinyl esters are moved across the mucosal membrane and hydrolyzed to retinol within the cell to be then re-esterified by cellular RBP II and packaged into chylomicra, which then enter the mesenteric lymphatic system and pass into the systemic circulation. A small amount of the ingested retinoid is converted into retinoic acid in the intestinal cell.

Carotenoids, also in micellar form, are absorbed into the duodenal mucosal cells by passive diffusion. Once inside the mucosal cell, β-carotene is principally converted to retinal by the enzyme β-carotene-15,15′-dioxygenase. Retinal is converted by retinal reductase to retinol and esterified. The newly synthesized retinyl esters, from both preformed vitamin A and carotenoids, along with exogenous lipids and nonhydrolyzed carotenoids, then pass with chylomicrons via the lymphatic system to the liver, where uptake by parenchymal cells again involves hydrolysis. In the liver, retinol is bound with RBP and transthyretin (TTR or thyroxine-binding prealbumin) in a 1:1:1 complex of sufficient size to prevent loss by glomerular filtration and is returned to the circulation, or stored as esters within the stellate cells.

TABLE 27.3 Median Plasma Vitamin Concentrations According to C-Reactive Protein Concentrations

CRP Concentration (mg/L)	Vitamin A (μmol/L)	Vitamin D (nmol/L)	Vitamin B_6 (μmol/L)	Vitamin C (μmol/L)
<5	2.0	34	48	23
>5–10	2.0	**33**	**27**	18
>10–20	**1.8**	**31**	**32**	**17**
>20–40	**1.6**	**27**	**24**	**8**
>40–80	**1.4**	**23**	**18**	**6**
>80	**1.0**	**20**	**15**	**5**

Data in bold depict significant ($P < .05$) decrease compared to CRP category <5 mg/L.
Conversion factors for traditional units: Vitamin A—divide by 0.0349 to μg/dL; Vitamin D—divide by 2.496 to ng/mL; Vitamin B_6—divide by 4.046 to ng/mL; Vitamin C—divide by 56.78 to mg/dL.
CRP, C-reactive protein.
Modified from Duncan, A., Talwar, D., McMillan, D. C., Stefanowicz, F., & O'Reilly, D. S. (2012). Quantitative data on the magnitude of the systemic inflammatory response and its effect on micronutrient status based on plasma measurements. *The American Journal of Clinical Nutrition, 95*, 64–71.

Functions

The participation of retinal in vision is considered the most important physiologic function of vitamin A. All-*trans*-retinol is the predominant circulating form of vitamin A. Cells of the retina isomerize this to the 11-*cis* alcohol that is reversibly dehydrogenated to 11-*cis* retinal, which combines with proteins (e.g., opsin) to generate photosensitive pigments, such as rhodopsin. Illumination of such pigments causes photoisomerization and the release of all-*trans*-retinal and the protein, a process that couples the large conformational change with ion flux and optic nerve transmission. The all-*trans*-retinal is isomerized to the 11-*cis* isomer, which combines with the liberated protein to reconstitute the photo pigment in a visual cycle, as shown in Fig. 27.2.

Other functions of vitamin A include its role in reproduction, growth, embryonic development, and immune function; many of these functions are mediated through the binding of retinoic acid to specific nuclear receptors that regulate genomic expression.

Deficiency

Vitamin A deficiency primarily affects infants and children, and its prevalence is subject to World Health Organization (WHO) surveillance. Risk factors include poverty, low birth weight, poor sanitation, malnutrition, infection, and parasitism. As hepatic accumulation of vitamin A occurs during the last trimester of pregnancy, preterm infants are relatively vitamin A deficient at birth.

Neonatal vitamin A supplementation has been a matter of some controversy, with trials giving mixed results dependent on the population studied. At the present time neonatal vitamin A supplementation is not recommended as a public health intervention. Fat malabsorption, particularly caused

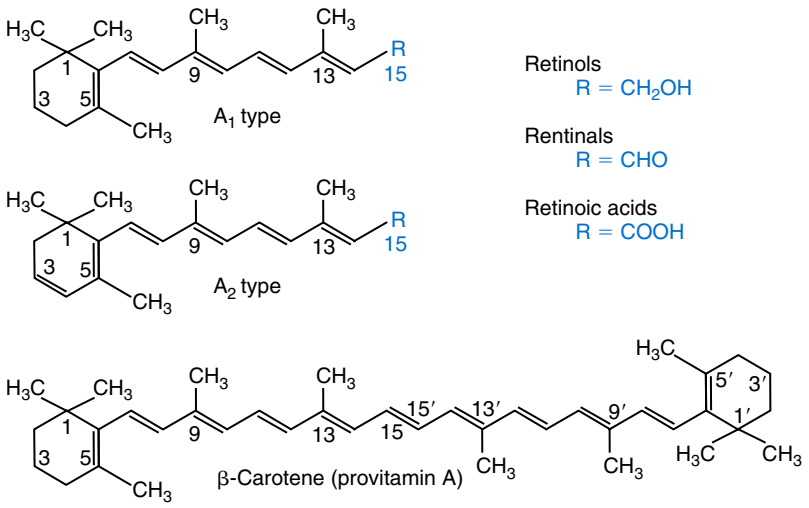


Fig. 27.1 Vitaminic forms of A_1, A_2, and β-carotene.

by celiac disease or chronic pancreatitis, and protein-energy malnutrition predispose to vitamin A deficiency. Liver disease diminishes RBP synthesis, and ethanol abuse leads to both hepatic injury and competition with retinol for alcohol dehydrogenase, which is necessary for the oxidation of retinol to retinal and retinoic acid. Vitamin A deficiency may lead to anemia, although the precise mechanism is not known.

Clinical features of vitamin A deficiency include degenerative changes in eyes and skin, and poor dark adaptation or *night blindness* (**nyctalopia**) followed by degenerative changes in the retina. **Xerophthalmia** refers to the spectrum of ocular manifestations due to vitamin A deficiency, ranging from night blindness to conjunctival dryness to the appearance of small gray plaques with foamy surfaces *(Bitot spots)*. These lesions are reversible with vitamin A administration. More serious effects of deficiency are known as *keratomalacia*, and cause ulceration and necrosis of the cornea that lead to perforation, prolapse, endophthalmitis, and blindness. Usually associated skin changes include dryness, roughness, papular eruptions, and follicular hyperkeratosis.

Toxicity

Toxic effects of hypervitaminosis A occur mainly as a result of ingestion of excess vitamin or as a side effect of inappropriate therapy. The elderly are more susceptible to vitamin A toxicity at lower doses, as exposure to retinyl esters is longer because of delayed postprandial clearance of lipoproteins.

Symptoms of acute toxicity present as abdominal pain, nausea, vomiting, severe headaches, dizziness, sluggishness, and irritability, followed within a few days by desquamation of the skin and recovery. Chronic toxicity from moderately high doses taken for protracted periods is characterized by bone and joint pain, hair loss, dryness and fissures of the lips, anorexia, benign intracranial hypertension, weight loss, and hepatomegaly.

Epidemiologic and experimental evidence has supported the view that high vitamin A intake in humans, acting via 13-*cis*-retinoic acid, is teratogenic. A further intriguing association, supported in part by epidemiologic studies, is that observed between excessive vitamin A intake and reduction in bone mineral density (BMD). Hypervitaminosis A is also a known cause of hypercalcemia, especially in chronic kidney disease.

Carotenemia results from chronic excessive intake of carotene-rich foods, principally carrots, and is usually reported in infants and children. This condition, in which yellowing of the skin is observed, is benign because the excess carotene is deposited rather than converted to vitamin A. Carotenemia has also been linked to amenorrhea, but the mechanism behind this association remains unknown. Elevated concentrations have also been found in hypothyroid patients, in whom conversion to vitamin A is decreased, and in patients with hyperlipemia associated with diabetes mellitus.

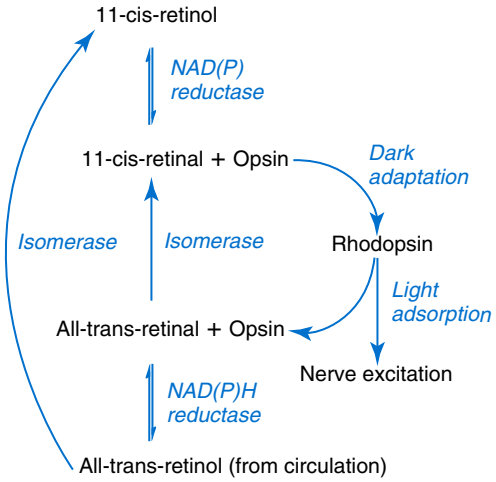


Fig. 27.2 Participation of A vitamers in the visual cycle. *NAD*, Nicotinamide-adenine dinucleotide.

Laboratory Assessment of Status

Although measurement of the plasma concentration of vitamin A is the most convenient and widely used assessment of vitamin A status, it is not an ideal indicator because it does not decline until liver stores become critically depleted, which is thought to occur at a concentration of approximately 20 μg/g liver.

Vitamin A status is assessed by the measurement of retinol concentration. Retinol circulates in plasma as a 1:1:1 complex with RBP and TTR. The circulating concentration of RBP is determined by dietary protein and zinc, which are necessary for RBP synthesis. Thus protein malnutrition, liver disease, and zinc deficiency resulting in RBP deficiency will lead to hypovitaminosis A. In contrast, renal failure resulting in decreased excretion of RBP has been reported to result in hypervitaminosis A. Another confounding factor in the assessment of vitamin A status is the effect of inflammation as discussed above. Both RBP and TTR are negative acute-phase proteins; thus inflammatory changes will result in transient falls in both proteins and plasma retinol. To distinguish inflammatory from nutritional causes of reduced plasma retinol concentrations, it may be necessary to measure CRP.

Currently high-pressure liquid chromatography (HPLC; see Chapter 12) after solvent extraction with fluorometric or spectrophotometric detection is mainly used to measure vitamin A. HPLC has brought enhanced specificity, lowered limits of detection, improved accuracy using primary standards, reference materials, and quality assurance schemes, and made acceptable reproducibility achievable (between batch coefficients of variation [CV] of <15% for both vitamin A and β-carotene). In the normal-phase HPLC, compounds to be separated are adsorbed to microparticulate silica gel and are eluted in the order of least polar to most polar. Reversed-phase HPLC is preferable for acid-sensitive compounds such as 5,6-epoxyretinoic acid. Photometric, electrochemical, and mass spectrophotometric detectors have all been used. Refer to Chapter 12 for general principles of chromatography. Vitamin A is light-sensitive, and samples should be protected from light.

Vitamin D

Vitamin D plays an essential role as a hormone in the control of calcium and phosphorous metabolism and bone physiology. It is discussed in Chapter 39.

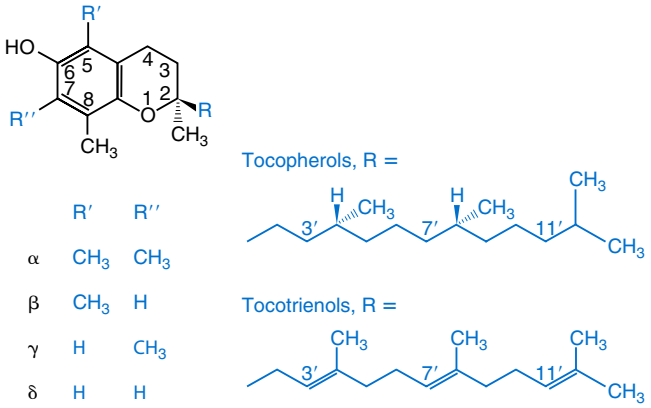


Fig. 27.3 Vitaminic forms of vitamin E.

Vitamin E

Vitamin E is the nutritional term for the group of tocopherols and tocotrienols that have biological activity similar to RRR-α-tocopherol (formerly D-α-tocopherol). The Greek letter prefixes α, β, γ, and δ indicates the presence or absence of methyl groups at positions 5 and 7 (Fig. 27.3).

Absorption, Transport, Metabolism, and Excretion

In the presence of bile, vitamin E is absorbed from the small intestine. Most forms of vitamin E are absorbed non-selectively and are secreted in chylomicron particles along with triacylglycerol and cholesterol. Some of this chylomicron-bound vitamin E is transported and delivered to the peripheral tissue (mainly adipose tissue) with the aid of lipoprotein lipase. The liver takes up the chylomicron remnants where α-tocopherol is incorporated into very low-density lipoproteins (VLDLs) by α-tocopherol transfer protein (α-TTP), enabling further distribution of α-tocopherol throughout the body. Vitamin E is excreted via the bile and in the urine as tocopheronic acid and its β-glucuronide conjugate.

Functions

The inhibition of free radical mediated lipid peroxidation is the main role for vitamin E. Tocopherols and tocotrienols inhibit lipid peroxidation largely because they scavenge lipid peroxyl radicals faster than the radical can react to adjacent fatty acid side chains or membrane proteins. The resultant tocopheryl or tocotrienyl radicals may then react with additional peroxyl radicals to produce tocopherones (nonradicals), or they may be regenerated by transfer of an electron to ascorbate to form the ascorbyl radical. Thus, vitamins E and C act synergistically to reduce lipid peroxidation (Fig. 27.4). α-Tocopherol also induces inhibition of cell proliferation, platelet aggregation, and monocyte adhesion, which are thought to be the result of direct interaction of α-tocopherol with cell components. There is some evidence that vitamin E may have anti-inflammatory properties. A recent study has shown that serum vitamin E is a determinant of bone mass by stimulating osteoclast fusion.

Deficiency

Premature and low birth weight infants are particularly susceptible to the development of vitamin E deficiency because placental transfer is poor and infants have such limited adipose tissue where much of the vitamin is normally stored. Signs of deficiency include (1) irritability, (2) edema, and (3) *hemolytic anemia*. Anemia reflects the shortened life span of erythrocytes with fragile membranes; it does not respond to iron therapy, which may aggravate the condition. Although symptoms of vitamin E deficiency are rare in children and adults, deficiency can occur in some conditions, such as fat malabsorption states including cystic fibrosis and chronic cholestasis in children that in turn cause neuropathy and hemolytic anemia. The genetic disorder *abetalipoproteinemia* (vitamin E is transported on lipoproteins) can also confer vitamin E deficiency (see Chapter 23).

Toxicity

Vitamin E toxicity is usually only ever achieved by excessive dietary supplementation. The US Food and Nutrition Board has recommended a tolerable upper limit of 1000 mg/day of vitamin E for adults 19 years and older.

Laboratory Assessment of Status

HPLC is currently the method of choice for quantification of tocopherols in serum, as it offers the advantages of accuracy (through the use of primary standards) and reproducibility (between-batch CV of <5%) and the ability to quantitate multiple analytes, including vitamin A and some carotenoids, in a single analytical run. Both α- and γ-tocopherols are the principal **vitamers** seen, although others may be detected with minor modifications to the analytical conditions. HPLC-mass spectrometry methods have also been developed and are increasingly being used, given the widespread adoption of this technique by clinical laboratories. Vitamin E is light-sensitive, and it is recommended samples should be protected from light.

Vitamin K

Vitamin K is the common generic name for a group of compounds with a methylated naphthoquinone structure (2-methyl-1,4-napthoquinones), which are substituted with side chains at carbon 3. *Phylloquinone* (K_1 type) synthesized in plants and *menaquinones* (K_2 type) of bacterial origin are the two principal natural classes of vitamin K (Fig. 27.5). Several synthetic analogs and derivatives have been used in human nutrition; most relate to or derive from *menadione* (K_3).

Absorption, Transport, Metabolism, and Excretion

As for other fat-soluble vitamins, the absorption of natural vitamin K from the small intestine into the lymphatic system is facilitated by bile. Vitamins K_1 and K_2 are bound to chylomicrons for transport from mucosal cells to the liver. Menadione (K_3) is more rapidly and completely absorbed from the gut before entering the portal blood. In the liver, intracellular distribution is seen mostly in the microsomal

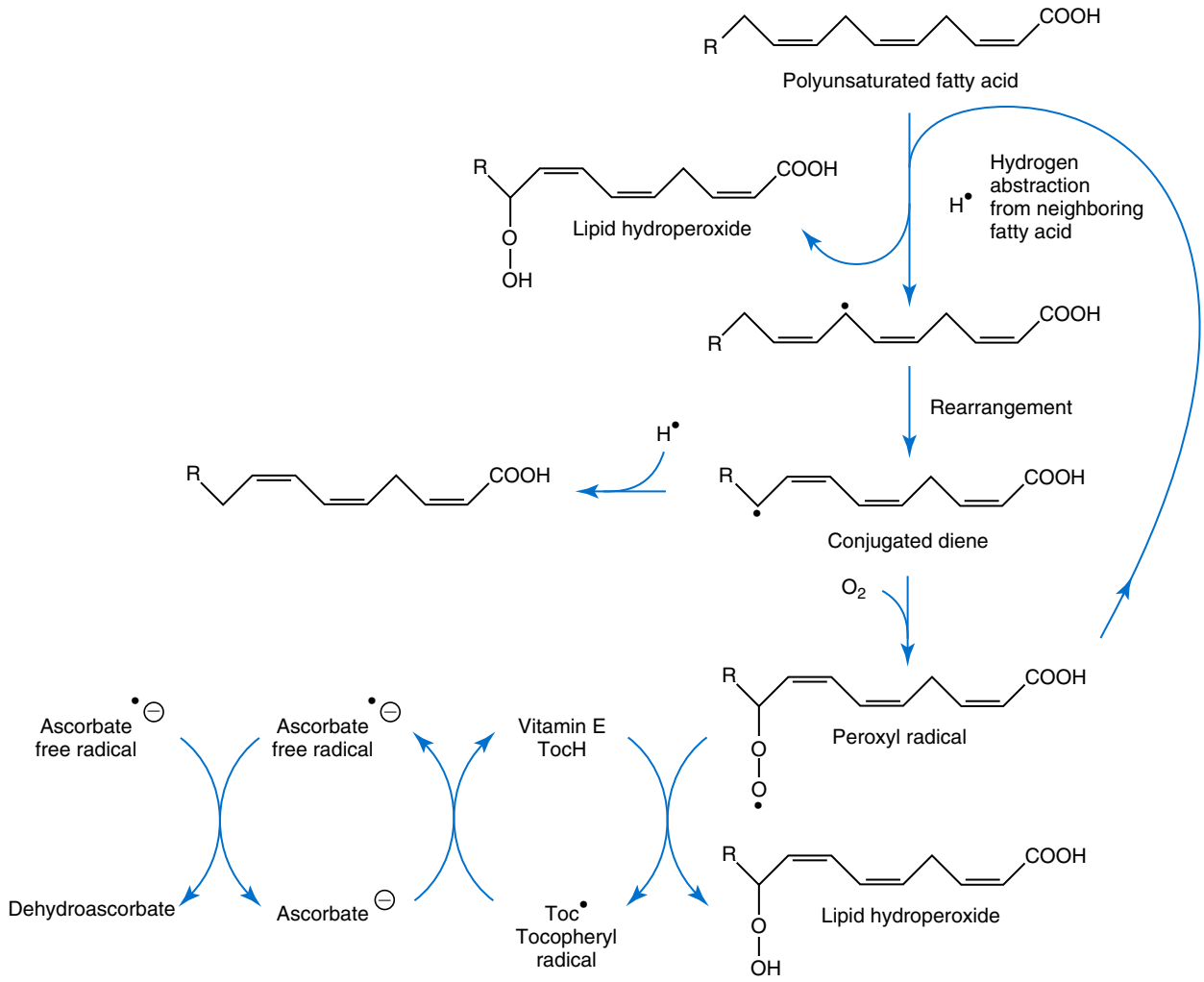


Fig. 27.4 Lipoperoxidation and synergistic action of vitamin E and vitamin C.

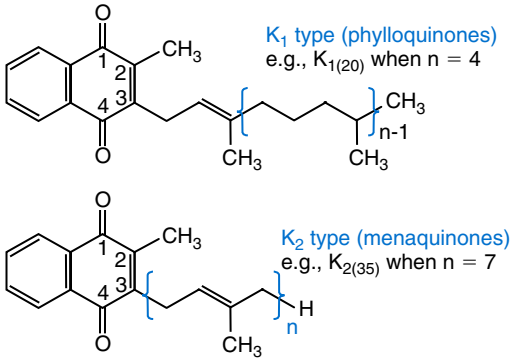


Fig. 27.5 Vitaminic forms of vitamin K.

fraction, where phenylation of menadione to form K_2 occurs. Release of vitamin K to the bloodstream allows association with circulating β-lipoproteins for transport to other tissue. Only traces of urinary metabolites of vitamins K_1 and K_2 appear in urine.

Functions

The main role of vitamin K is as a cofactor to vitamin K–dependent carboxylase, an enzyme necessary for the posttranslational conversion of specific glutamyl residues in target proteins to γ-carboxyglutamyl (Gla) residues. This γ-carboxylation increases the affinity of these proteins for calcium. The antihemorrhagic function of vitamin K depends on the formation of the Gla proteins prothrombin (factor II), proconvertin (factor VII), plasma thromboplastin component (factor IX), and Stuart factor (factor X), which together with two other hemostatic vitamin K–dependent proteins, proteins C and S, and Ca^{2+} initiate a process to form thrombin that then catalyzes the conversion of fibrinogen to a fibrin clot.

Proteins that contain γ-carboxyglutamyl are also abundant in bone tissue, with osteocalcin accounting for up to 80% of the total γ-carboxyglutamyl content of mature bone. Epidemiologic studies have shown an association between low vitamin K intakes and hip fracture risk. Evidence indicates that vitamins K and D may act synergistically in maintaining bone density.

Deficiency

Although vitamin K deficiency in the adult is uncommon, the risk is increased with fat malabsorption states such as (1) bile duct obstruction, (2) cystic fibrosis, and (3) chronic pancreatitis and liver disease. Risk is also increased by the use of drugs

that interfere with vitamin K metabolism, such as the coumarin anticoagulants (e.g., warfarin) and antibiotics containing the *N*-methylthiotetrazole side chain (e.g., cephalosporin). Other at-risk groups are hospitalized patients with poor nutrient intakes or those receiving TPN, when fat-soluble vitamin supplements may not fully meet requirements. Conversely, ingestion of supraphysiologic doses of vitamins A and E has been reported to induce vitamin K deficiency, probably through competitive mechanisms. Defective blood coagulation and demonstration of abnormal noncarboxylated prothrombin are at present the only well-established signs of vitamin K deficiency.

Hemorrhagic disease of the newborn can develop because of (1) poor placental transfer of vitamin K, (2) hepatic immaturity leading to inadequate synthesis of coagulation proteins, and (3) the low vitamin K content of early breast milk. Severe diarrhea and antibiotics used to suppress diarrhea exacerbate the situation, so prothrombin concentrations can drop significantly, resulting in bleeding. This condition is routinely prevented by the prophylactic administration of phylloquinone intramuscularly immediately after birth.

Toxicity

The use of high doses of naturally occurring vitamin K (K_1 and K_2) appears to have no known toxic effect; however, menadione (K_3) treatment can lead to the formation of erythrocyte cytoplasmic inclusions known as Heinz bodies and hemolytic anemia.

Laboratory Assessment of Status

Because of its relatively low plasma concentration, vitamin K has long presented an analytical challenge. For this reason, vitamin K status has traditionally been assessed by functional methods, primarily by its effect on clotting time. The *prothrombin time (PT)* is assessed by adding a portion of tissue thromboplastin to recalcified plasma and measuring the clotting time against a normal control sample. In vitamin K deficiency, the PT may rise above 30 seconds (normal, 10 to 14 seconds), and at least 2 seconds beyond the control time. Attempts at cross-laboratory standardization led to the introduction of the international normalized ratio (INR), by which PT can be expressed as a fraction of the control time.

A more sensitive assessment of vitamin K status with respect to prothrombin can be made by the immunoassay of des-γ-carboxy prothrombin, or undercarboxylated prothrombin, PIVKA-II (protein induced by vitamin K absence or antagonism). PIVKA-II has proved to be a useful marker of subclinical vitamin K deficiency. Another measurement of deficient γ-carboxylation, plasma undercarboxylated osteocalcin, has been shown to correlate individually with PIVKA-II and plasma phylloquinone concentrations and has a better correlation with plasma phylloquinone than PIVKA-II.

Direct measurement of plasma phylloquinone is probably the best indicator of vitamin K status and has been shown to correlate with intake. HPLC methods are the mainstay of vitamin K measurement. HPLC-mass spectrometry methods have also been developed.

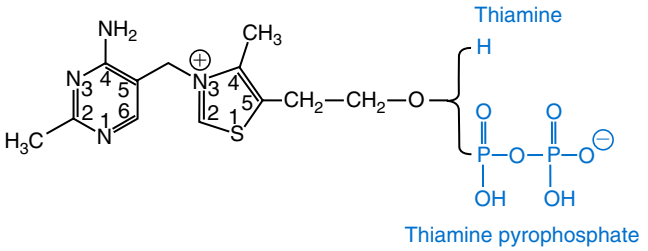


Fig. 27.6 Thiamine and the pyrophosphate coenzyme.

Vitamin K is light-sensitive, and it is recommended samples should be protected from light.

Vitamin B_1

The structure of *thiamine* (3-[4-amino-2-methyl-pyrimidyl-5-methyl]-4-methyl-5-[β-hydroxyethyl]thiazole) is that of a pyrimidine ring, bearing an amino group, linked by a methylene bridge to a thiazole ring (Fig. 27.6). The thiazole has a primary alcohol side chain at C5, which can be phosphorylated in vivo to produce thiamine phosphate esters, the most common of which is thiamine pyrophosphate (TPP; also known as thiamine diphosphate [cocarboxylase]). Monophosphate and triphosphate esters also occur.

Absorption, Transport, Metabolism, and Excretion

Thiamine absorption occurs primarily in the proximal small intestine by a saturable (thiamine transporter) process at low concentration (1 μmol/L or lower) and by simple passive diffusion beyond that, although percentage absorption diminishes with increased dose. Absorbed thiamine undergoes intracellular phosphorylation, mainly to the pyrophosphate, but at the serosal side, 90% of transferred thiamine is present in the free form. Thiamine uptake is enhanced by thiamine deficiency and is reduced by thyroid hormone, diabetes, and ethanol ingestion. The free vitamin is present in plasma, but the **coenzyme**, TPP, is the primary cellular component. Approximately 30 mg is stored in the body, with 80% as pyrophosphate, 10% as triphosphate, and the rest as thiamine and its monophosphate. About half of body stores are found in skeletal muscle, with much of the remainder in heart, liver, kidneys, and nervous tissues (including the brain, which contains most of the triphosphate).

Functions

Thiamine is required by the body as the pyrophosphate (TPP) in two general types of reactions: (1) the oxidative decarboxylation of 2-oxo acids catalyzed by dehydrogenase complexes, and (2) the formation of α-ketols (ketoses) as catalyzed by transketolase and as the triphosphate (TTP) within the nervous system. TPP functions as the Mg^{2+}-coordinated coenzyme for so-called active aldehyde transfers in multienzyme dehydrogenase complexes that affect decarboxylative conversion of α-keto (2-oxo) acids to acyl-coenzyme A (acyl-CoA) derivatives, such as pyruvate dehydrogenase and α-ketoglutarate dehydrogenase. These are often localized in the mitochondria, where efficient use in the Krebs tricarboxylic acid (citric acid) cycle follows.

Transketolase is a TPP-dependent enzyme found in the cytosol of many tissues, especially liver and blood cells, in which principal carbohydrate pathways exist. In the pentose phosphate pathway, which additionally supplies reduced nicotinamide-adenine dinucleotide phosphate (NADPH) necessary for biosynthetic reactions, this enzyme catalyzes the reversible transfer of a glycoaldehyde moiety from the first two carbons of a donor ketose phosphate to the aldehyde carbon of an aldose phosphate. TTP is also thought to be involved in the regulation of ion channels—specifically chloride channels of large unitary conductance, the so-called maxi-Cl channels.

Deficiency

Causes of thiamine deficiency include inadequate intake caused by diets largely dependent on milled, nonenriched grains such as rice and wheat, or the ingestion of raw fish containing thiaminases, which hydrolytically destroy the vitamin in the gastrointestinal tract. Tea may also contain antithiamine factors. Chronic alcoholism often leads to thiamine deficiency caused by reduced intake, impaired absorption, impaired use, and reduced storage, and may lead clinically to the *Wernicke-Korsakoff syndrome*. **Wernicke encephalopathy** classically presents with ophthalmoplegia, ataxia, and confusion. Other at-risk groups include those receiving parenteral nutrition without adequate thiamine supplementation, elderly patients taking diuretics, and patients undergoing long-term renal dialysis.

Beriberi (origin: Sinhalese from a word meaning weakness) is the disease resulting from thiamine deficiency. Clinical signs of thiamine deficiency primarily involve the nervous and cardiovascular systems. In the adult, symptoms most frequently observed include mental confusion, anorexia, muscular weakness, ataxia, peripheral paralysis, ophthalmoplegia, edema *(wet beriberi)*, muscle wasting *(dry beriberi)*, tachycardia, and an enlarged heart. In infants, symptoms appear suddenly and severely, often involving cardiac failure and cyanosis. Commonly, the distinction between wet (cardiovascular) and dry (neuritic) manifestations of beriberi relate to duration and severity of the deficiency, the degree of physical exertion, and caloric intake. The wet or edematous beriberi results from severe physical exertion and high carbohydrate intake, whereas the dry or polyneuritic beriberi stems from relative inactivity with caloric restriction during the chronic deficiency. Nervous system involvement includes peripheral neuropathy, Wernicke encephalopathy, and the amnesic psychosis of Korsakoff syndrome. More rarely, but especially in seriously ill patients in hospitals, an acute form of cardiac failure has been described *(Shoshin beriberi)*, which may be fatal but can be successfully and rapidly reversed with high-dose intravenous thiamine.

Toxicity

No reports have described adverse effects from consumption of excess thiamine from food and supplements.

Laboratory Assessment of Status

The free or phosphorylated forms of thiamine can be measured directly in a suitable body fluid or tissue, or its properties as an enzymatic cofactor can be exploited in a functional assay.

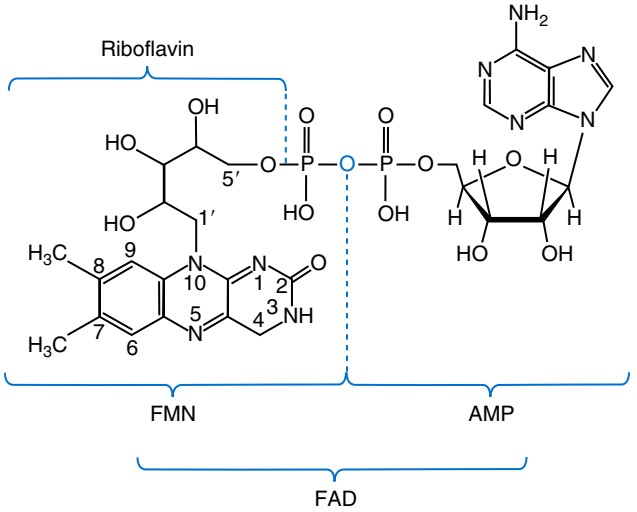


Fig. 27.7 Riboflavin and flavin mononucleotide *(FMN)* as components of flavin adenine dinucleotide *(FAD)*.

The most commonly used enzyme for the functional assay is transketolase. Transketolase catalyzes two reactions in the pentose phosphate pathway. As an enzyme within the erythrocyte, transketolase is independent of nonspecific changes in the extracellular plasma. As vitamin B_1 deficiency becomes more severe, (1) thiamine becomes limiting in the body cells, (2) the amount of the coenzyme is depleted, and (3) transketolase activity subsequently diminishes. The *TPP effect* measures the extent of depletion of the transketolase enzyme for coenzyme by assaying enzyme activity before and after TPP supplementation.

Circulating thiamine concentration may be directly measured in plasma, erythrocytes, or whole blood. The plasma (or serum) concentration is thought to reflect recent intake and is mainly unphosphorylated thiamine at low concentration (around 10 to 20 nmol/L). Because the erythrocyte contains approximately 80% of the total thiamine content of whole blood (mainly as the pyrophosphate) and erythrocyte thiamine stores deplete at a similar rate to other major organs, HPLC measurement of TPP in erythrocytes is a good indicator of body stores. Whole blood samples may be analyzed in a similar manner to washed erythrocytes and may provide the advantage of simpler sample handling, but they are subject to variable plasma dilution. However, a good correlation has been obtained between erythrocyte and whole blood TPP concentrations, particularly when whole blood TPP included a correction for hemoglobin (Hb).

Vitamin B_2

Riboflavin, commonly known as vitamin B_2, is the precursor of all biologically important flavins, notably flavin mononucleotide, FMN (riboflavin-5′-phosphate) and flavin adenine dinucleotide, FAD (Fig. 27.7). FMN is formed from riboflavin by flavokinase-catalyzed phosphorylation, and FAD is formed from FMN and ATP by the action of FAD synthetase, also called *pyrophosphorylase*. FAD is further converted by covalent bonding to form various tissue flavoproteins.

Absorption, Transport, Metabolism, and Excretion

Most dietary riboflavin is taken in as a complex of proteins with the coenzymes FMN and FAD. These coenzymes are released from noncovalent attachment to proteins as a consequence of gastric acidification. Nonspecific action of pyrophosphatase and phosphatase on the coenzyme occurs in the upper gut. The vitamin is primarily absorbed in the proximal small intestine by a saturable active transport system. Bile salts appear to facilitate uptake, and a modest amount of the vitamin circulates via the enterohepatic system. The transport of flavins in human blood involves loose binding to albumin and tight binding to numerous immunoglobulins. Pregnancy increases the concentration of carrier protein for riboflavin, which results in a higher rate of riboflavin uptake at the maternal surface of the placenta. Uptake of riboflavin into the cells of organs such as liver is facilitated, possibly requiring a specific carrier at physiologic concentrations, but it can occur by diffusion at higher concentrations.

Functions

Riboflavin and its coenzyme derivatives are involved in a large variety of chemical reactions. These derivatives are capable of one- and two-electron transfer processes and play a pivotal role in coupling the two-electron oxidation of most organic substrates to the one-electron transfer of the respiratory chain, thus being involved in energy production. They also function as electrophiles and nucleophiles, with covalent intermediates of flavin and substrate frequently involved in catalysis. Flavoproteins catalyze dehydrogenation reactions, hydroxylations, oxidative decarboxylations, deoxygenations, and reductions of oxygen to hydrogen peroxide. Other major functions of riboflavin include drug metabolism in conjunction with the cytochrome P450 enzymes and lipid metabolism. Flavins also have pro-oxidative and antioxidative functions. They are thought to contribute to oxidative stress through their ability to produce superoxide and to catalyze the production of hydrogen peroxide. As an antioxidant, FAD is a coenzyme to glutathione reductase in the regeneration of reduced glutathione from oxidized glutathione, which is necessary for the removal of lipid peroxides. Riboflavin deficiency is associated with increased lipid peroxidation. FAD is a cofactor to methylenetetrahydrofolate reductase (MTHFR) in the remethylation of homocysteine.

Deficiency

Although riboflavin has a wide distribution in foodstuffs, many people live for long periods on low intakes; consequently, minor signs of deficiency are common in many parts of the world. In addition to poor intake, other causes of deficiency include hypothyroidism and adrenal insufficiency, which inhibit the conversion of riboflavin to its coenzyme derivatives; drugs such as chlorpromazine, imipramine, and amitriptyline, which have a similar tricyclic structure to riboflavin; the anticancer drug doxorubicin; and the antimalarial quinacrine. Excess ethanol ingestion interferes with both digestion and absorption of riboflavin.

Because flavin coenzymes are widely distributed in intermediary metabolism, the consequences of deficiency may be widespread. Riboflavin coenzymes are involved in the metabolism of vitamin B_{12} and folic acid (irreversible reduction of 5,10-methylenetetrahydrofolate to 5-methyltetrahydrofolate) and therefore are a determinant of plasma homocysteine concentration; pyridoxine (PN; conversion to pyridoxal 5-phosphate [PLP]); and niacin (conversion of 5-hydroxytryptamine to tryptophan that is required for niacin synthesis). Therefore deficiency will affect enzyme systems other than those requiring flavin coenzymes per se.

The deficiency syndrome is characterized by (1) sore throat; (2) hyperemia; (3) edema of the pharyngeal and oral mucous membranes; (4) cheilosis; (5) angular stomatitis; (6) glossitis (magenta tongue); (7) seborrheic dermatitis; and (8) normochromic, normocytic anemia associated with pure red blood cell (RBC) aplasia of the bone marrow.

Toxicity

Probably as a result of its limited solubility and limited gastric absorption, no adverse effects have been associated with ingestion of riboflavin appreciably above RDA amounts.

Laboratory Assessment of Status

Riboflavin status can be assessed by (1) determination of urine riboflavin excretion, (2) by a functional assay measuring the erythrocyte glutathione reductase activation coefficient (EGRAC), which is the ratio between enzyme activity determined with and without the addition of the cofactor, FAD, or (3) direct measurement of riboflavin or its metabolites in plasma or erythrocytes.

Riboflavin status is commonly assessed by the determination of FAD-dependent glutathione reductase activity in freshly lysed erythrocytes. This enzyme-based assay has been chosen for most surveys of riboflavin status. Most methods measure the rate of change of absorbance at 340 nm caused by the oxidation of NADPH and have been automated to give rapid throughputs and CVs of less than 2% within the run, although some have used fluorescence detection with increased sensitivity. Direct measurement of riboflavin, FMN, and FAD in plasma or erythrocytes is undertaken by HPLC, usually with fluorescence detection after protein precipitation, or by capillary zone electrophoresis with laser-induced fluorescence detection (CZE-LIF). In critically ill patients, plasma FAD, which is bound to albumin, may fall as a result of the systemic inflammatory response and due to redistribution of FAD, as there is increased tissue requirement whereas red cell FAD is unaffected. Therefore the measurement of red cell FAD is more sensitive and recommended, especially in critically ill patients. As vitamin B_2 is light-sensitive, samples should be protected from light.

Vitamin B_6

The vitamin B_6 group comprises three natural forms, *pyridoxine* (pyridoxol; *PN*), *pyridoxamine (PM)*, and *pyridoxal (PL)*, which are 4-substituted 2-methyl-3-hydroxyl-5-hydroxymethyl pyridines (Fig. 27.8). During metabolic conversion, each vitamer

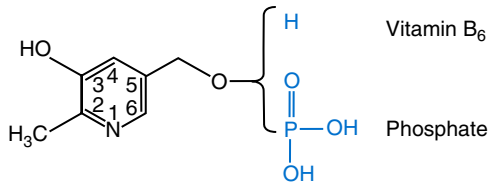


Fig. 27.8 Free and phosphorylated forms of vitamin B_6. R = CH_2OH for pyridoxine, CH_2NH_2 for pyridoxamine, and CHO for pyridoxal.

becomes phosphorylated at the 5-hydroxymethyl substituent. PLP is the coenzyme form that participates in a large number of B_6-dependent enzyme reactions.

Absorption, Transport, Metabolism, and Excretion

The phosphorylated sources are hydrolyzed by the intraluminal action of intestinal alkaline phosphatase to enable cellular uptake. The nonphosphorylated forms are readily absorbed by the mucosal cells through a process of passive diffusion. In other cells requiring vitamin B_6, the unphosphorylated vitamers may be "metabolically trapped" as phosphorylated forms by cytoplasmic PL kinase, which is responsible for catalyzing the ATP-dependent phosphorylation of all three vitamin forms. Transport to the liver via the portal vein is done by the unphosphorylated form.

Release of free vitamin, mainly PL, occurs when physiologic nonsaturating concentrations of vitamin are absorbed. Here the phosphates are hydrolyzed by nonspecific alkaline phosphatase located on the plasma membrane of cells. Some PLP is released into the circulation by the liver and circulates complexed to proteins—mostly albumin. PLP is the principal tissue form of vitamin B_6, whereas PL constitutes much of the circulating vitamin. The main catabolite excreted in urine is 4-pyridoxic acid (4-PA), which is formed by the action of the FAD-dependent general liver aldehyde oxidase.

Functions

As coenzyme PLP, vitamin B_6 functions in more than 100 reactions that embrace the metabolism of macronutrients, such as proteins, carbohydrates, and lipids. These enzymes include aminotransferases; decarboxylases that lead to the formation of various functional amines, including epinephrine and norepinephrine; phosphorylases; and cystathionine β-synthase in the trans-sulfuration pathway of homocysteine. The biosynthesis of heme depends on the early formation of 5-aminolevulinate from PLP-dependent condensation of glycine and succinyl-CoA, followed by decarboxylation. In lipid metabolism, PLP-dependent condensation of L-serine with palmitoyl-CoA forms 3-dehydrosphinganine, a precursor of sphingomyelins. Therapeutically, vitamin B_6 has been used for the treatment of some intractable seizures in neonates and infants and for the treatment of other vitamin B_6–responsive inborn errors of metabolism and the carpal tunnel syndrome.

Deficiency

Vitamin B_6 deficiency in isolation is rare; it is more usual in association with deficits in other vitamins of the B-complex. Deficiency has been observed with the antituberculosis drug isoniazid (isonicotinic acid hydrazide). Penicillamine (β-dimethyl cysteine), used in the treatment of patients with Wilson disease, inactivates PLP by forming a thiazolidine derivative. Other drugs that can cause vitamin B_6 deficiency include the antiparkinsonian drugs benserazide and carbidopa, which react by forming hydrazones, and theophylline.

Several vitamin B_6–responsive inborn errors of metabolism are known, including homocystinuria, xanthurenic aciduria, primary cystathioninuria, and pyridoxine-responsive epilepsy. Low vitamin B_6 status (together with low vitamin B_{12} and folate status) in humans has been linked to hyperhomocysteinemia and is an independent risk factor for cardiovascular disease, although clinical trials have been inconclusive.

Toxicity

Although no adverse effects have been observed with high intakes of vitamin B_6 from food sources, high oral supplemental doses have been found to have neurotoxic and photosensitive effects. Based on the end point of development of sensory neuropathy, a tolerable upper intake amount of 100 mg/day for adults is recommended.

Laboratory Assessment of Status

As with the other B vitamins that act as coenzymes, biochemical assessment of vitamin B_6 can be made by direct chemical analysis of the vitamer or its metabolites, or by functional means. Measurements that have been used are PLP in plasma or red cells, its metabolite 4-PA in urine or plasma, the activity and activation coefficient of the red cell aminotransferases (aspartate and alanine), and the tryptophan load metabolite excretion test. Because no single marker adequately reflects status, a combination of these markers offers the best approach.

Plasma PLP and plasma or urine 4-PA are most commonly measured by HPLC, PLP with fluorescence detection following precolumn fluorophore formation as a semicarbazone or a pyridoxic acid phosphate, and 4-PA with its natural fluorescence. Using ion-pair reversed-phase chromatography, plasma vitamin B_6 vitamers (PLP, PL, PN, PMP, PM, and 4-PA) can be measured.

Functional assessment of vitamin B_6 status may be made by measuring the activity of red cell aspartate (or alanine) aminotransferase and its activation coefficient on incubation with PLP, although because the **apoenzyme** is highly unsaturated with PLP, the results obtained have greater variability than those derived by corresponding methods for vitamins B_1 and B_2 and thus are considered less useful.

Vitamin B_6 is light-sensitive, samples should be protected from light.

Vitamin B_{12}

Vitamin B_{12} is one of the most structurally complex small molecules produced by nature, and the only known carbon-metal bond (involving cobalt) found in a biologically active molecule. The generic term *vitamin B_{12}* refers to a group of physiologically active substances chemically classified as cobalamins or corrinoids. They are composed of tetrapyrrole

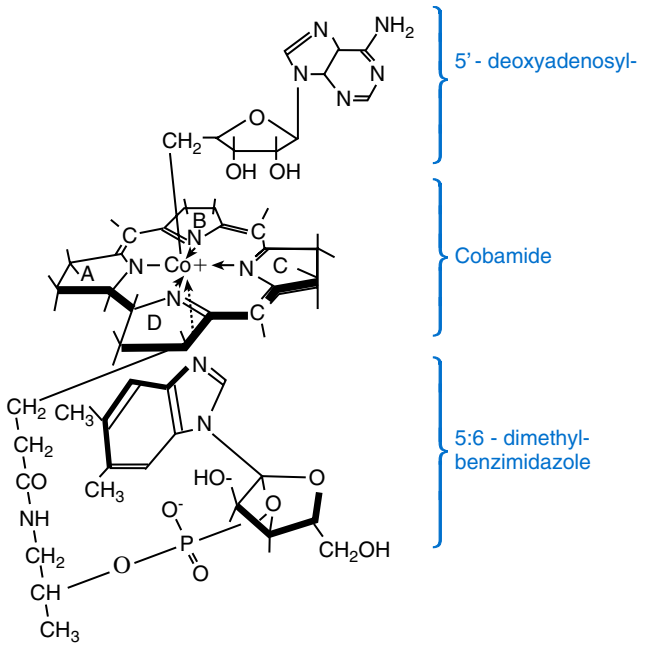


Fig. 27.9 The structure of 5′-deoxyadenosyl cobalamin.

rings surrounding central cobalt atoms and nucleotide side chains attached to the cobalt atoms. The cobalamin tetrapyrrole ring, exclusive of cobalt and other side chains, is called a *corrin*. All compounds containing this corrin nucleus are corrinoids. The cobalt-corrin complex is termed *cobamide* (Fig. 27.9). Cobalamins differ in the nature of additional side groups bound to cobalt. Examples include methyl (methylcobalamin), 5′-deoxyadenosine (deoxyadenosyl [short form, adenosyl], cobalamin, or coenzyme B_{12}), hydroxyl (hydroxocobalamin), H_2O (aquocobalamin, or vitamin B_{12b}), and cyanide (cyanocobalamin).

Absorption, Transport, Metabolism, and Excretion

Vitamin B_{12} is known to be tightly bound to proteins and has to be released from food by acid and the enzyme pepsin present in the stomach. The synthetic form is free and therefore readily available for metabolism. Thus acid-blocking drugs such as proton pump inhibitors (e.g., omeprazole) can cause vitamin B_{12} deficiency. The free vitamin B_{12} molecule is bound to haptocorrin (HC, R protein) and travels with it into the duodenum, where the HC is digested by pancreatic enzymes. Liberated vitamin B_{12} then binds to intrinsic factor (IF), a glycoprotein with a molecular weight of approximately 50 kDa that is produced by the gastric parietal cells in the fundus and body of the stomach. One molecule of IF binds one molecule of vitamin B_{12}. Gastric secretion of IF is stimulated by food, histamine, and gastrin produced by the antrum portion of the stomach; it is inhibited by vagal blockade. When the vitamin B_{12}–IF complex reaches the distal ileum, it is bound by specific receptors known as the cubam complex, which consists of two subunits, namely cubilin and amnionless on the surface of mucosal epithelial cells, and it is internalized. The vitamin B_{12}–IF complex is dissociated within the mucosal epithelial cells by lysosomes and released into blood.

In circulation, vitamin B_{12} then binds to two circulating binding proteins, HC (referred to as holohaptocorrin when bound to vitamin B_{12} and previously known as transcobalamin [TC] I) and TC (holotranscobalamin when bound to vitamin B_{12} and previously known as TC II); however, it is only the TC-bound fraction for which there is a receptor-mediated cellular uptake as described above and is therefore the bioactive fraction. The function of HC that is released by granulocytes is currently unknown, and low haptocorrin concentrations, found in approximately 15% of persons with low serum cobalamin, could be one of the most common causes of low cobalamin concentrations. HC accounts for 80% to 94% of endogenous plasma vitamin B_{12}, whereas TC accounts for the remainder, but the latter is the more important vitamin B_{12} transport protein in plasma. TC transports vitamin B_{12} to receptors on cell membranes throughout the body. The TC–vitamin B_{12} complex enters the cell by pinocytosis. Lysosomal proteolysis degrades TC and releases the vitamin B_{12}. Unbound vitamin B_{12} can enter the tissue cells, but the process is very inefficient.

As discussed above, only vitamin B_{12} bound to TC is capable of cellular uptake and as a result malabsorption of TC-bound vitamin B_{12} leads to deficiency in a large proportion of cases. Almost all vitamin B_{12} (bound to TC) is taken up by hepatocytes as the blood in the portal vein passes through the liver where it is stored, and released into blood to meet physiologic demands. If the quantity of vitamin B_{12} exceeds the capacity of hepatocyte receptors, most of the excess is excreted by the kidneys. Normally, approximately 1 mg of vitamin B_{12} is stored in the liver—a quantity equivalent to the daily metabolic requirement for 2000 days. Thus when the dietary supply of vitamin B_{12} is interrupted or mechanisms of absorption are impaired, vitamin B_{12} deficiency does not become evident for several years.

Vitamin B_{12} is continually secreted in the bile, but most is reabsorbed and is available for metabolic functions. If circulating vitamin B_{12} concentrations exceed the binding capacity of the blood, the excess will be excreted in the urine, but in most circumstances, the highest losses of vitamin B_{12} occur through the feces.

Functions

Vitamin B_{12} is a cofactor or coenzyme for various enzyme systems. In humans it is required in both adenosylcobalamin and methylcobalamin. Adenosylcobalamin is a coenzyme to L-methylmalonyl-CoA mutase in the conversion of L-methylmalonyl CoA to succinyl-CoA. The conversion of L-methylmalonyl-CoA to succinyl-CoA links propionyl-CoA, which is formed from branched-chain amino acids such as valine, isoleucine, and methionine with the tricarboxylic acid (TCA) cycle. Congenital defects of mutase synthesis or the inability to synthesize adenosylcobalamin results in life-threatening methylmalonic aciduria and metabolic ketoacidosis. Methylcobalamin is a coenzyme to methionine synthase in the conversion of homocysteine to methionine. In this reaction (Fig. 27.10), methylcobalamin serves as an intermediate in the transfer of a methyl group from 5-methyltetrahydrofolate to homocysteine

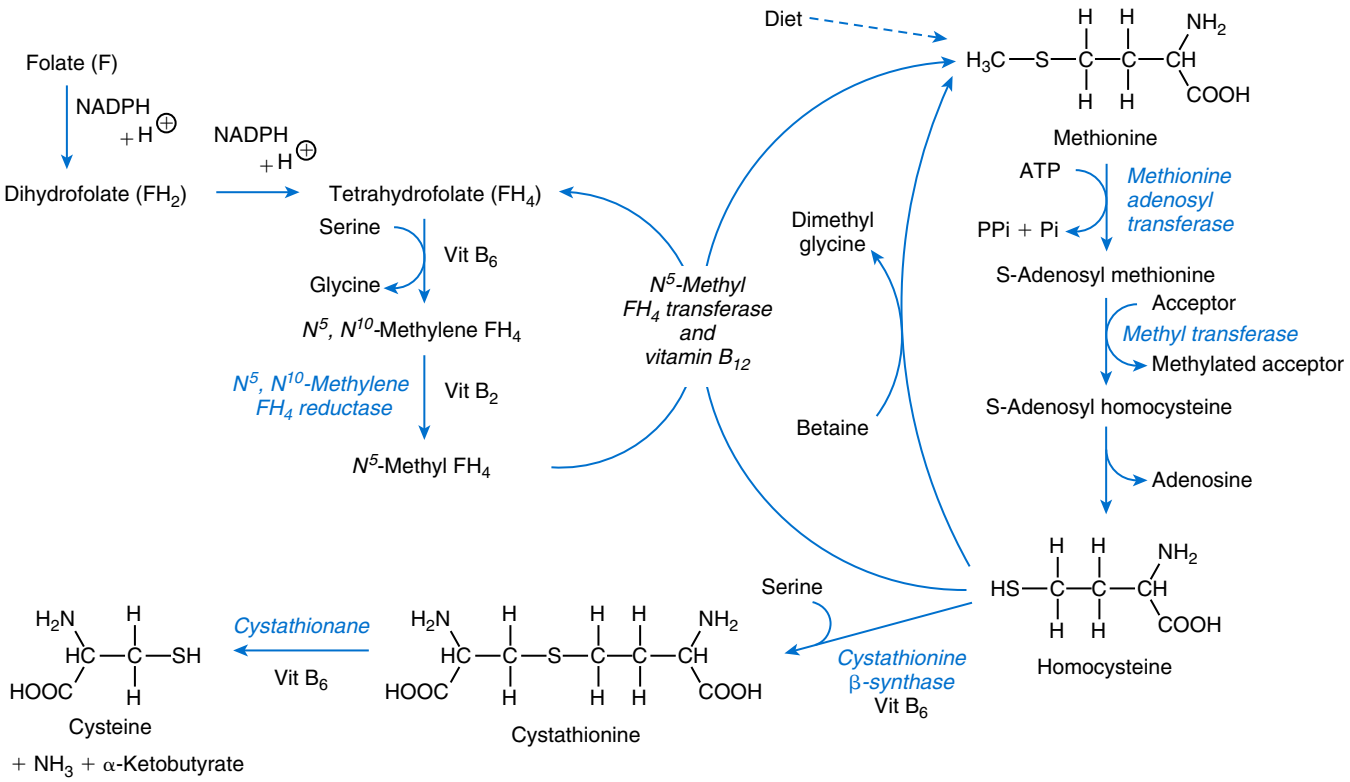


Fig. 27.10 Metabolism of homocysteine and methionine. *NADPH,* Nicotinamideadenine dinucleotide phosphate.

for the formation of methionine. Methionine is required for protein synthesis and as the methyl donor, *S*-adenosylmethionine. Congenital defects in methionine synthase or the synthesis of methylcobalamin result in severe hyperhomocysteinemia. As discussed under the section on vitamin B_6, defects in the enzyme cystathionine β-synthase or its cofactor vitamin B_6 also results in hyperhomocysteinemia.

Deficiency

Deficiency of vitamin B_{12} in humans is associated with macrocytosis, megaloblastic anemia, and neuropathy. The most common cause of vitamin B_{12} deficiency is **pernicious anemia**, an autoimmune disease in which chronic atrophic gastritis in the fundus and body region of the stomach results from autoantibodies to gastric parietal cells directed against gastric parietal cell H^+/K^+-ATPase, which is responsible for secreting acid (H^+) in exchange for potassium (K^+). Loss of parietal cells also leads to decreased production of IF. In addition to IF deficiency, blocking autoantibodies that bind the Vitamin B_{12}-binding sites of IF prevent the formation of Vitamin B_{12}-IF complex required for recognition by the cubam complex in the distal ileum. At-risk groups include (1) those older than 65 years of age; (2) those with malabsorption; (3) vegetarians or vegans; (4) those with autoimmune disorders (pernicious anemia usually occurs as part of the autoimmune polyglandular syndrome type 3B that includes autoimmune thyroiditis); and (5) those taking prescribed medication known to interfere with vitamin absorption or metabolism, including nitrous oxide (also known as laughing gas that inactivates vitamin B_{12} by oxidizing its cobalt atom), phenytoin, dihydrofolate reductase inhibitors, metformin, and proton pump inhibitors; as well as (6) infants with suspected metabolic disorders.

Intestinal malabsorption of vitamin B_{12} may be caused by gastrectomy or ileal resection, with an inverse relationship noted between the length of ileum resected and absorption of vitamin B_{12}. Other causes of malabsorption include tropical sprue, inflammatory disease of the small intestine such as celiac disease, intestinal stasis with overgrowth of colonic bacteria, which consume vitamin B_{12} ingested by the host, and human immunodeficiency virus (HIV) infection. Another cause of vitamin B_{12} malabsorption is failure to extract cobalamin from food. This is particularly a problem in patients with compromised gastric status or early in the course of development of pernicious anemia.

Vegetarians have a lower intake of vitamin B_{12} than omnivores, and although clinical signs of deficiency are uncommon, biochemical markers of status indicate functional vitamin B_{12} deficiency.

A large number of disorders are associated with cobalamin deficiency in infancy or childhood. Of these, the most commonly encountered is the **Imerslund-Grasbeck syndrome**, a condition that is characterized by inability to absorb vitamin B_{12}, with or without IF, and proteinuria. It appears to be due to an inability of intestinal mucosa to absorb the vitamin B_{12}–IF complex as a result of mutations in cubilin or amnionless. The second most common of these is congenital deficiency of gastric secretion of IF. Very rarely, congenital deficiency of vitamin B_{12} in a breast-fed infant is due to deficiency of vitamin B_{12} in maternal breast milk resulting from unrecognized

pernicious anemia in the mother. This is rare because most women with undiagnosed and untreated pernicious anemia are infertile. Methylmalonic acidemias (acidurias) and homocysteinemias may be due to vitamin B_{12} deficiency and, depending on the underlying mutation, may or may not be responsive to supplementation with the vitamin.

The hematologic effects of vitamin B_{12} deficiency are indistinguishable from those of folate deficiency. Classical morphologic changes in the blood, in approximate order of appearance, are as follows: hypersegmentation of neutrophils, macrocytosis, anemia, leukopenia, and thrombocytopenia, with megaloblastic changes in bone marrow accompanying peripheral blood changes.

In addition to hematologic changes, vitamin B_{12} deficiency can lead to a demyelinating disorder of the central nervous system in man. Serious and often irreversible neurologic disorders can occur, such as burning pain or loss of sensation in the extremities, weakness, spasticity and paralysis, confusion, disorientation, and dementia. This condition has been given the name *subacute combined degeneration of the spinal cord*. Neurologic symptoms may occur without any discernible hematologic changes in the blood; indeed an intriguing inverse relationship between the hematologic and the neurologic has been observed. Vitamin B_{12} deficiency may be associated with other mainly gastrointestinal complications, such as glossitis of the tongue, appetite and weight loss, flatulence and constipation, mental changes, and infertility.

Depending on the laboratory and the procedure used, reference intervals vary widely. A recent WHO consultation defined a serum vitamin B_{12} concentration less than 203 ng/L (150 pmol/L) as deficient. Vitamin B_{12} concentrations within the reference interval may not necessarily reflect adequate vitamin B_{12} status, because serum concentrations may be maintained at the expense of tissue stores. Conversely, low serum vitamin B_{12} concentrations may not be indicative of vitamin B_{12} deficiency.

Toxicity

No adverse effects have been associated with excess vitamin B_{12} intake from food or supplements in healthy people.

Laboratory Assessment of Status

Both direct and indirect (functional) methods are available for assessment of vitamin B_{12} status. Indirect tests include assays for urinary and serum concentrations of methylmalonic acid (MMA), plasma homocysteine, the deoxyuridine suppression test, and the now rarely used vitamin B_{12} absorption (Schilling) test. Cytochemical staining of RBC precursors and the test for IF blocking antibodies are other ancillary methods of assessing vitamin B_{12} status.

Multiple automated and semiautomated systems are available for measuring vitamin B_{12} using, for example, chemiluminescence as a signal. Most immunometric methods use solid-phase separation by immobilizing the IF antibodies on beads or magnetic particles. Spurious elevations of vitamin B_{12} have been reported in patients with pernicious anemia when using automated analyzers based on the competitive binding of serum vitamin B_{12} with reagents using IF. This has been attributed to the high levels of IF-blocking autoantibodies in these patients interfering in the assay. Up to 70% of patients with pernicious anemia have IF-blocking antibodies. Assay manufacturers are aware of this problem and have taken steps to inactivate the IF-blocking antibodies. However, at the present time the scale of the problem is unknown, and alternative methods should be used if in doubt, including the measurement of IF-blocking antibodies.

Indirect tests assess the functional adequacy of vitamin B_{12}. Serum MMA concentration is increased when lack of adenylcobalamin causes a block in the conversion of methylmalonyl-CoA to succinyl-CoA. It is a sensitive test of status, being often the first analyte to be raised in subclinical vitamin B_{12} deficiency. It has a further advantage in that it is unaffected by folate deficiency. Unfortunately MMA is increased by renal insufficiency and in persons above the age of 65 years. Plasma total homocysteine concentration is also a sensitive indicator of vitamin B_{12} status because methylcobalamin is required for the remethylation of homocysteine to methionine, but it is not specific, being elevated in deficiencies of folate, vitamin B_{12}, vitamin B_2, and vitamin B_6. There are several preanalytical variables affecting it: high protein meal and venous stasis may increase it; it is decreased in the supine position as it is bound to albumin. EDTA plasma is the most widely recommended sample type and should be centrifuged within 1 hour or kept cold by collecting on ice until centrifugation.

The measurement of holotranscobalamin is potentially useful as a specific marker of biologically available vitamin B_{12}, because it is the only vitamin B_{12} moiety that is specifically available for uptake by all cells and has been shown to have the best diagnostic accuracy for vitamin B_{12} deficiency. However, assays for holotranscobalamin are not widely available at present, and their sensitivity and specificity in the diagnosis of vitamin B_{12} deficiency remains to be established.

Folic Acid

Vitamin B_9, folate, and folic acid are generic terms for a family of compounds that function as coenzymes in the processing of one-carbon units and that are derived from pteroic acid, to which one or more molecules of glutamic acid are attached. Pteroic acid is composed of a pteridine ring joined to a *p*-aminobenzoic acid residue (Fig. 27.11). When pteroic acid is conjugated with one molecule of L-glutamic acid, pteroylglutamic acid is formed; this can be reduced to dihydrofolic acid (DHF) with hydrogens in positions 7 and 8, or to tetrahydrofolate (THF) with hydrogens in positions 5, 6, 7, and 8. Only the reduced forms are biologically active. Other folate derivatives have multiple glutamic acid residues ranging from 1 to 7. As a summary, folate is converted by specific enzymes to dihydrofolate (DHF), which is converted to THF and is in turn converted to methylene-THF and finally to 5-methyl-THF.

Absorption, Transport, Metabolism, and Excretion

Folate is absorbed from dietary sources mainly as reduced methyl- and formyl-tetra-hydropteroylpolyglutamates. The

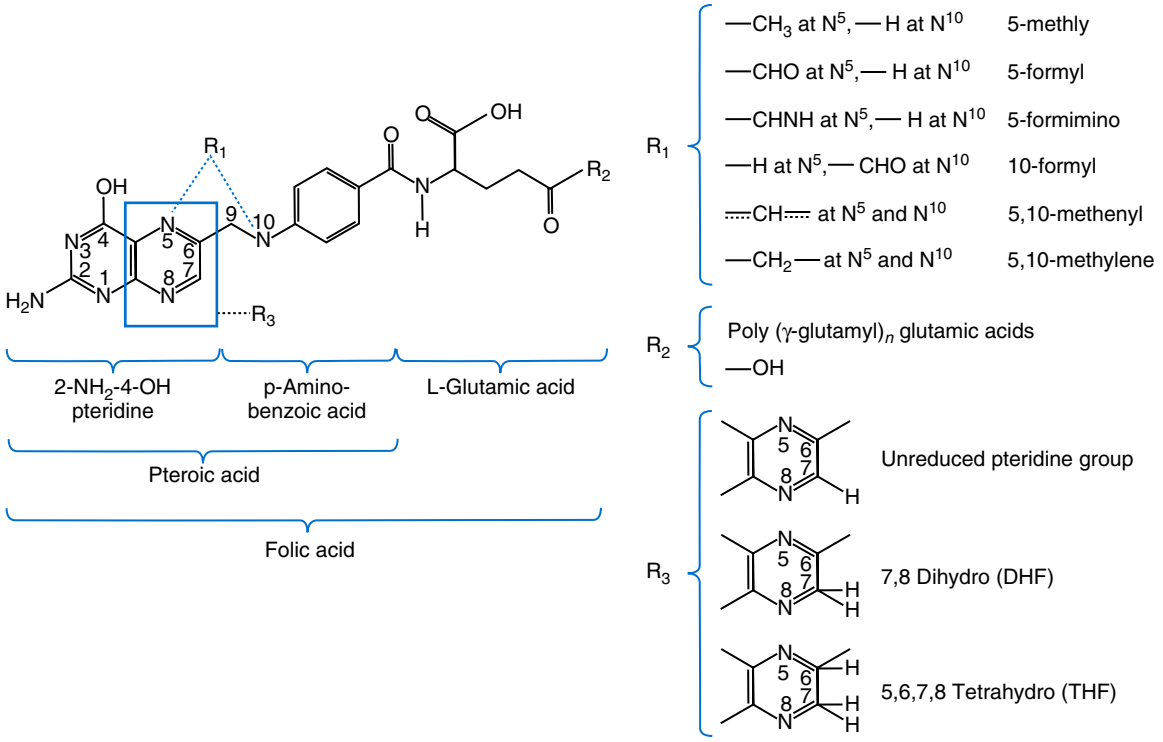


Fig. 27.11 Structure and relationships of folic acid and its derivatives.

bioavailability of folate from food sources is variable and is dependent on factors such as incomplete release from plant cellular structure, entrapment in food matrix during digestion, inhibition of deglutamation by other dietary constituents, and possibly the degree of polyglutamation. Polyglutamate forms of folate present in food are first converted to monoglutamates, by pteroylpolyglutamate hydrolase, in the intestinal mucosa. Absorption of monoglutamyl folates at low concentration occurs through a saturable transport process at an acidic pH of around 5, with an additional, apparently nonsaturable absorption mechanism when intestinal folate concentrations exceed 5 to 10 μmol/L. After cellular uptake, most of the folate is reduced and methylated and enters the circulation as THF, circulating loosely bound to albumin or to a lesser degree to a high-affinity folate-binding protein (FBP). Uptake by certain cells (kidney, placenta, and choroid plexus) occurs by membrane-associated FBPs that act as folate receptors, and the reduced folate carrier, a member of the SLC19 family, facilitates uptake by most tissue. Once THF is demethylated and converted to the polyglutamyl form by folylpolyglutamate synthase, it is retained within the cell. For release into the circulation, the polyglutamates are reconverted to monoglutamates by polyglutamate hydrolase.

Folic acid and vitamin B_{12} metabolism is linked by the reaction that transfers a methyl group from THF to cobalamin. In cases of cobalamin deficiency, folate is "trapped" as methyl-THF and is "metabolically dead." It cannot be recycled as THF back into the folate pool to serve as the main one-carbon unit acceptor for many biochemical reactions. Eventually, cellular depletion of methylene-THF ensues, causing a reduction in thymidylic acid synthesis, which, in turn, results in megaloblastic anemia and neuropathies. This concept is supported by the fact that THF corrects megaloblastic anemia in patients with congenital methylmalonic aciduria and homocystinuria, whereas it is not corrected with methyl-THF.

Protein-free plasma folate is filtered at the glomerulus, and most is reabsorbed by the proximal renal tubules. Most of the biliary excreted folate is reabsorbed in the enterohepatic circulation. Fecal losses are similar to urinary losses.

Functions

Folate coenzymes, together with coenzymes derived from vitamins B_{12}, B_6, and B_2, are essential for one-carbon metabolism. Biochemically, a carbon unit from serine or glycine is transferred to THF to form methylene-THF, which then is (1) used in the synthesis of thymidine (and incorporation into DNA), (2) oxidized to formyl-THF for use in the synthesis of purines (precursors of RNA and DNA), or (3) reduced to methyl-THF, which is necessary for the methylation of homocysteine to methionine. Much of this methionine is converted to *S*-adenosylmethionine, a universal donor of methyl groups to DNA, RNA, hormones, neurotransmitters, membrane lipids, and proteins. Some of these reactions are illustrated in Fig. 27.12.

Deficiency

Deficiency of folate may result from (1) absence of intestinal microorganisms (gut sterilization), (2) poor intestinal absorption (e.g., after surgical resection, in celiac disease or sprue), (3) insufficient dietary intake (including chronic alcoholism), (4) excessive demands (as in pregnancy, liver disease, and malignancies), (5) administration of antifolate drugs (e.g., methotrexate), and (6) anticonvulsant therapy

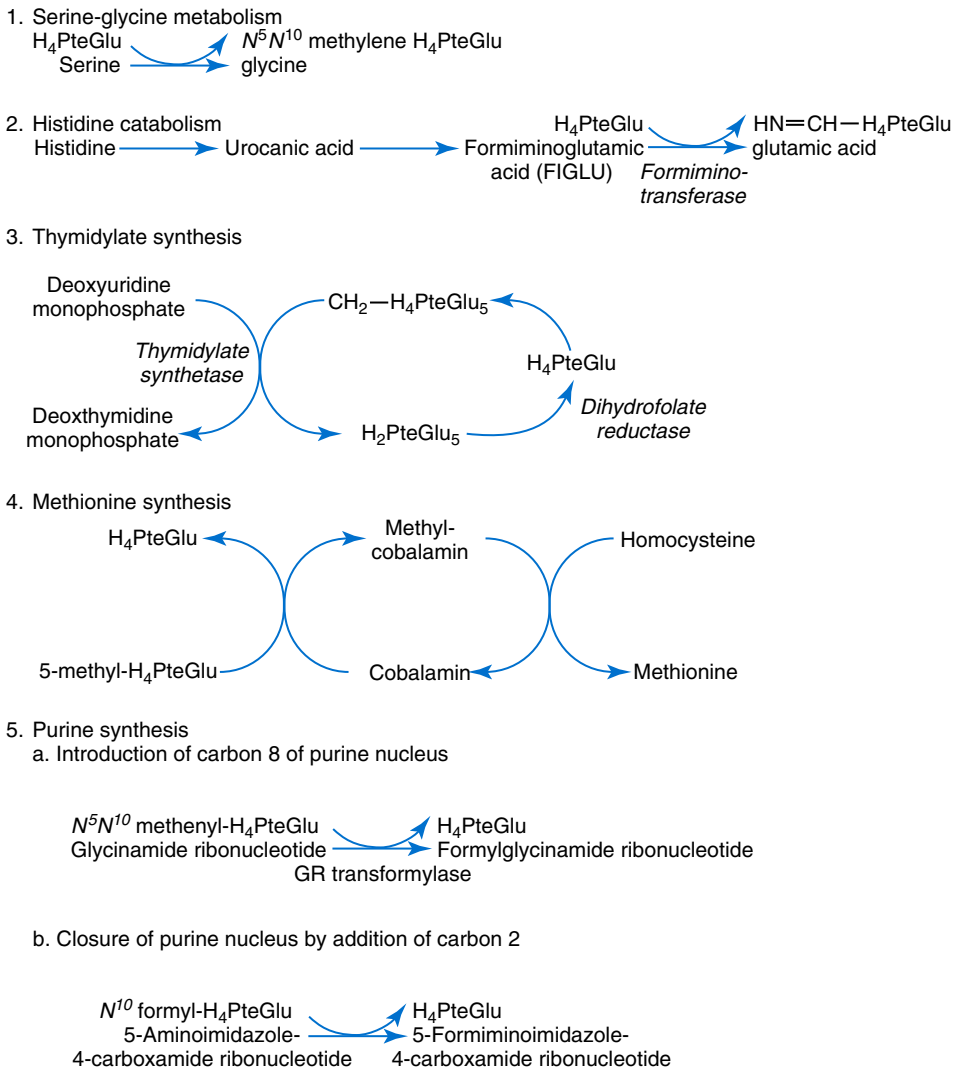


Fig. 27.12 The five major metabolic functions of folate in human cells.

(that increases folate requirements, especially during pregnancy). Inadequate folate intake leads first to decreased serum folate concentration, then to a decrease in RBC folate concentration and an increase in plasma homocysteine, and then to megaloblastic changes in the bone marrow and other tissues. Megaloblastic anemia (characterized by large, abnormally nucleated erythrocytes in the bone marrow) is the major clinical manifestation of folate deficiency, although sensory loss and neuropsychiatric changes may also occur. Deficiencies of folate and iron may coexist in malnourished people, in which case the latter may mask the expected macrocytic and megaloblastic changes.

Pregnancy brings increased demand to folate stores because of increased DNA synthesis, and one-carbon transfer reactions and low serum folate concentrations in pregnancy are associated with adverse outcomes, including preterm delivery, infant low birth weight, and fetal growth retardation. In addition, many observational studies have confirmed a reduction in risk of neural tube defects (NTDs) with periconceptual folic acid supplementation. Current suggestions are that women planning pregnancy should take at least 400 μg/day until the 12th week of pregnancy, although a daily intake of 5 mg of folic acid is recommended, especially for those with a previous history of NTD. Recent meta-analyses have strongly suggested a significant association of the variant 5,10-methyleneterahydofolate reductase (MTHFR) C677T polymorphism with increased risk of NTDs.

The increase in plasma homocysteine concentration has been postulated to be an independent risk factor for coronary artery disease and cerebrovascular disease. The involvement of folate in its coenzyme forms with homocysteine and methionine metabolism is summarized in Fig. 27.10. Folate is the principal micronutrient determinant of homocysteine status, and supplementation with folate (0.5 to 5.0 mg/day) has been used as a treatment modality to reduce circulating homocysteine concentrations. However, the most recent Cochrane meta-analyses of trials of homocysteine-lowering therapy with folate, vitamins B_6 or B_{12} given alone or in combination have not shown a reduction in cardiovascular disease.

Folate appears to have a protective effect on colorectal cancer development but is associated with increased risk of gastric cancer, prostate cancer, and may be associated with lung

cancer. One meta-analysis of 10 randomized clinical trials (RCTs) showed a borderline significant increase in frequency of overall cancer in the folic acid-supplemented group compared with controls.

Toxicity

No adverse effects have been reported from the consumption of folate-fortified foods; thus any signs of toxicity are associated with supplemental folate. Most of the limited evidence suggests that excessive folate supplementation can precipitate or exacerbate neuropathy in vitamin B_{12}–deficient subjects, and it is this end point that has been used to set a tolerable upper intake concentration of 1 mg/day from fortified foods or supplements for adults. One recognized complication of folate supplementation is that it "masks" vitamin B_{12} deficiency, because the associated anemia responds to folate. This may delay treatment of any neurologic abnormalities. It has been recommended that if a low serum folate is present concomitantly with low vitamin B_{12}, the latter should be corrected first.

Laboratory Assessment of Status

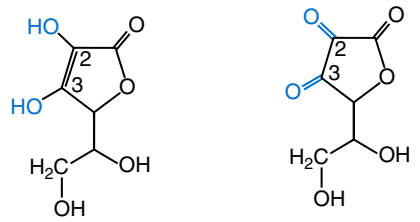
Folate status may be reliably assessed by direct measurement of serum and RBC or whole blood concentrations using immunometric assays, and its metabolic function as a coenzyme may be assessed by metabolite concentrations such as plasma homocysteine. Serum folate concentrations are considered indicative of recent intake and not of tissue stores, but serial measurements have been used to confirm adequate intake. Whole blood and RBC folate concentrations are more indicative of tissue stores over the lifetime of RBCs and are therefore a better indicator of longer term folate status than serum folate. Because folate is taken up only by the developing RBC in the bone marrow and not by the mature cell, RBC concentrations reflect folate status over the 120-day life span of the cell. However, in general serum and RBC folate provide equivalent information regarding folate status.

Plasma homocysteine is considered a sensitive indicator of folate status and is strongly correlated with serum folate concentration in the lower range, which is below 10 nmol/L (4.5 μg/L). A homocysteine concentration above 15 μmol/L is indicative of folate deficiency. However, as the concentration of homocysteine is dependent on age, gender, renal function, genetic factors, and the status of other vitamins (B_6 and B_{12}), it is not a specific marker of folate status.

Due to the instability of folate in serum, it is therefore recommended that if delay in transportation is anticipated, serum should be separated from RBCs by centrifugation and stored frozen until analysis. As there is a significant amount of folate in RBCs, hemolysis invalidates serum folate results.

Vitamin C

The term *vitamin C* refers to all molecules that exhibit antiscorbutic properties (derived from Latin word for scurvy, scorbutus) in humans and includes both ascorbic acid and its oxidized form, dehydroascorbic acid (DHA), as shown in Fig. 27.13.



L-Ascorbic acid Dehydroascorbic acid

Fig. 27.13 L-Ascorbic and dehydroascorbic acids.

Absorption, Transport, Metabolism, and Excretion

Gastrointestinal absorption of ascorbic acid occurs throughout the ileum via a combination of sodium-dependent active transport at low concentrations and simple diffusion at high concentrations. Vitamin C is found in most tissues including the brain, but glandular tissues, such as pituitary, adrenal cortex, corpus luteum, and thymus, have the highest amounts, and the eye lens has 20 to 30 times the plasma concentration. In plasma, vitamin C exists predominantly as the ascorbate ion. Excretion of unchanged ascorbate occurs increasingly with increased dosage, with almost all of an injected dose of more than 500 mg excreted over 24 hours.

Functions

Ascorbic acid, working in synchrony with vitamin E, is one of the most effective antioxidants in biological fluids and is capable of scavenging physiologically important reactive oxygen species and reactive nitrogen species. Vitamin C has been shown to also exhibit pro-oxidant properties, and although pro-oxidants tend to have deleterious effects, pharmacologic doses of ascorbate may have beneficial pro-oxidant activity in the treatment of certain diseases especially cancer. It is increasingly being recognized that the balance between antioxidant and pro-oxidant activity is important for health, and too much antioxidant might actually be harmful.

Ascorbic acid acts as a cofactor for a number of mixed function oxidases in hydroxylation processes in which it promotes enzyme activity by maintaining metal ions in their reduced form (particularly iron and copper). Vitamin C has a functional role in the formation of collagen by acting as a cofactor for protocollagen hydroxylase, the enzyme responsible for hydroxylation of prolyl and lysyl residues within nascent peptides in connective tissue proteins. Among these are collagen and related proteins, which make up the intercellular material of cartilage, dentin, and bone.

Ascorbate is also involved in (1) carnitine biosynthesis, serving as a cofactor to 6-*N*-trimethyl-L-lysine hydrolase; (2) γ-butyrobetaine hydrolase, which converts γ-butyrobetaine to carnitine; (3) degradation of tyrosine via 4-OH phenylpyruvate dioxygenase; (4) synthesis of adrenal hormones via dopamine β-hydroxylase; (5) biosynthesis of corticosteroids and aldosterone; (6) hydroxylation of cholesterol in the formation of bile acids; and (7) folate metabolism and leukocyte functions.

Vitamin C also has an important role in the intestinal absorption of nonheme iron by maintaining it in its reduced

ferrous (Fe^{2+}) form. More recently identified functions include nucleic acid and histone dealkylation and proteoglycan deglycanation.

Deficiency

Scurvy results due to deficiency of vitamin C and still occurs in developed countries, albeit rarely either as part of general malnutrition or in isolation. Those most at risk of the disease include (1) elderly men, particularly those who live alone; (2) those with alcohol dependence and smokers; (3) those taking unbalanced diets especially populations who have limited access to fresh fruits and vegetables either due to weather conditions or disasters both natural and manmade; (4) some mentally ill patients; (5) renal failure patients undergoing peritoneal dialysis or hemodialysis; (6) those with increased requirements, such as the sick and pregnant women; and (7) some patients with cancer.

Lack of vitamin C or scurvy causes an inability to form adequate intercellular cement substance in connective tissue, bones, and dentin resulting in swollen, tender, and often bleeding or bruised loci at joints and in other areas where structurally weakened tissue cannot withstand stress. Bone formation is usually impaired causing poor growth and lesions. Infantile scurvy, also known as *Barlow disease*, exhibits a bayonet-rib syndrome. The gums are livid and swollen, cutaneous bleeding often begins on the lower thighs as perifollicular hemorrhages, and large spontaneous bruises (ecchymoses) may arise almost anywhere on the body. This condition may be misconstrued as child abuse. Ocular hemorrhages, drying of salivary and lacrimal glands, parotid swelling, femoral neuropathy, edema of the lower extremities, and psychologic disturbances have also been described. Some scorbutic patients may develop anemia, display radiologic changes characteristic of osteoporosis, or die suddenly from heart failure.

Vitamin C deficiency has also been linked to anemia, diabetes mellitus, various types of cancer, cataract formation, and osteoporosis, but there is no clear evidence for a causal link. A meta-analysis found that regular vitamin C supplementation did not significantly reduce the incidence of common cold in the general population.

Toxicity

Rare adverse effects include increased oxalate excretion and kidney stone formation, increased uric acid excretion, excess iron absorption, lowered vitamin B_{12} concentrations, systemic conditioning, and "rebound" scurvy. The tolerable upper intake amount for vitamin C is 2 g/day for adults older than 19 years.

Laboratory Assessment of Status

At present, no useful functional tests of vitamin C adequacy are available; thus laboratory assessment of status is made by direct measurement of plasma, urine, or tissue concentrations of ascorbic acid, total vitamin C, or (rarely) metabolite. Plasma ascorbate concentration is considered a reliable indicator of vitamin C intake and has been measured photometrically by oxidation with 2,4-dinitrophenylhydrazine to form the red *bis*-hydrazone, or with 2,4-dichlorophenol-indophenol, which is reduced to a colorless form. A more specific approach is to use the enzyme ascorbate oxidase to convert ascorbate to dehydroascorbate, which then is coupled with *o*-phenylene diamine to form a product that is measured fluorometrically or at 340 nm on an automated analyzer. HPLC methods offer the potential advantage of specificity but generally are time-consuming. Detection may be done by precolumn derivatization to the fluorescent quinoxaline, or by electrochemical or coulometric means. Care must be taken during the analysis to prevent oxidation of the ascorbate before detection by using dithiothreitol or homocysteine added to the sample and mobile phase.

Vitamin C in plasma or serum is known to be readily degraded by oxidation caused by high temperature, light, neutral pH, pro-oxidant material such as certain enzymes, and interaction with iron or copper. Plasma samples should be treated with a metal-chelating and protein-precipitating acid, such as metaphosphoric or sulphosalicylic acids, soon after a phlebotomy followed by prompt freezing. Dithiothreitol may also be used as a preservative.

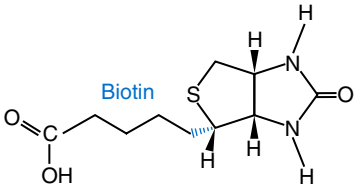


Fig. 27.14 Structure of biotin.

Biotin

Biotin is *cis*-tetrahydro-2-oxothieno[3,4-D]-imidazoline-4-valeric acid (Fig. 27.14). The vitamin in most organisms occurs mainly bound to protein.

Absorption, Transport, Metabolism, and Excretion

Biotin in the diet is largely linked to lysine residues on protein, and digestion of these proteins by gastrointestinal proteases and peptidases produce biocytin (ε-*N*-biotinyl lysine) and biotinyl peptides, and the latter may be further hydrolyzed by intestinal biotinidase to release biotin. Avidin, a protein found in raw egg whites, binds biotin tightly and prevents its absorption, and therefore raw egg consumption has been known to result in biotin deficiency.

A biotin carrier, the sodium-dependent multivitamin transporter (SMVT) for which pantothenic acid and lipoate compete, is located in the intestinal brush border membrane and transports biotin against a sodium ion concentration gradient. At higher biotin doses (>25 μmol/L), there is passive diffusion across cell membranes. By contrast, transport of biotin across the basolateral membrane is sodium-independent and electrogenic. The enzyme biocytinase (biotin amidohydrolase) in plasma and erythrocytes catalyzes the hydrolysis of biocytin to yield free biotin. Biotin is cleared from the circulating blood more rapidly in deficient than in normal mammals; it is taken up by such tissues as liver, muscle, and kidney and is localized in cytosolic and mitochondrial carboxylases.

Functions

The principal biochemical function of biotin in humans is as a cofactor for carboxylation reactions. Five carboxylases are currently found in human tissue; one of these, an acetyl-CoA carboxylase, is inactive and may act as a storage vehicle for biotin. The others are carboxylases for acetyl-CoA, propionyl-CoA, β-methylcrotonyl-CoA, and pyruvate. Recent advances have implicated biotin in the epigenetic regulation of gene expression.

Deficiency

Biotin deficiency is uncommon but may be seen (1) with prolonged consumption of raw egg white, which contains the protein avidin that binds biotin in the intestines preventing absorption, (2) in those on TPN without biotin supplementation, and (3) in those with inborn errors of biotin metabolism. The first two situations may be complicated by effects on gut flora that produce biotin.

Symptoms of biotin deficiency include anorexia, nausea, vomiting, glossitis, pallor, conjunctivitis, ataxia, hypotonia, depression, dry scaly dermatitis, and developmental delay in infants and children. Significantly lowered urinary excretion or circulating blood concentrations have been found in alcoholic individuals, in patients with achlorhydria, and among the elderly and some athletes. Biotin has also been linked to immune function and lipid metabolism.

Two inborn errors of biotin metabolism include (1) deficiency of biotinidase leading to the inability to utilize protein-bound biotin in the diet or to recycle endogenous biotin from holocarboxylases and also (2) an autosomal recessive condition known as biotin-responsive basal ganglia disease, which presents in childhood with subacute episodes of encephalopathy usually triggered by febrile illnesses and disappears without neurological manifestations, if biotin is given.

Toxicity

No adverse effects of biotin in doses up to 300 times normal dietary intake have been reported, which has been administered to patients with biotinidase deficiency.

Laboratory Assessment of Status

Traditionally biotin has been measured in biological samples by microbiological assays, but HPLC methods are mainly used at present. Lymphocyte propionyl-CoA carboxylase has been shown to be an early and sensitive indicator of biotin deficiency.

Niacin

The term *niacin* (also known as niacinamide or vitamin B_3) refers to nicotinic acid (pyridine-3-carboxylic acid), its amide nicotinamide, and derivatives that show the same biological activity as nicotinamide. Structures of both vitamers and the two coenzyme forms containing the nicotinamide moiety are shown in Fig. 27.15.

Absorption, Transport, Metabolism, and Excretion

Dietary NAD and NADP are hydrolyzed by enzymes, such as NAD glycohydrolase, in the intestinal mucosa to release

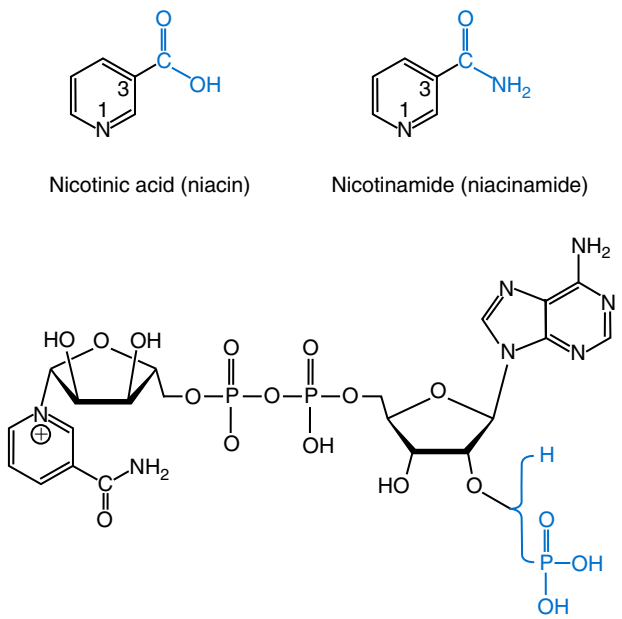


Fig. 27.15 Niacin, niacinamide, and coenzyme.

nicotinamide, which, together with any nicotinic acid, is rapidly absorbed in the stomach and the intestine by Na^+-dependent facilitated diffusion at low concentrations and passive diffusion at higher concentrations. Nicotinamide is the main circulating form in the plasma post absorption or following release from hydrolyzed liver NAD; it can be taken up by most tissues requiring NAD by simple diffusion. Once inside blood, kidney, brain, and liver cells, both nicotinic acid and nicotinamide are converted to coenzyme forms. Little storage of niacin as such occurs.

Functions

Niacin is essential as the coenzymes NAD and NADP in which nicotinamide acts as an electron acceptor or a hydrogen donor in a large number of redox reactions. Many enzymes function as dehydrogenases and catalyze such diverse reactions as the conversion of alcohols (often sugars and polyols) to aldehydes or ketones, hemiacetals to lactones, aldehydes to acids, and certain amino acids to keto acids.

Nicotinic acid, when used as a pharmaceutical agent, has important antiatherogenic properties. It effectively lowers triglycerides, raises HDL cholesterol, and shifts LDL particles to a less atherogenic phenotype (see Chapter 23).

Deficiency

Tryptophan is a precursor of niacin. As much as two-thirds of niacin required by adults can be derived from tryptophan metabolism via nicotinic acid ribonucleotide to NAD and NADP. It has been shown that 60 mg of tryptophan can provide the equivalent of 1 mg of niacin in the adult.

Pellagra is the classic deficiency disease presenting with dementia, diarrhea, and dermatitis. Niacin deficiency and therefore pellagra may occur in (1) people who consume corn-based staple diets in less developed countries, although this is now rare except during war or famine; (2) *carcinoid syndrome*, in which up to 60% of tryptophan is catabolized to

5-hydroxytryptophan and serotonin resulting in niacin deficiency; (3) **Hartnup disease**, an autosomal recessive disorder in which several amino acids, including tryptophan, are poorly absorbed in the intestines, resulting in inadequate niacin for body requirements; and (4) treatment with the antituberculous drug isoniazid, which competes with PL-5-phosphate (vitamin B_6), required as a cofactor in the conversion of tryptophan to niacin.

Toxicity

High doses are commonly associated with vascular dilation or "flushing," a burning, tingling sensation of the face (that may be reddened), arms, and chest that is thought to be mediated by prostaglandins. The symptoms of flushing have been taken as an end point sign in the formulation of a tolerable upper intake amount for niacin. This has been set at 35 mg/day for adults 19 years and older, with lower amounts for children and adolescents.

Laboratory Assessment of Status

At present, no blood markers are commonly used as indicators of niacin status. Most assessments of niacin nutriture have been based on measurement of the two urinary metabolites, *N′*-methylnicotinamide and *N′*-methyl-2-pyridone-5-carboxamide. HPLC-based methods are currently the methods of choice, although some capillary electrophoresis methods have been developed. Another approach that may prove valuable is the ratio of erythrocyte niacin coenzymes NAD to NADP. NAD concentrations respond to the niacin status, whereas NADP concentrations remain relatively constant under different conditions of niacin status. The niacin number is NAD/NADP × 100; a niacin number less than 130 is indicative of a risk of developing niacin deficiency.

Pantothenic Acid

Pantothenic acid, also known as vitamin B_5 (name derived from Greek *pantothen* meaning from everywhere), is of ubiquitous occurrence in nature, where it is synthesized by most microorganisms and plants. It is the amide formed by the linkage between pantoic acid (D-2,4-dihydroxy-3,3-dimethylbutyric acid) derived from L-valine, and β-alanine derived from L-aspartate.

Absorption, Transport, Metabolism, and Excretion

Pantothenic acid is taken in as dietary CoA compounds and 4′-phosphopantetheine, and is hydrolyzed by pyrophosphatase and phosphatase in the intestinal lumen to dephospho-CoA, phosphopantetheine, and pantetheine, which are further hydrolyzed to pantothenic acid. The vitamin is primarily absorbed as pantothenic acid through a saturable process at low concentrations and through simple diffusion at higher ones. After absorption, pantothenic acid enters the circulation and is taken up by cells in a manner similar to its intestinal adsorption. Pantothenic acid is excreted in the urine after hydrolysis of CoA compounds by enzymes that cleave phosphate and the cysteamine moieties.

Functions

Pantothenic acid has two major metabolic roles—the first as part of CoA, and the second as the prosthetic group of the acyl-carrier protein (ACP). In the former role, CoA is primarily involved in acetyl and acyl transfer reactions in catabolic processes of carbohydrate, lipid, and protein chemistry.

Deficiency

The widespread availability of pantothenic acid in food is commensurate with its many roles and makes dietary deficiency of pantothenate unlikely in humans. Symptoms have been produced in a few volunteers who received ω-methylpantothenic acid as an antagonist and in people fed semisynthetic diets virtually free of pantothenate. Subjects became irascible and developed postural hypotension and rapid heart rate on exertion, epigastric distress with anorexia and constipation, numbness and tingling of the hands and feet, hyperactive deep tendon reflexes, and weakness of finger extensor muscles.

Toxicity

No reports have described adverse effects, with the exception of occasional mild diarrhea, with oral pantothenic acid given in doses as high as 20 g/day.

Laboratory Assessment of Status

No convenient or reliable functional tests of pantothenic acid status are currently available; thus assessment is made by direct measurement of whole blood or urine pantothenic acid concentrations. Methods that have been used to measure pantothenic acid in human samples include radioimmunoassay, liquid chromatography coupled to mass spectrometry, gas chromatography–mass spectrometry, and a stable isotope dilution assay based on liquid chromatography–mass spectrometry. CoA and ACP have been measured by enzymatic methods.

TRACE ELEMENTS

The term **trace element** was originally used to describe the residual amount of inorganic analyte quantitatively determined in a sample. More sensitive methods allow accurate determination of most inorganic micronutrients present at very low concentrations in body fluids and tissue. Those present in body fluids (μg/L) and in tissue (mg/kg) are, however, still widely referred to as "trace elements." The corresponding dietary requirements are quoted in mg/day or μg/day, respectively. The biological effects of deficiency disease define the essential trace elements. For example, an element is considered essential when signs and symptoms induced by a deficient diet are uniquely reversed by an adequate supply of the particular trace element under investigation.

Dose-Effect Relationships

At low intakes of essential trace elements, deficiency disease may be seen. With increasing supply, a plateau region of optimal status is reached. Still higher intakes result in adverse effects. The concentration window separating beneficial from

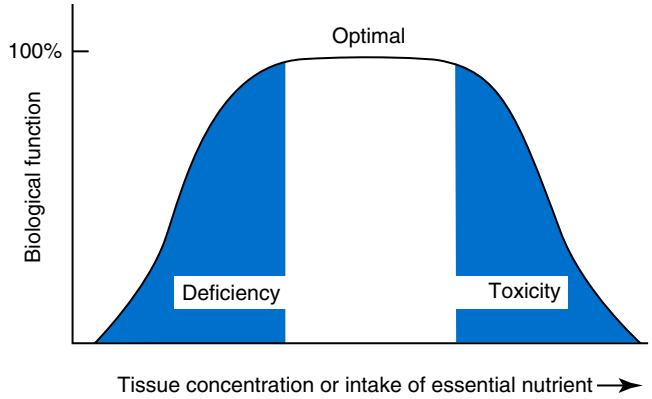


Fig. 27.16 Model of the relationship between tissue concentration and intake of an essential nutrient and dependent biological function.

toxic intakes varies depending on the element in question. This is similar to the dose-effect relationship described for organic micronutrients (Fig. 27.16). Therefore the RDA is set at amounts that are sufficient to prevent deficiency. A tolerable upper intake value that will prevent toxicity has been proposed for the inorganic micronutrients of known importance to human health.

Trace elements discussed in this chapter include copper, selenium, zinc, iodine, manganese, chromium, cobalt, and molybdenum. Iron is discussed in Chapter 28. Dietary sources, RDAs, and intravenous intakes are shown in Table 27.1. The reference intervals are given in Table 27.4.

TABLE 27.4 Reference Intervals for Trace Elements

Trace Element	Mass Unit	SI Units
Chromium (serum)	0.1–0.15 μg/L	<6 nmol/L
Chromium (urine)	<1–1.0 μg/24 h	2–19 nmol/24 h
Cobalt (serum)	<0.6 μg/L	<10 nmol/L
Cobalt (urine)	<1 μg/g creatinine	<17 nmol/L
Copper (serum)	70–140 μg/dL	10–22 μmol/L
Concentrations increase during pregnancy and oral contraception		
Copper (urine)	<50 μg/24 h	<0.7 μmol/24 h
Wilson disease	>200 μg/24 h	>3 μmol/24 h
Wilson disease postpenicillamine	>1700 μg/24 h	>25 μmol/24 h
Copper (liver)	8–40 μg/g dry weight	
Wilson disease	>250 μg Cu/g dry weight	
Iodine (urine)	76–546 μg/L	0.60–4.30 μmol/L
WHO-recommended urinary concentrations for various population groups are given in (WHO/NMH/NHD/EPG/13.1)		
Manganese (serum)	0.6–2.3 μg/L	11–42 nmol/L
Manganese (blood)	5–15 μg/L	90–270 nmol/L
Molybdenum (serum)	0.4–1.2 μg/L	4–12 nmol/L
Molybdenum (urine)	14.4–119 μg/24 h	0.15–1.24 μmol/24 h
Selenium (serum)	60–125 μg/L	0.75–1.57 μmol/L
Adult (UK)	26–77 μg/L	0.33–0.97 μmol/L
<1.5 years old	40–88 μg/L	0.51–1.12 μmol/L
1.5–3 years old	48–100 μg/L	0.6–1.29 μmol/L
4–18 years old		
Zinc (serum)	63–134 μg/dL	9.6–20.5 μmol/L
Zinc (urine)	0.2–1.3 mg/24 h	3–21 μmol/24 h

WHO, World Health Organization.

LABORATORY ASSESSMENT OF TRACE ELEMENT STATUS

Specimen Requirements

Direct determination of trace elements is done in many types of specimens—blood, plasma or serum, and urine. Less often, other sample types (e.g., leukocytes, saliva) may be analyzed. Tissue samples obtained by needle biopsy (e.g., liver, bone) or following an autopsy, and also hair and nail samples offer a noninvasive means of sampling tissue. Measurements of hair and nails for essential elements may be of value on a group basis during studies of severely depleted populations but are of limited value in the investigation of individual hospital patients. Problems of external contamination from environmental pollution, cosmetics, shampoos, and other sources are difficult to control.

Preanalytical Factors

Numerous variables can affect trace element determinations before analysis of the sample and should be considered. Age, sex, ethnic origin, time of sampling in relation to food intake, time of day, medication, and other factors should be noted. In patients with infection, after injury or postsurgery, the systemic inflammatory response affects the concentration of essential elements in circulating blood, independently of nutritional status.

Collection Equipment

Contamination from collection equipment can be a problem. Trace metal-free vacutainers are commercially available. To avoid contamination from the needle during venous sampling, at least the third specimen from the order of draw should be used. For some elements, collection is via plastic cannulae with the sample then placed into acid-washed containers. Urine collection bottles may be washed thoroughly with ultrapure water. For random urine samples and tissue biopsy samples, a plain plastic container with no added preservative is preferable.

ANALYTICAL METHODS FOR MEASURING TRACE ELEMENTS

Analytical methods should provide sufficient sensitivity to reliably determine the concentrations in the limited specimen

usually available. Techniques that are appropriate include atomic absorption spectrometry (AAS), inductively coupled plasma-optical emission spectrometry (ICP-OES), and inductively coupled plasma-mass spectrometry (ICP-MS). UV-vis spectrophotometry is possible for elements where concentrations are higher. Other techniques have been applied to elemental analysis. Using laser ablation ICP-MS or X-ray fluorescence spectrometry, 2D or even 3D images of trace element distributions within biological materials can be produced. Speciation methods involve techniques to separate compounds of individual elements in a sample. A typical procedure involves HPLC coupled to ICP-MS, to separate methyl mercury from inorganic mercury. Details of these procedures are given in Chapters 9 and 13.

Copper

Absorption, Transport, Metabolism, and Excretion

Copper absorption occurs mainly in the small intestine, although some gastric uptake has been shown. The extent of intestinal copper absorption varies with dietary copper content and is around 50% at low intakes (<1 mg/day) but only 20% at higher intakes (>5 mg/day). Absorption is reduced by other dietary components such as zinc (which induces metallothionein that binds Cu^{2+}), vitamin C (which converts Cu^{2+} to insoluble nonabsorbable Cu^{+}), molybdate, and iron, and is increased by amino acids and by dietary sodium.

Absorbed copper is transported to the liver in portal blood, bound to albumin, where it is incorporated by hepatocytes into cuproenzymes and other proteins, and is then exported as ceruloplasmin via peripheral blood to tissues. Although two-thirds of the 80 to 100 mg total body content is located in the skeleton and muscle, the liver is the key organ in copper homeostasis. The ceruloplasmin molecule contains 6 to 8 atoms of copper per molecule, with 6 atoms of copper involved in the protein's ferroxidase and free radical scavenging activities.

Ceruloplasmin is a positive acute-phase reactant, and concentrations increase during infection, after tissue injury, and also in pregnancy and during the use of oral contraceptives. The increase is accompanied by a rise in serum copper concentration. A smaller amount of copper in plasma (<10%) is bound to albumin by specific peptide sequences, and this copper is in equilibrium with plasma amino acids. This fraction may be important for cellular uptake.

Copper is predominantly excreted via bile into feces. Patients with cholestatic jaundice or other forms of liver dysfunction are at risk for copper accumulation caused by failure of excretion.

Functions

Copper is a catalytic component of numerous enzymes and a structural component of other important proteins.

Energy Production. Cytochrome *c* oxidase contains copper and iron. Located on the external face of mitochondrial membranes, the enzyme catalyzes four-electron reduction of molecular oxygen, establishing a high-energy proton gradient across the inner mitochondrial membrane that is necessary for ATP production.

Connective Tissue Formation. Protein-lysine 6-oxidase (lysyl oxidase) is a cuproenzyme essential for stabilizing extracellular matrixes by cross-linking of collagen and elastin.

Iron Metabolism. Copper-containing enzymes, ferroxidase I (ceruloplasmin) and ferroxidase II, and hephaestin in the enterocyte oxidize ferrous, to ferric, iron. This allows incorporation of Fe^{3+} into transferrin and eventually into Hb.

Central Nervous System. Dopamine mono-oxygenase is an enzyme that requires copper as a cofactor and uses ascorbate as an electron donor. This enzyme catalyzes the conversion of dopamine to norepinephrine, an important neurotransmitter. Monoamine oxidase, one of the numerous amine oxidases, is a copper-containing enzyme that catalyzes the degradation of serotonin in the brain and is involved in the metabolism of catecholamines.

Melanin Synthesis. Tyrosinase is a copper-containing enzyme that is present in melanocytes and catalyzes the synthesis of the melanin biopigments pheomelanin and eumelanin.

Antioxidant Functions. Both intracellular and extracellular superoxide dismutases (SODs) are Cu- and Zn-containing enzymes, able to convert superoxide radicals to hydrogen peroxide, which can be removed subsequently by catalase and other antioxidant defenses.

Regulation of Gene Expression and Intracellular Copper Handling. Copper-dependent proteins act as transcription factors for specific genes, such as those regulating SOD, and catalase. Synthesis of metallothionein, important in regulating the intracellular distribution of copper, is controlled by copper-responsive transcription factors. Additional specialized proteins act as "copper chaperones" to deliver copper to intracellular sites and to prevent oxidative damage by free copper ions.

Aceruloplasminemia is another genetic defect resulting in failure of hepatic synthesis of ceruloplasmin leading to secondary iron overload, and insulin-dependent diabetes developing in the 4th to 5th decade of life.

Deficiency

Anemia. Copper deficiency is a rare but treatable cause of hematologic abnormalities. Usual hematologic features include anemia, leucopenia, and rarely thrombocytopenia. The mechanism is not clear.

Zinc-Induced Hypocupremia. Zinc intake (e.g., via dental fixatives) induces metallothionein expression in the intestinal mucosa, which sequesters copper blocking its absorption leading to copper deficiency.

Menkes Syndrome. This is a rare (1/100,000 live births), X-linked mutation disorder due to impaired intestinal transport of copper caused by a mutation in the ATP7A gene that leads to severe copper deficiency. It occurs in male infants at 2 to 3 months who present with hypotonia, seizures, failure to thrive, kinky hair (pili torti), facial changes, and neurologic abnormalities. Low concentrations of copper in plasma, liver, and brain occur because of the impaired intestinal absorption. Findings include very low plasma copper and ceruloplasmin concentrations.

Deficiency of the copper enzyme dopamine mono-oxygenase in CSF may allow early diagnosis. Therapy with parenteral copper histidine is rarely successful, and patients usually die following aortic rupture.

Cardiovascular Disease. The role of copper in human cardiovascular disease is controversial. The possible importance illustrates the Janus phenomenon of duality, with evidence suggesting both protective and harmful scenarios. Evidence for a positive link with low dietary copper intake comes from epidemiologic surveys showing that increased plasma copper values are a positive cardiovascular risk factor.

Neuropathy. Copper deficiency–associated myeloneuropathy is an increasingly recognized disorder. Deficiency secondary to celiac disease and excessive zinc intake are among the causes suggested.

Toxicity

Toxicity can arise from ingestion of contaminated diet or drinking water. Symptoms of acute poisoning include bloody vomiting, melena, and hypotension. Fatal renal failure has been recorded following accidental or intentional ingestion of copper sulfate. Indian childhood cirrhosis is recognized in children who may be genetically sensitive to copper in drinking water.

Wilson disease, the genetic disorder due to a mutation in the ATP7B gene, affects copper incorporation into ceruloplasmin and excretion into bile causing increased copper accumulation in tissues. With an incidence of 1/30,000 live births, Wilson disease is one of the more common genetic disorders. The presentation is highly variable. Primarily, it affects either the liver or the brain, but rare cases are diagnosed with renal tubular disease or hemolytic anemia. Severe liver disease typically develops in young children while neurological signs tend to appear in teenagers and early adults, but onset and diagnosis may be made at any age. Initial investigations include plasma copper and ceruloplasmin, which will usually be low. Although the total plasma copper is decreased, the non-ceruloplasmin-bound free fraction is increased. However, methods for analysis of this fraction are not particularly reliable. Other signs and symptoms include copper deposits in the eye (Kayser-Fleischer rings), abnormal liver function tests, and increased urine copper excretion. In cases that are difficult to diagnose, analysis of liver biopsies for copper may help and a test involving an oral dose of stable ^{65}Cu isotope has been developed. Genetic testing is possible, but because several hundred mutations exist, this may not always be informative. Transplantation is required for cases of acute liver disease, while the chronic form of Wilson disease is treated by oral chelating agents (penicillamine) that remove excess copper from tissue and increase urinary excretion. Maintenance therapy may involve oral administration of zinc salts or ammonium molybdate that block intestinal absorption of copper.

Laboratory Assessment of Status

Serum copper and ceruloplasmin concentrations are measured, but factors that increase the hepatic synthesis of ceruloplasmin, such as the APR, pregnancy or estrogen-containing contraceptive pill, will increase plasma copper independently of dietary copper intake. Urinary excretion of copper is decreased during dietary deprivation, but the change from an already low basal value is small, which makes this of limited use.

Selenium

Absorption, Transport, Metabolism, and Excretion

Intestinal absorption of various dietary forms of selenium is efficient and is not regulated. The inorganic salts, selenite (Se^{IV}) and selenate (Se^{VI}), used in some dietary supplements and in food fortification are almost completely absorbed, but much of the selenate ion is rapidly excreted in urine. Selenium from inorganic salts is more rapidly incorporated into glutathione peroxidase and other selenoproteins than selenium from organic sources containing selenomethionine. However, selenium-enriched yeast containing the organic forms is considered less toxic and is widely used as a dietary supplement.

Whole body selenium is about 15 mg, with higher concentrations in the kidney and liver, followed by the other organs. The concentrations of selenium in whole blood and in plasma are related to dietary intake. About 50% to 60% of total plasma selenium is present as the protein selenoprotein P, a highly basic protein having multiple histidine residues and about 10 atoms of selenium per molecule. Around 30% of plasma selenium is present as glutathione peroxidase (GSHPx-3), the remainder being incorporated into albumin as selenomethionine.

The major route of excretion is via the urine and reflects recent dietary selenium intake. The amounts excreted vary widely, ranging from less than 20 μg/L to more than 1000 μg/L.

Functions

More than 30 biologically active selenocysteine-containing proteins have been identified and their biological function investigated. Some of these are discussed below.

Glutathione Peroxidase (GSHPx). This enzyme has at least eight isoforms with GSHPx-3 present in plasma. These enzymes use the reducing power of glutathione to remove an oxygen atom from hydrogen peroxide and lipid hydroperoxides.

Iodothyronine Deiodinase. Type I (Dio1), II (Dio2), and III (Dio3) isoforms of this enzyme are responsible for conversion of the precursor hormone T_4 to the active hormone T_3. Type 1 is located in the thyroid, liver, kidney, and muscle and is responsible for more than 90% of plasma T_3 production. Pituitary, brain, and brown adipose tissue contain the type II and III deiodinases.

Thioredoxin Reductases. Three isoforms catalyze the NADPH-dependent reduction of thioredoxin and are important in maintaining the intracellular redox state.

Selenophosphate Synthetase. This enzyme is required for the intracellular synthesis of selenoproteins via a monoselenium phosphate intermediate.

Selenoprotein P. This protein is the major selenium-containing protein in blood plasma; it may be a transport protein for the element, and it has an antioxidant function.

Deficiency

Selenium enters the food chain mainly as selenomethionine from plants. The soil content of selenium is highly variable, and the geographic source of plant and animal foodstuffs determines the dietary intake. Soils in parts of China and New Zealand are particularly low in selenium, while wheat and other cereal products from North America are good sources.

Keshan Disease. Conclusive evidence of a role for selenium in human nutrition came with the results from large-scale trials in China that showed the protective effects of selenium supplementation in children and young adults suffering from an endemic cardiomyopathy. This was observed in areas of the country (Keshan region) with low soil selenium concentrations.

Kashin–Beck Disease. A type of severe arthritis described in parts of China and neighboring areas of Russia.

Nutritional Depletion in Hospital Patients. The need for selenium supplementation, with monitoring, during nutritional support is well established, and few cases of symptomatic deficiency are now reported.

Thyroid Function. As noted above, deiodinase enzymes are selenoproteins and are necessary for normal thyroid function. Children with biochemical and clinical signs of hypothyroidism were successfully treated with oral selenium. A report of a study in the United Kingdom concluded that supplementation had no effect on thyroid status.

Immune Function. Deficiency of selenium is accompanied by loss of immunocompetence, and this is related to the reduction of selenoproteins in the liver, spleen, and lymph nodes.

Inflammatory Conditions. Many conditions associated with inflammation and increased oxidative stress could be influenced by selenium status.

Cardiovascular Disease. A protective role for selenium against cardiovascular disease has been extensively investigated, but with no conclusive findings.

Viral Virulence. It is possible that an unusually virulent strain of the Coxsackie virus is part of the cause of cardiomyopathy in selenium-depleted regions of China.

Cancer Chemoprevention. Epidemiologic surveys suggest an association between cancer incidence and soil selenium. However, a large selenium and vitamin E study known as the Selenium and Vitamin E Cancer Prevention Trial (SELECT) concluded that selenium or vitamin E, alone or in combination at the doses and formulations used, did not prevent prostate cancer in healthy men.

Toxicity

Clinical signs of selenosis include garlic odor on the breath, hair loss, nail damage, nervous system and skin disorders, poor dental health, and paralysis in extreme cases. The tolerable upper limit has been set at 400 μg/day for adults and less than 280 μg/day for children. Cases of toxicity and even death from self-administration or homicidal dosages have been reported. Some parts of the world, generally where there are volcanic soils, are naturally selenium-rich, and locally produced food contains high concentrations of selenium. Consumption of these foods leads to chronic selenosis typified by cases reported from Enshi City in the Hubei region of China.

Laboratory Assessment of Status

Selenium concentrations in whole blood or plasma/serum are the usual markers of selenium status. Samples may be analyzed by electrothermal-AAS or by ICP-MS (note the interference due to the Gd^{2+} ion). Urinary selenium is measured where there is occupational exposure. Measurement of glutathione peroxidase (GSHPx-3) activity in whole blood or serum is an additional biomarker of selenium status. Results are dependent on the substrate and other assay conditions so that comparisons among different studies require careful interpretation. Plasma selenoprotein P may be determined, but methods are not readily available. Plasma selenoprotein P, plasma GSPHx-3, and total plasma selenium concentration are all lowered by the APR to injury or infection. This effect should be considered when results are being interpreted. Hair and nail selenium concentrations can be useful in population studies.

Zinc

Zinc is second to iron as the most abundant trace element in the body.

Absorption, Transport, Metabolism, and Excretion

Intestinal absorption of zinc is controlled in the face of variable dietary zinc input and is usually from 20% to 50% of the dietary content. Uptake from the gut lumen into intestinal cells is achieved and regulated by zinc transporter proteins. Several factors such as dietary phytate, fiber calcium, and iron reduce intestinal absorption by forming insoluble complexes. Profound zinc deficiency in humans was first recognized in a population where geophagia (eating soil) was practiced; because zinc is bound to the soil components, deficiency resulted.

Absorbed zinc is transported to the liver by the portal circulation, where incorporation into metalloenzymes and plasma proteins occurs. Plasma contains less than 1% of the total body content of zinc and lies within a narrow concentration interval. About 80% of plasma zinc is associated with albumin, and most of the rest is tightly bound to α_2-macroglobulin. The zinc on albumin is in equilibrium with plasma amino acids (mostly histidine and cysteine), and this small (<1%) ultrafiltrable fraction may be important in cellular uptake mechanisms (Fig. 27.17). Red cell zinc concentrations are about 10 times higher than in plasma. Fecal excretion includes both unabsorbed dietary zinc and zinc resecreted into the gut. The amount normally equals the total dietary intake and is about 10 to 15 mg/day. Urine excretion of zinc is normally about 0.5 mg/day, but this can increase markedly during catabolic illness.

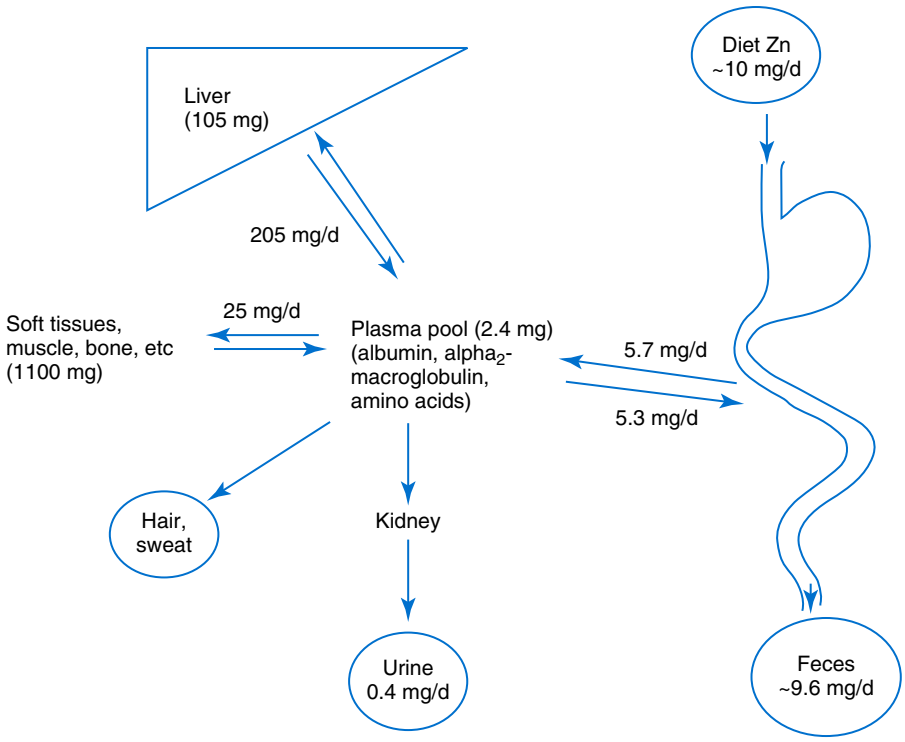


Fig. 27.17 Zinc metabolism.

Functions

More than 300 zinc metalloenzymes have been described. Some important examples include carbonic anhydrase, alkaline phosphatase, RNA and DNA polymerases, and alcohol dehydrogenase. The key roles of zinc in protein and nucleic acid synthesis explain the failure of growth and impaired wound healing observed in individuals with zinc deficiency. In some enzymes, such as Cu and Zn SOD, structural stability is ensured by zinc protein binding and the catalytic activity of the enzyme by the active copper site. Proteins can form domains that bind tetrahedral zinc atoms by coordination with histidine and cysteine to form folded structures known as "zinc fingers." These molecules have important roles in gene expression and hormone function.

Deficiency

Increased amounts are required during pregnancy and lactation. Strict vegetarians may need as much as 50% more zinc per day because of increased phytic acid and fiber in their diet. As might be expected from the multiple biochemical functions of zinc, the clinical presentation of deficiency disease is varied, nonspecific, and related to the degree and duration of depletion. Signs and symptoms include depressed growth with stunting; increased incidence of infection, possibly related to alterations in immune function; diarrhea; altered cognition; defects in carbohydrate use; reproductive teratogenesis; skin lesions; alopecia; eyesight defects; and other adverse clinical outcomes.

Acrodermatitis enteropathica is an autosomal recessive disorder resulting in failure to absorb zinc characterized by severe periorificial and acral dermatitis, alopecia, and diarrhea. Patients with this disorder have abnormally low blood zinc concentrations (<30 μg/dL). Symptoms are reversed by oral zinc supplementation. Regular monitoring of serum concentrations is recommended.

Immune Function. Zinc depletion impairs immunity and has a direct effect on the gastrointestinal tract that increases the severity of enteric infections. A review of controlled trials of zinc supplementation of children in low-income countries found significant clinical benefits with substantial reduction in the incidence of severe diarrhea and respiratory infection.

Subclinical Effects of Deficiency. Marginal zinc deficiency, not severe enough to cause clinical signs and symptoms, may be associated with "subclinical" effects. These include susceptibility to sickle cell disease, impaired immune response, and mental well-being.

Toxicity

Zinc toxicity is unusual and is generally associated with ingestion of zinc-contaminated food or water. Symptoms include abdominal pain, diarrhea, nausea, and vomiting. Chronic ingestion can lead to a zinc-induced deficiency of copper.

Laboratory Assessment of Status

Although plasma zinc determination is insensitive to dietary zinc intake and is subject to a variety of influences, it remains the most widely used laboratory test to confirm severe deficiency and to monitor adequacy of zinc provision, especially when interpreted together with changes in serum albumin and the APR. A clinical and biochemical response to zinc supplementation supports a diagnosis of deficiency. It is

important to recognize that a low plasma zinc concentration may not mean deficiency. Plasma albumin is a negative acute-phase reactant that redistributes, together with zinc, into interstitial space from the plasma pool during infection, after trauma, and in chronic disease. It is essential to consider plasma zinc results along with plasma albumin and a marker of the APR. Increased urine zinc is an important source of loss in the severely injured catabolic patient. Urine output increases with amino acid infusion given during TPN.

Iodine

Absorption, Transport, Metabolism, and Excretion

Absorption of dietary iodide is not entirely understood. The Na^+/I^- symporter (NIS) on the apical surface of enterocytes of the small intestine is responsible for uptake into vesicles within the cell, but movement into the circulation is not clear. Under normal circumstances, 90% to 95% of the intake is absorbed. Subsequent transport, movement into thyroid cells (which also involves an NIS), synthesis of thyroid hormones, and release is discussed in Chapter 42. Iodine is primarily excreted in the urine.

Functions

The essential function for iodine relates to thyroid physiology (see Chapter 42). While iodine is concentrated in thyroid tissue and hormones, around 70% of the body's iodine is distributed in many other tissues. Its role in mammary tissue relates to fetal and neonatal development, but its role in other tissues is partially unknown.

Deficiency

Deficiency of iodine is a major health problem throughout the world, occurring in regions where the soil and groundwater are deficient in the element. Deficiency may also develop following prolonged use of special diets that may have low iodine content. The most serious adverse effect of iodine deficiency is impaired development of nervous tissue in the fetus. Thyroid hormones are required for neuronal migration and myelination of the fetal brain. Severe maternal iodine deficiency during pregnancy increases the risk of stillbirths and congenital abnormalities. A decrease in intelligence quotient (IQ) points has been shown in children who are iodine deficient.

Toxicity

Iodine is a strong oxidizing agent and acts as an acid corrosive. Iodine vapor irritates the eyes, skin, and mucous membranes. Common causes of acute toxicity include accidental or deliberate ingestion, resulting in corrosive damage to the gastrointestinal tract, cardiovascular collapse, and renal failure; liberal use of iodine to open wounds; large injections of iodine-containing radio-contrast media; and amiodarone treatment for cardiac arrhythmias. Iodine toxicity causes a rapid decrease in release of thyroxine and 3,5,3′-triiodothyronine from thyroglobulin, decreased uptake of iodide by the thyroid gland, and hence decreased synthesis of thyroid hormones. This culminates in acute hypothyroidism.

Laboratory Assessment of Status

Urine is the matrix of choice for suspected iodine toxicity and for the assessment of recent iodine *intake*. For epidemiological studies, such as population-based surveys, the *median* urinary iodine concentration can be used to classify that *population's* iodine status. Adequate iodine status is indicated by a population median urinary iodine concentration greater than 0.79 μmol/L (100 μg/L). However, the urine iodine test is not an appropriate test to diagnose iodine deficiency in individuals. Urine iodine levels have a low predictive value in an individual because urinary iodine concentration varies substantially between days and seasons, as well as within a day (up to threefold) as a consequence of circadian rhythm and due to differences in fluid intake. Because the thyroid gland can store large amounts of iodine (12 to 16 mg in an iodine sufficient individual), a "low" urine iodine concentration no more indicates iodine deficiency than a low urinary sodium indicates sodium deficiency. It is recommended, however, that iodine status can be reliably estimated if 10 repeat random urine samples are collected. Iodine concentration in serum or whole blood provides no useful information relevant to iodine nutritional status.

Manganese

Absorption, Transport, Metabolism, and Excretion

Dietary manganese is absorbed from the small intestine. Diets high in iron, calcium, magnesium, phosphates, fiber, phytic acid, oxalate, and tannins from tea can reduce the absorption. Once absorbed, manganese is transported to the liver bound to albumin; it is then exported to other tissues bound to transferrin and possibly to α_2-macroglobulin. Excretion of manganese occurs primarily via bile into feces, with urine output being very low and not sensitive to dietary intake.

Functions

Manganese is a constituent of many important metalloenzymes. Mn^{2+} ions can be replaced by Mg^{2+}, Co^{2+}, and other cations during the activation of some enzymes. Some important manganese-dependent enzymes are discussed.

Superoxide Dismutase (SOD). Manganese-dependent SOD is a mitochondrial enzyme that is an important factor in limiting oxygen toxicity. The enzyme catalyzes the breakdown of the superoxide radical O_2^- to H_2O_2, which is then removed by catalase and glutathione peroxidase.

Pyruvate Carboxylase. This enzyme is required to catalyze the formation of phosphoenolpyruvate (PEP), from pyruvate, a key reaction in the hepatic synthesis of glucose.

Arginase. Arginase is the terminal enzyme in the urea cycle, hydrolyzing L-arginine to urea and ornithine and completing the deamination of amino acids. Arginase is most concentrated in the liver, but is also found in other tissues. The activity of arginase affects the production of nitric oxide by limiting the availability of L-arginine required for synthesis of nitric oxide synthetase.

Glycosyl Transferases. These enzymes are responsible for the sequential addition of carbohydrate molecules to proteins to form proteoglycans, and ultimately connective tissue and cartilage.

Deficiency

Overt manganese deficiency has not been documented in humans eating natural diets. Symptoms in experimentally induced manganese deficiency in animal studies include impaired growth and reproductive function, skeletal abnormalities, impaired glucose tolerance, and cholesterol synthesis.

Prolidase deficiency in infants is a rare genetic disorder that is known to be associated with abnormalities of manganese biochemistry. One study suggests there is low arginase activity due to a defect in the supply of manganese for enzyme activation.

Toxicity

Prolonged exposure to manganese-containing dust or fumes at work is well recognized, with neurologic symptoms resembling Parkinson disease developing slowly over a period of months or years. Patients with severe liver disease may have neurologic and behavioral signs of manganese neurotoxicity resulting from failure to excrete manganese in bile. Manganese deposition in the globus pallidus results in T_1-weighted magnetic resonance imaging (MRI) signal hypersensitivity. By causing deficits in neurotransmitter production, manganese ions may be partially responsible for symptoms of postsystemic hepatic encephalopathy. Deposition of manganese in the brain has been demonstrated in children with biliary atresia, in adult cirrhotic patients and patients receiving manganese intravenously during TPN, especially those with cholestasis. A study in adults that compared the effects of increasing doses of manganese in patients receiving home parenteral nutrition showed a good correlation between blood manganese, MRI intensity, and T_1 values in the globus pallidus.

Increases in serum manganese to greater than 1.6 μg/L (>30 nmol/L) or in blood manganese to greater than 20 μg/L (>360 nmol/L) are indices of manganese retention.

Laboratory Assessment of Status

Manganese status should be monitored in at-risk patients. Serum manganese, lymphocyte Mn SOD activity, and blood arginase are potentially useful when possible nutritional depletion is assessed. Blood or serum manganese in combination with brain MRI scans and neurologic assessment detect excessive exposure.

Chromium

Chromium occurs in biology as Cr^{3+} or Cr^{6+}, each having markedly different properties. The element has no redox or acid-base properties, but in the Cr^{6+} form it is a strong oxidant that can cause tissue damage (e.g., chromic acid), although it is normally rapidly reduced to Cr^{3+}.

Absorption, Transport, Metabolism, and Excretion

Absorption is increased marginally by ascorbic acid, amino acids, oxalate, and other dietary factors. After absorption, chromium binds to plasma transferrin with an affinity similar to that of iron. It then concentrates in human liver, spleen, other soft tissue, and bone. Urine chromium is the main route for excretion.

Functions

While symptoms have been reported in severely chromium-deficient rats (impaired growth, reduced life span, corneal lesions, alterations in carbohydrate, lipid, and protein metabolism), studies by other groups yielded inconsistent results.

Deficiency

Chromium is proposed to play a role in impaired glucose tolerance, diabetes, and cardiovascular disease. Clinical signs attributed to chromium deficiency were described in a few patients receiving TPN for a prolonged period using fluids that did not supply sufficient chromium. All had similar presentations, with previously stable patients developing insulin-resistant glucose intolerance, weight loss, and in some cases neurologic deficits. Addition of substantial amounts of Cr^{3+} to the intravenous fluids (150 to 200 μg/day) reversed glucose intolerance and reduced insulin requirements with eventual improvement in neurologic symptoms. These cases influenced the US Food and Nutrition Board to designate chromium as an essential trace element, although this is not universally accepted.

Toxicity

Hexavalent chromium is a recognized carcinogen, and industrial exposure to fumes and dusts containing this metal is associated with increased incidence of lung cancer, dermatitis, skin ulcers, and hypersensitivity reactions. Chromium picolinate is a widely used dietary supplement, and this compound may cause renal and hepatic damage when used at high doses.

In addition, markedly increased concentrations (up to 1000-fold) have been observed in both plasma and urine in patients with problem hip prostheses.

Laboratory Assessment of Status

Measurement of chromium in blood plasma or serum, the diet, in oral and intravenous nutritional support regimens requires great care to prevent contamination before and during analysis. Increased amounts of chromium in urine confirm recent occupational or environmental exposure. Monitoring urine chromium during trials that use pharmacologic dosages of chromium confirms compliance and detects potential toxicity.

Cobalt

Cobalt (Co) is essential for humans as an integral part of vitamin B_{12} (cobalamin). No other function in the human body is known. Non-vitamin B_{12} cobalt does not interact with the body vitamin B_{12} pool.

Increased exposure from industrial uses, particularly with hard metal blades, leads to high mean urinary Co concentrations. Increases are also associated with hip prostheses for which concentrations can be markedly raised with no apparent clinical evidence of overt toxicity; however, there is some emerging evidence of cobalt toxicity with wear from metal-on-metal hip prostheses.

Erythropoiesis can be stimulated by inorganic cobalt ions (Co^{2+}), even in nonanemic subjects, and cobalt chloride was

once administered at daily doses of 25 to 50 mg for use as an antianemic agent and Co^{2+} therapy proved effective in stimulating erythropoiesis in both nonrenal and renal anemia. The mode of action is to stabilize hypoxia-inducible transcription factors that increase the expression of the erythropoietin gene. Because the mass of Hb is a critical factor in aerobic sports, use of cobaltous salts is tempting for athletes and horse racing trainers to stimulate erythropoiesis. But there are risks to health with severe organ damage, primarily in the heart and the sensory systems, the gastrointestinal tract, and the thyroid.

Molybdenum

Absorption, Transport, Metabolism, and Excretion

Molybdenum is efficiently absorbed over a wide range of dietary intakes, mainly as molybdate, although competitive inhibition of absorption by sulfate reduces intestinal uptake. Concentrations in whole blood are about 1.0 μg/L (10 nmol/L), and some 80% to 90% or more of molybdenum is bound to red cell proteins. Transport of small amounts in plasma may involve α_2-macroglobulin. Urine output directly reflects the dietary intake of molybdenum.

Functions

Several important mammalian enzymes, sulfite oxidase, xanthine dehydrogenase, and aldehyde oxidase require molybdenum as a cofactor. The organic component is a molybdopterin complex. Sulfite oxidase is probably the most important enzyme in relation to human health, catalyzing the last step in the degradation of sulfur amino acids by oxidizing sulfite to sulfate and transferring electrons to cytochrome *c*. Xanthine dehydrogenase and aldehyde oxidase hydroxylate heterocyclic substances, such as purines and pteridines.

Deficiency

Molybdenum deficiency has not been observed in healthy people consuming a normal diet. A single case report described a patient receiving prolonged parenteral nutrition during treatment for severe Crohn disease who developed an intolerance to intravenous amino acids, especially L-methionine. Clinical signs included tachycardia, visual defects, neurologic irritability, and eventually coma. Symptoms improved on discontinuation of amino acid infusion. Biochemical abnormalities included high methionine and low uric acid concentrations in plasma. Treatment with ammonium molybdate (300 μg/day, IV) improved the clinical and biochemical abnormalities.

Very rare, usually fatal, recessive inherited diseases result from defects in the biosynthesis of molybdenum cofactor. Symptoms include failure to thrive and seizures with lens dislocations and cerebral atrophy developing later. Diagnosis is made by detection of excess sulfite in urine.

Toxicity

Molybdenum compounds have low toxicity in humans. Reports have described increased blood uric acid in those with occupational exposure and in Armenian populations that have an abnormally high dietary intake (10 to 15 mg/day). A single report of acute toxicity from self-administration of 300 to 800 μg/day with a cumulative total of 13.5 mg over 18 days led to acute psychosis and seizures but is inconsistent with another study in healthy men given as much as 1500 μg/day for 24 days with no adverse effects reported. Excess molybdenum intake prevents copper absorption through formation of an insoluble thiomolybdate-copper complex. This is the basis for using ammonium molybdate in the management of Wilson disease.

Laboratory Assessment of Status

Whole blood and serum or plasma molybdenum concentrations are too low to be reliably used to detect deficiency. However, urinary output is responsive to increases or decreases in input. Measuring urate or sulfite in the urine is the most valuable means of confirming molybdenum cofactor disorder or possible molybdenum deficiency.

POINTS TO REMEMBER

Vitamins		
Vitamin A	Retinol	Antioxidant; important role in vision
Vitamin E	Tocopherols	Antioxidant; multiple functions including bone metabolism
Vitamin K	Phylloquinones	Coagulation; bone metabolism
Vitamin B_1	Thiamine, aneurin	Coenzyme for decarboxylation and transketolation reactions; involved in carbohydrate metabolism
Vitamin B_2	Riboflavin	Prosthetic group for oxidation-reduction reactions
Vitamin B_3	Niacin	Hydrogen acceptors (as NAD and NADP) in many oxidation-reduction reactions
Vitamin B_5	Pantothenic acid	Coenzyme A required for metabolism of carbohydrates, fats, and proteins
Vitamin B_6	Pyridoxine	Coenzyme for enzymes involved in metabolism of various amino acids
Vitamin B_7	Biotin, vitamin H	Coenzyme for carboxylation reactions involved in gluconeogenesis and lipogenesis
Vitamin B_9	Folate	Required for the interconversions of amino acids and the biosynthesis of purines and pyrimidines
Vitamin B_{12}	Cobalamin	Required for erythropoiesis, methylation processes necessary for DNA metabolism, and is a cofactor for various enzymes
Vitamin C	Ascorbic acid	Connective tissue formation; antioxidant

POINTS TO REMEMBER—cont'd

Trace Elements	
Copper	Catalytic component of numerous enzymes and has structural roles
Selenium	Component of numerous selenocysteine-containing proteins such as glutathione peroxidase and iodothyronine deiodinase
Zinc	Component of numerous metalloenzymes and proteins including transcription factors
Iodine	Important role in thyroid hormone synthesis
Manganese	Constituent of many important metalloenzymes such as superoxide dismutase
Chromium	Implicated roles in impaired glucose tolerance, diabetes, and cardiovascular disease
Cobalt	Integral component of vitamin B_{12} (cobalamin)
Molybdenum	Cofactor for several enzymes such as sulfite oxidase and xanthine dehydrogenase

REVIEW QUESTIONS

1. Which enzyme activity is decreased as a result of vitamin B_1 deficiency?
   a. Glutathione reductase
   b. Glutathione peroxidase
   c. Transketolase
   d. Methyltetrahydrofolate reductase
2. Which of the following statements best describes the effects of systemic inflammatory response (SIR) on vitamin status?
   a. SIR causes a decrease in the vitamins A, E, B_2, B_6, C, D, and carotenoids.
   b. SIR causes an increase in the vitamins A, E, B_2, B_6, C, D, and carotenoids.
   c. SIR does not have an appreciable effect on any vitamins in blood circulation.
   d. SIR causes a decrease in albumin and CRP that affect vitamin status.
3. Which of the following statements best accounts for vitamin A deficiency?
   a. It is dependent on renal functioning.
   b. It is not caused by liver disease.
   c. It may result due to decreased retinol binding protein (RBP) concentration in circulating blood.
   d. Circulating retinol always correlates with total body vitamin A stores.
4. Which of the following is a major cause of fat-soluble vitamin deficiency?
   a. Decreased fat intake
   b. Hemolytic disease
   c. Renal insufficiency
   d. Malabsorption syndromes
5. Which of the listed vitamin deficiencies is not associated with the disease state?
   a. Xerophthalmia—vitamin A
   b. Beriberi—vitamin B_1
   c. Wernicke-Korsakoff syndrome—vitamin B_2
   d. Hartnup disease—niacin
6. Which of the following enzymes does not require selenium for functionality?
   a. Glutathione peroxidase
   b. Selenophosphate synthetase
   c. Glutathione reductase
   d. Iodothyronine deiodinase
7. Which of the following best facilitates the diagnosis of copper dysfunction as in Wilson disease?
   a. Liver biopsy
   b. Urine copper output
   c. Urine copper post penicillamine challenge
   d. Serum copper
8. Which of the following conditions does not result in low serum copper and ceruloplasmin?
   a. Menkes syndrome
   b. Wilson disease
   c. Keshan disease
   d. Aceruloplasminemia
9. How would you investigate a patient suspected of taking excess supplements of essential trace elements such as Cu, Zn, and/or Se? By measuring:
   a. different organo inorganic species.
   b. urine output corrected for creatinine.
   c. impact on other analytes for, for example, renal and/or liver function.
   d. whole blood and plasma content of the named traced elements.

SUGGESTED READINGS

1. Shenkin, A. (2003). The role of vitamins and minerals. *Clinical Nutrition*, *22*, S29–S32.
2. Bates, C. J. (1997). Vitamin analysis. *Annals of Clinical Biochemistry*, *34*, 599–626.
3. Taylor, A. (1996). Detection and monitoring of disorders of essential trace elements. *Annals of Clinical Biochemistry*, *33*, 486–510.
4. Gibson, R. S. (2005). *Principles of Nutritional Assessment* (2nd ed.). Oxford: Oxford University Press.
5. Rosenfeld, L. (1997). Vitamine—vitamin. The early years of discovery. *Clinical Chemistry*, *43*, 680–685.
6. Battersby, A. R. (1994). How nature builds the pigments of life: the conquest of vitamin B_{12}. *Science*, *264*, 1551–1557.
7. Ayling, R., & Marshall, W. D. (2007). *Nutrition and Laboratory Medicine*. London: ACB Venture Press.
8. Roberts, N. B., Taylor, A., & Sodi, R. (2018). Vitamins and trace elements. In R. Nader, A. R. Horwarth, & C. T. Wittwer (Eds.), *Tietz Textbook of Clinical Chemistry and Molecular Diagnostics* (6th ed.). (pp. 639–718). St. Louis: Elsevier.

28

Hemoglobin, Iron, and Bilirubin

Christina Ellervik, Maria Domenica Cappellini, Stanley Lo, Dorine W. Swinkels

OBJECTIVES

1. Define the following:
   - Anemia of chronic disease
   - Bilirubin
   - Conjugated (direct) bilirubin
   - Ferritin
   - Hereditary hemochromatosis
   - Hemoglobin
   - Hemoglobinopathy
   - Jaundice
   - Kernicterus
   - Soluble transferrin receptor (sTfR)
   - Thalassemia
   - Transferrin
   - Transferrin saturation
   - Transfusion induced iron overload
   - Unconjugated (indirect) bilirubin
   - Urobilinogen
2. Describe the structure and physiological role of hemoglobin.
3. State the function of hemoglobin and its significance in health and disease.
4. List four different types of hemoglobin and state the makeup of the globin chains in each.
5. List four different types of thalassemia and state the globin deficiency and genetic defect in each.
6. Compare the differences between hemoglobinopathy and thalassemia with regard to causes and laboratory values.
7. List analytical methods used to assess the presence of a hemoglobinopathy or thalassemia.
8. Describe the principle of the cyanmethemoglobin method of hemoglobin determination.
9. Describe iron food sources.
10. Describe iron uptake, transport, distribution, and storage.
11. Describe systemic and cellular iron regulation.
12. List the three major disorders of iron metabolism.
13. Describe and contrast causes, clinical and laboratory findings for iron deficiency, hereditary hemochromatosis, and iron distribution disorder (anemia of chronic disease).
14. Compare the causes of primary, secondary, and juvenile hemochromatosis.
15. State the principle of chromogenic serum iron determination and the clinical significance of the results.
16. Define total iron-binding capacity (TIBC) and state how TIBC value is determined.
17. Describe how to calculate transferrin saturation (TSAT).
18. List analytical methods used to assess serum ferritin, iron, transferrin, hepcidin, and soluble transferrin receptor.
19. Describe how biological and pre-analytical variation impact laboratory findings for serum ferritin, iron, transferrin.
20. Describe the biochemistry of bilirubin including transport in blood, conjugation in the hepatocyte, and hydrolysis in the intestine.
21. List and describe two causes of hyperbilirubinemia in the neonate.
22. State the principle of the diazo method of total bilirubin analysis; state what is required in this reaction to measure direct bilirubin.

KEY WORDS AND DEFINITIONS

Anemia of chronic disease Anemia of inflammation.

Biliary atresia A condition characterized by failure of a fetus to develop an adequate pathway for bile to drain from the liver to the intestine.

Bilirubin A yellow bile pigment that is a breakdown product of heme mainly formed from the degradation of erythrocyte hemoglobin in reticuloendothelial cells.

Carboxyhemoglobin A form of hemoglobin in which the sites usually bound to oxygen are bound to carbon monoxide.

Complete blood count (CBC) The determination of the quantity of each type of blood cell in a milliliter of blood, often including the amount of hemoglobin, the hematocrit, and the proportions of various white cells. In certain countries it is referred to as the "full blood count."

Conjugated bilirubin Bilirubin that has been taken up by the liver cells and conjugated to form the water-soluble bilirubin diglucuronide.

Direct bilirubin The fraction of bilirubin that reacts with the diazo reagent in the absence of alcohol.

Dubin-Johnson syndrome An autosomal recessive disorder that causes an increase of conjugated bilirubin in the serum.

Ferritin The iron/apoferritin complex, which is one of the chief forms in which iron is stored in the body; it occurs in the (1) gastrointestinal mucosa, (2) liver, (3) spleen, (4) bone marrow, and (5) reticuloendothelial cells.

Gilbert syndrome An inborn error of bilirubin metabolism causing predominantly unconjugated hyperbilirubinaemia.

Heme Any quadridentate chelate of iron with the four pyrrole groups of a porphyrin, further distinguished as ferroheme or ferriheme, referring to the chelates of Fe(II) and Fe(III), respectively.

Hemochromatosis A rare genetic disorder caused by deposition of hemosiderin in the parenchymal cells, resulting in tissue damage and dysfunction of the liver, pancreas, heart, and pituitary. Also called *iron overload disease*.

Hemoglobin(Hb) The oxygen-carrying pigment of the erythrocytes, formed by the developing erythrocyte in bone marrow. It is a conjugated protein containing four heme groups and globin, with the property of reversible oxygenation.

Hemoglobinopathy Any inherited disorder caused by abnormalities of hemoglobin, resulting in conditions such as sickle cell anemia, hemolytic anemia, or thalassemia.

Hemosiderin An intracellular storage form of iron; the granules consist of an ill-defined complex of ferric hydroxides, polysaccharides, and proteins having an iron content of about 33% by weight.

Hemosiderosis A focal or general increase in tissue iron stores without associated tissue damage. Hepatic and pulmonary forms of hemosiderosis are characterized by abnormal quantities of hemosiderin in the liver and lungs, respectively.

Hepcidin Hepcidin is a peptide hormone produced by the liver and regulates systemic iron.

Hereditary hemochromatosis A genetically heterogeneous group of inherited disorders of iron metabolism characterized by failure to prevent excessive amounts of iron from entering the circulatory pool and accumulating in the tissues.

Hereditary persistence of fetal hemoglobin A condition characterized by continued production of fetal hemoglobin beyond the point when it is normally replaced by hemoglobin A.

Hyperbilirubinemia Excessive concentrations of bilirubin in the blood, which may lead to jaundice; the hyperbilirubinemias are classified as conjugated or unconjugated, according to the predominant form of bilirubin in the blood.

Icterus A condition characterized by hyperbilirubinemia and the deposition of bile pigments in the skin, mucous membranes, and sclera, with resulting yellow appearance of the patient; also called *jaundice*.

Indirect bilirubin Free bilirubin that has not been conjugated with glucuronic acid.

Jaundice A condition characterized by hyperbilirubinemia and deposition of bile pigment in the skin, mucous membranes, and sclera, with resulting yellow appearance of the patient; also called *icterus*.

Kernicterus A clinical syndrome of the neonate resulting from high concentrations of unconjugated bilirubin that passes the immature blood–brain barrier of the newborn and causes degeneration of cells of the basal ganglia and hippocampus.

Ligandin A hepatic transport protein; its measurement in serum and urine may be a means of estimating the severity of hepatocellular necrosis.

Methemoglobin A form of hemoglobin where its iron atom is changed from the ferrous to the ferric state.

Myoglobin A heme-containing protein found in red skeletal muscle.

Sickle cell anemia An autosomal dominant type of hemolytic anemia that is caused by the presence of hemoglobin S with abnormal sickle-shaped erythrocytes *(sickle cells)*.

Soluble transferrin receptor A circulating truncated form of the transferrin receptor used in the diagnosis of iron deficiency.

Thalassemia A heterogeneous group of hereditary hemolytic anemias having a decreased rate of synthesis of one or more hemoglobin polypeptide chains and classified according to the chain involved (α, β, δ); the two major categories are α- and β-thalassemia.

Thalassemia major The homozygous form of β-thalassemia, a severe condition evident from the neonatal period with complete absence of hemoglobin A.

Thalassemia minor The heterozygous form of β-thalassemia; it is usually asymptomatic, although hemoglobin A synthesis may be retarded and there is sometimes moderate anemia and splenomegaly.

Transferrin A beta globulin that carries iron in the blood.

Transferrin saturation The ratio of serum iron to total iron binding capacity, expressed as a ratio or percentage.

Transfusion-induced iron overload Iron overload caused by transfusion.

Unconjugated bilirubin Free bilirubin that has not been conjugated with glucuronic acid.

Urobilinogen A colorless compound formed in the intestines by the reduction of bilirubin.

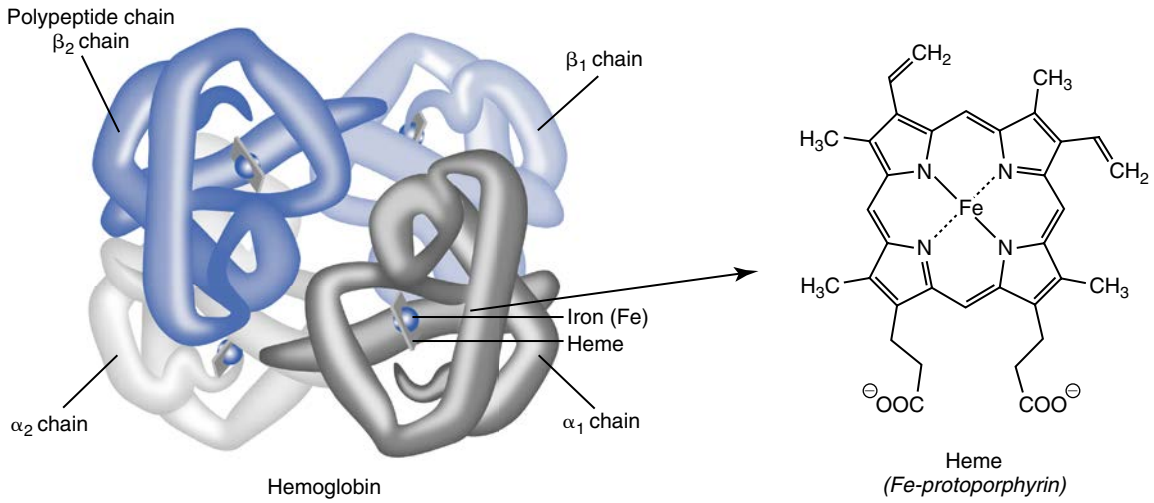


Fig. 28.1 Model of the hemoglobin tetramer with the α-chain subunits facing the reader. Each subunit contains a molecule of heme attached to an atom of iron.

HEMOGLOBIN

Biochemistry

Hemoglobin (Hb) is a globular protein with a diameter of 6.4 nm and a molecular mass of approximately 64,500 Da. Hb consists of four globin subunits: two α- and two non–α-chains (β-, γ-, or δ-chains). The **heme** portion is an iron-containing chelate with the globin portion consisting of two pairs of polypeptide globin chains, designated α and non-α. The four globin chains form a thick-walled globular shell surrounding a central cavity, in which the heme portion, attached by interaction of the histidine residues of the globin chains to the iron in heme, is suspended (Fig. 28.1). The heme iron is normally in the ferrous (2+) oxidation state.

Globin Structure of Normal and Fetal Hemoglobin

In healthy human adults, Hb A is composed of two normal α- and two normal β-polypeptide chains and is represented symbolically as $\alpha_2\beta_2$; it represents at least 96% of the total Hb contained in a sample of whole blood. Hb A_2 is typically approximately 2.5% to 3.0% of total Hb; it contains two α- and two δ-chains and is designated as $\alpha_2\delta_2$. HbA2′ is considered a variant of Hb A_2 and is the result of a glycine-to-arginine substitution in the 16th position of the δ-chain. It rarely forms more than 3% of the total Hb. Fetal Hb (Hb F) predominates during fetal life but rapidly diminishes during the first year of postnatal life. In healthy adults, less than 1% of Hb is Hb F. It consists of two α- and two γ-chains ($\alpha_2\gamma_2$).

Modified Hemoglobins

In addition to the Hbs discussed previously, carboxyhemoglobin, **methemoglobin**, and sulfhemoglobin are other Hbs whose structure has been environmentally or chemically modified.

Each of the modified Hbs has a characteristic spectral pattern, as shown in Fig. 28.2. These spectral characteristics form the basis of analysis in the many co-oximeters and blood gas analyzers that provide, in a single analysis, the simultaneous quantitative measurement of carboxyhemoglobin, methemoglobin, and sulfhemoglobin. The spectral scans are performed using multidiode arrays covering a number of wavelengths, followed by patented calculations that discriminate between normal and modified Hbs.

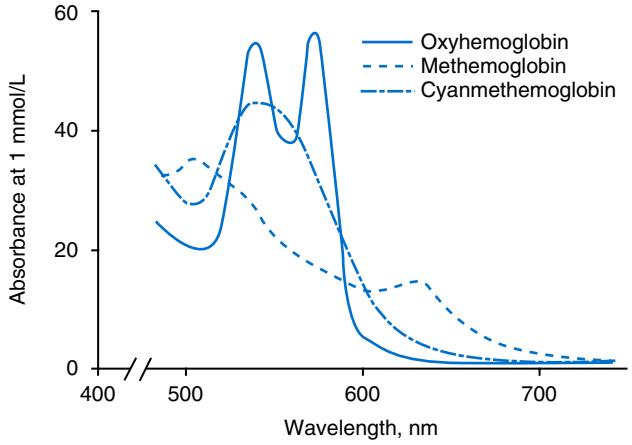


Fig. 28.2 Spectrophotometric absorption curves for oxyhemoglobin, methemoglobin, and cyanmethemoglobin. Oxyhemoglobin and cyanmethemoglobin are used in measuring the concentration of hemoglobin *(Hb)*. The peak at 630 nm, which is distinctive for methemoglobin, is abolished by the addition of cyanide, and the resultant decrease in absorbance is directly proportional to the methemoglobin concentration. All heme proteins exhibit their maximum absorbance in the Soret band region of 400 to 440 nm. Because the absorbance of Hb in the Soret region is approximately 10 times the absorbance at 540 nm, the Soret peaks have been omitted from this diagram. The absorbance curve for methemoglobin is greatly influenced by small changes in pH. The curve given here was obtained at a pH of 6.6.

Carboxyhemoglobin. **Carboxyhemoglobin** is formed by the preferential attachment of carbon monoxide instead of oxygen to Hb, compromising the oxygen-carrying ability of hemoglobin. Carboxyhemoglobin concentrations are increased in smokers and employees with certain

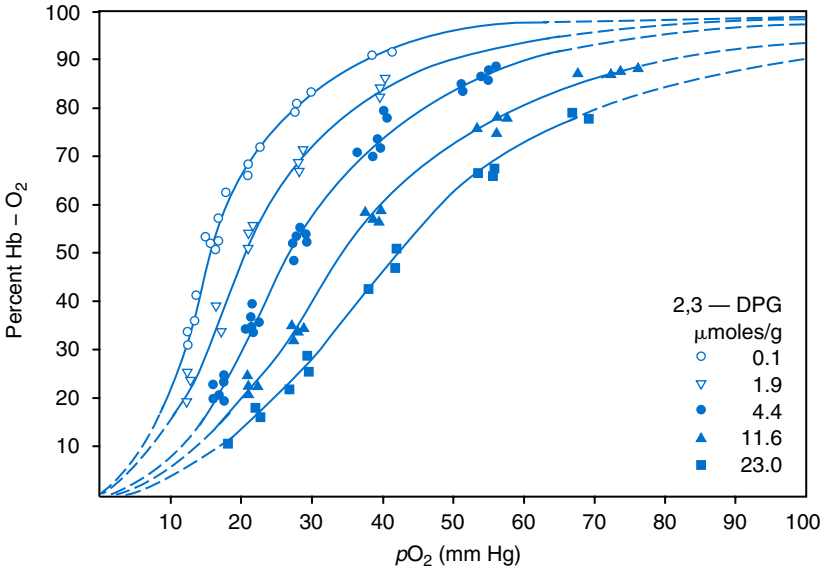


Fig. 28.3 Normal oxygen dissociation curve of hemoglobin *(Hb)*. Changes in 2,3-diphosphoglycerate *(2,3-DPG)* concentration in the erythrocyte greatly influence the position of the curve. As the concentration of 2,3-DPG increases, the curve shifts to the right. pO_2, Partial pressure of oxygen. (From Duhm, J. [1972]. The effect of 2,3-DPG and other organic phosphates on the Donnan equilibrium and the oxygen affinity of human blood. In M. Roth, & P Astrup [Eds], *Oxygen affinity of hemoglobin and red cell acid base status [Alfred Benzon Symposium, IV]*. Copenhagen, Denmark: Alfred Benzon Foundation.)

environmental exposures. Carboxyhemoglobin saturation that varies from 15% to 25% may be associated with dizziness, headaches, and nausea, and greater than 50% saturation is considered life threatening. The half-life of carboxyhemoglobin is 4 to 6 hours.

Methemoglobin. The Fe of heme is normally in the reduced ferrous state (Fe^{2+}). Under alkaline conditions, the Fe is oxidized to the ferric state (Fe^{3+}) by toxic agents, such as nitrates (found in some well waters), aniline dyes, chlorates, drugs (e.g., quinones, phenacetin, and sulfonamides), or local anesthetics (e.g., procaine, benzocaine, and lidocaine). This oxidation converts the heme to hematin and the Hb to methemoglobin. Patients with methemoglobin are cyanotic because methemoglobin is unable to reversibly bind oxygen. Methemoglobin is normally reduced to Hb in the cell by the reduced form of the nicotinamide-adenine dinucleotide–cytochrome reductase system.

Sulfhemoglobin. Sulfhemoglobin is produced by the reaction of sulfur-containing compounds with heme to form an irreversible chemical alteration and oxidation of Hb by the introduction of sulfur in one or more of the porphyrin rings. The most common cause of sulfhemoglobinemia is exposure to drugs, such as phenacetin and sulfonamides. Sulfhemoglobin cannot transport oxygen, and cyanosis is noted at low concentrations.

Physiological Role

The Fe of heme is in the ferrous state (Fe^{2+}) and is able to combine reversibly with oxygen to act as the major oxygen-carrying moiety. The term *cooperativity* is used to describe the interaction of globin chains in such a way that oxygenation of one heme group enhances the probability of oxygenation of the other heme group. The *Bohr effect* refers to reduction of oxygen affinity, with a decrease in pH from the physiologic range (7.35 to 7.45) to 6.0 and is another result of this cooperativity. Because the pH of the tissue decreases as a result of the presence of the end products of anaerobic metabolism, carbon dioxide (CO_2) and carbonic acid, the delivery of oxygen to the exercising tissue is enhanced. The oxygen dissociation curve of normal blood Hb is sigmoidal (Fig. 28.3). Physiologically, the CO_2 reversibly combines with the amino terminal groups of Hb to form carbamated Hb, which facilitates the removal of approximately 10% of the CO_2 that forms because of metabolism in the tissue to the lungs. Removal and transport of CO_2 from the tissue are enhanced by the preference for the attachment of more CO_2 by carbamated Hb.

Clinical Significance

The **thalassemia** syndromes and hemoglobinopathies are clinical disorders related to Hb pathophysiology. Although they may have some similar clinical manifestations, they form two distinct disease groups of genetic origin. The thalassemias originate from insufficient or absent globin chain production due to gene deletion or nonsense mutations, which are mutations that affect the transcription or stability of mRNA products. The name *thalassemia* is derived from the Greek word for "sea," *thalassa,* because early cases of β-thalassemia were described in children of Mediterranean origin.

Hemoglobinopathies, the most common single gene disorder in the world, are structural Hb variants arising from mutations in the globin genes, which result in substitutions or disruptions in the normal amino acid residue sequence in one or more of the globin chains of Hb.

The thalassemias and hemoglobinopathies occur worldwide and are particularly prevalent in South East Asia, Southern China, Mediterranean countries (particularly Greece and the Greek Cypriot part of Cyprus), India, the Middle East, Africa, and the islands of the South Pacific. With increasing migration, the thalassemias—once considered rare genetic diseases in Northern Europe, Australia, and North America—are now becoming more common all over the world.

Thalassemia Syndromes

Thalassemias are identified by the globin chain in which a production deficiency occurs. For example, α- and β-thalassemias result from a deficiency in α- or β-globin chain production, respectively. They are further clinically classified depending on the extent of globin chain production and the resultant severity of the anemia. All the thalassemias have a similar pattern of inheritance: in most cases the gene defects are transmitted in a Mendelian autosomal fashion. Thus, the severe symptomatic varieties result from the interaction of more than one genetic determinant. The inheritance of α-thalassemia is more complicated because it involves the products of the linked pairs of α genes (αα).

Alpha-thalassemias. The α-thalassemias arise from deficiencies in production of the α-globin chains and are caused by deletions or (less frequently) point mutations in one or more of the four α-globin genes. There are two major classes of α thalassemias: α^0 in which both α genes are inactivated (– –/), and α^+ thalassemias, in which only one of the pair is defective either due to α deletion or to α mutation. The conventional nomenclature for the point mutations of an α-gene is "$\alpha^T\alpha$," and for deletion, it is "–α." The clinical spectrum of α thalassemias correlates well with the number of the affected α-genes (i.e., from normal to the loss of all four genes). The inheritance of a normal allele (ααα) with one of the α^+or α^0 results, most frequently, in an "*α-thalassemia minor*" (– –/αα; –α/αα; $-\alpha^T/\alpha\alpha$; $\alpha^T\alpha/-\alpha$; –α/–α) (Table 28.1). The **complete blood count (CBC)** in patients with *α-thalassemia minor* shows a mildly reduced Hb with low mean corpuscular volume (MCV) and mean corpuscular hemoglobin (MCH). Alpha-thalassemias are common in areas where β-thalassemia is also found at a high frequency. Thus, the coinheritance of α- and β-thalassemia traits may occur and even ameliorate the hematological parameters. *Alpha-Thalassemia silent* describes an α-thalassemia trait with a single α-globin chain gene deletion or mutation (–α/αα; $\alpha^T\alpha/\alpha\alpha$).

Hemoglobin Bart's results from deletion of all four α-globin genes, with the subsequent inability to produce any α-globin chains, leading to failure of synthesis of Hb A, F, or A_2. In the fetus, an excess number of γ-globin chains join together to form unstable tetramers known as Hb Bart's (γ^4). Mothers who carry a fetus with Hb Bart's usually present clinically between 20 and 26 weeks' gestation with pregnancy-induced hypertension, polyhydramnios, and hydrops fetalis.

TABLE 28.1 The Alpha-Thalassemias

Condition	Affected Number of Alpha-Globin Genes	Phenotype
Silent carrier	One	Asymptomatic or occasional low red blood cell indices
Alpha-thalassemia trait	Two	Mild microcytic and hypochromic anemia
HbH disease	Three	Mild-to-moderate anemia, splenomegaly, jaundice
Alpha-thalassemia major/Hb Bart's hydrops	Four genes	Most severe form Hb Bart's hydrops Death in utero or at birth

Hb, Hemoglobin.
Data from Vichinsky, E. P. (2009). Alpha thalassemia major–new mutations, intrauterine management, and outcomes. *Hematology American Society of Hematology Education Program*, 35–41.

Hemoglobin H disease. This disorder is usually caused by a three α-globin gene deletion (– –/–α) and is characterized by a chronic anemia of variable severity.

Beta-thalassemias. The β-thalassemias result from a reduction in the synthesis of the β-globin chain. The high frequency of β-thalassemia in the tropics is believed to reflect an advantage of heterozygotes against *Plasmodium falciparum* malaria.

The clinical classification of β-thalassemia includes: **thalassemia major** (TM; transfusion-dependent), thalassemia intermedia (TI; of intermediate severity, nontransfusion-dependent) and **thalassemia minor** (asymptomatic). The severity of the clinical manifestations correlates well with the degree of imbalance of globin chains, depending on the β-globin gene defects and their interaction. The production of β-globin chains is quantitatively reduced to different degrees, whereas the synthesis of α-globin continues as normal, resulting in an accumulation of excess unmatched α-globin chains in the erythroid precursors. Clinical manifestations of β-thalassemia range from mild anemia to severe life-threatening disease that requires lifelong transfusions (Fig. 28.4).

Beta-thalassemia (beta-thalassemia major). Beta-TM results from mutations that interfere with translation or are involved in the initiation, elongation, or termination of globin chain synthesis. Mutations that interfere with translation account for almost 50% of all β-thalassemia mutations. Included in this are frame shift or nonsense mutations that produce premature termination codons, which result in incomplete translation of the β-globin gene and nonproduction of the β-globin chain, leading to β-thalassemia. Clinical presentation usually occurs at younger than 1 year of age, with features such as small size for age, abdominal girth expansion, and failure to thrive. Physical examination of the

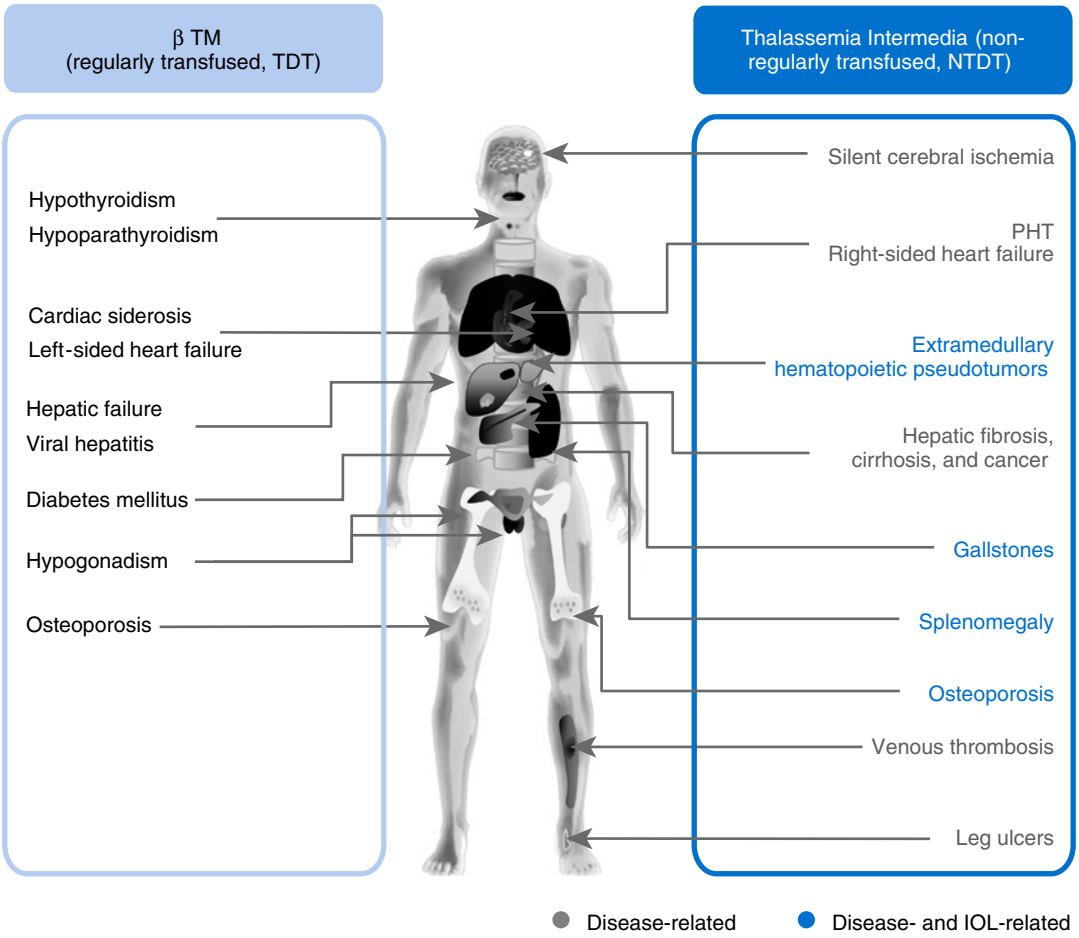


Fig. 28.4 Clinical complications in thalassemia major and in thalassemia intermedia. *IOL*, Iron overload; *NTDT*, non-transfusion-dependent thalassemia; *PHT*, pulmonary hypertension; *TDT*, transfusion-dependent thalassemia; *β-TM*, β-thalassemia major. (Modified from Musallam, K., Rivella, S., Vichinsky, E., & Rachmilewitz, E. A. [2013]. Non-transfusion-dependent thalassemias. *Haematologica, 98*, 833–844.)

patient may reveal frontal bossing (an unusually prominent forehead) caused by thickening of the cranial bones, pallor, and prominence of the cheek bones, which in older children obscures the base of the nose and exposes the teeth. These features are a result of marrow expansion (up to a 30-fold increase) caused by ineffective erythropoiesis with production of highly unstable α-globin tetramers, leading to a sequence of events responsible for bone marrow expansion, anemia, hemolysis, splenomegaly, and increased Fe absorption. Typical CBC results include severe anemia with Hb concentration between 3 and 6.5 g/dL (30 and 65 g/L), MCV of 48 to 72 fL, and MCH of 23 to 32 pg, and varying size of red blood cells (RBCs) (anisocytosis).

Beta-thalassemia intermedia. TI is a clinical term used to describe patients with anemia and splenomegaly, but who do not have the full spectrum of clinical severity found in TM. Most TI patients are homozygotes or compound heterozygotes for mild-to-moderate β-gene mutations (β^+/β^+; β/β^+), less commonly, only a single β-globin gene is affected. Mildly affected patients are almost completely asymptomatic until adult life, experiencing only mild anemia and spontaneously maintaining Hb levels between 7 and 10 g/dL (70 to 100 g/L). In adult life, Hb is significantly reduced to values between 6 and 10 g/dL (60 to 100 g/L)

Beta-thalassemia minor (beta-thalassemia trait). Patients with β-thalassemia minor are often asymptomatic, except at times of hematopoietic stress, such as infection or pregnancy, when they can become more anemic. The CBC on patients with a β-thalassemia trait shows mild anemia with low normal or decreased Hb concentration and hematocrit, decreased MCV (<80 fL) and MCH (<25 pg), and variable red cell distribution widths.

Delta-/beta-thalassemia. The δβ-thalassemias are the result of deletions that affect various parts of the β-globin locus.

Hereditary persistence of fetal hemoglobin. The term **hereditary persistence of fetal hemoglobin (HPFH)** is used to describe a group of genetic conditions in which the concentration of Hb F is increased above the upper limit of the reference interval because of a reduction in β-globin synthesis and a compensatory increase in δ-globin synthesis. HPFH is divided into *deletional* and *nondeletional* HPFH.

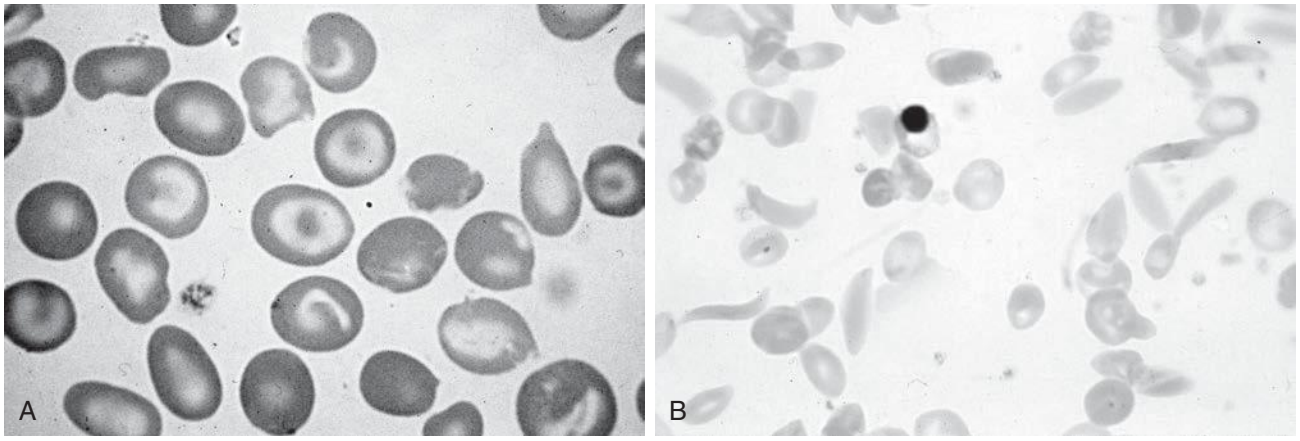


Fig. 28.5 Peripheral blood smear from patients with (A) homozygous hemoglobin (Hb) E and (B) homozygous Hb S.

Hemoglobinopathies

Nomenclature. Hb variants are named using (1) letters (Hbs S, D, E, and so on), (2) the family name of the index case (Hb Lepore), (3) the place of discovery of the variant or place of origin of the propositus (Hb Edmonton), or (4) the name of the river (Hb Saale) flowing through the city in which the propositus lived. In some cases, both a letter and a name are used, as in Hb J-Baltimore, indicating that the Hb is classified as having electrophoretic mobility similar to that of other J Hbs, but differs from them in amino acid sequence, and was originally discovered in Baltimore. The term *AS trait* (sometimes abbreviated to *S trait*) is used to describe a heterozygous state in which one of the β-globin chains is S and the other is A. In instances in which no normal β-globin chain is present (e.g., Hb SD), the β-globin chain present in the higher concentration is usually, although not always, placed first. A systematic nomenclature system is now used alongside the variant name to describe the affected chain and location on the chain and the amino acid substitution. For example, Hb Spanish Town [$\alpha 27(\beta 8)^{Glu \rightarrow Val}$], a Hb variant named after a district in Kingston, Jamaica, and found in Jamaicans of African descent, results from a substitution of valine for glutamic acid in position 27 of the α-globin chain, which is located in position 8 of the B helix of the α-chain.

Classification of hemoglobin variants. Hb variants are classified according to the type of mutation. Single point mutations in α-globin chains give rise to substitution of one amino acid residue. As an example, Hb San Diego [$\beta 109(G11)^{Val \rightarrow Met}$] has a methionine residue instead of the normal valine at position 109 of the β-chain.

Types of hemoglobin variants. Types of Hb variants include (1) deletion, (2) insertion, (3) deletion/insertion, (4) elongation, and (5) hybrid/fusion/crossover Hbs.

Hemoglobin S. Hemoglobin S [$\beta 6(A3)^{Glu \rightarrow Val}$] in the heterozygous or homozygous state is the most widespread of the Hb variants and arises from a substitution of valine for glutamic acid at position 6 in the A helix of the β-globin chain. Approximately 8% of African Americans are heterozygous for Hb S, and homozygous Hb S is found in 1 in 500 newborns in this group. The widespread distribution of the single point gene mutation responsible for the synthesis of Hb S in areas where *P. falciparum* malaria is endemic is due to the protection of Hb S heterozygotes from the worst manifestations of this malaria. The homozygous condition is described as **sickle cell anemia** or sickle cell disease because of the sickle-shaped RBCs that occur when a sickle cell crisis occurs (Fig. 28.5). It is sometimes written as $\beta^S\beta^S$.

Analytical Methods

Thalassemias and hemoglobinopathies are diagnosed and monitored using multiple assays. The methods used are divided into presumptive identification (e.g., CBC, high-performance liquid chromatography [HPLC], and electrophoresis) and definitive identification (DNA sequencing and mass spectrometry). In addition to these tests, the Fe status of the patient should be ascertained by the measurement of ferritin or by the Fe/total Fe-binding capacity/saturation index. Information on the ethnicity and/or nationality of the patient also provides useful information.

Ethylenediaminetetraacetic acid (EDTA) whole blood should be collected and testing should be performed within 5 days of collection, and samples should be stored at 4°C.

Techniques

Analytical techniques used to measure RBCs and their indices, Hb, and related compounds include (1) determination of CBC, (2) electrophoresis, (3) immunoassay, (4) separation techniques such as HPLC, capillary electrophoresis, and mass spectrometry, (5) molecular techniques such as DNA analysis, and (6) specific tests for specific variants.

Complete blood count. A CBC, also known as a full blood count, a full blood examination, or a blood panel, consists of (1) numbers of RBCs (erythrocytes), (2) numbers of white blood cells, (3) numbers of platelets, (4) a measure of Hb, (5) and various cell indices (Table 28.2). Knowledge of red cell indices and the information obtained from microscopic examination of a peripheral blood film is vital to the diagnosis of both α- and β-thalassemias.

TABLE 28.2 Definition of the Parameters That Constitute a Complete Blood Count

Parameter	Definition
WBC	The number of WBCs in the blood
WBC differential count	The number (or percentage) of each type of WBC present in the blood: neutrophils, basophils, eosinophils, lymphocytes, monocytes
RBC count	The number of RBCs in the blood
Hct	The Hct is the proportion of blood volume that is occupied by RBCs
Hb	The protein molecule in RBCs that carries oxygen
MCV	The MCV is the average volume of RBCs
Mean corpuscular Hb	The average amount of Hb in the average RBC
Mean corpuscular Hb concentration	The average concentration of Hb in a given volume of blood
Red cell distribution width	A measurement of the variability of RBC size
Platelet count	The calculated number of platelets in a volume of blood
Platelet distribution width	A measurement of the variability of platelet size
Plateletcrit	The proportion of blood volume that is occupied by platelets
Mean platelet volume	The average volume of platelets

Hb, Hemoglobin; *Hct,* hematocrit; *MCV,* mean corpuscular volume; *RBC,* red blood cell; *WBC,* white blood cell count.

Electrophoresis. Electrophoresis using agarose gel under alkaline conditions (pH 9.2) is the most common initial screening method for the detection and preliminary identification of hemoglobinopathies. Visualization of separated Hb bands is achieved by using a protein-binding stain, such as Amido Black or Ponceau S. Hb bands stain blue with Amido Black and reddish pink with Ponceau S.

At alkaline pH, Hbs migrate according to electrical charge, with Hb H moving the fastest (closest to the anode). The order of migration (fastest to slowest) is Hb H, Hb N, Hb I, Hb J, Hb A, Hb F, Hb S, and Hb C. Hb D and Hb G comigrate with Hb S, and Hb E, Hb O, and Hb A_2 comigrate with Hb C. Hemoglobin Constant Spring migrates slightly toward the cathode. An easy way to remember the sequence is Hb A goes to the anode, whereas Hb C migrates to the cathode. Hb F and Hb S follow after Hb A in alphabetical order.

Electrophoresis at pH 6.4 using a citrate buffer is performed when an abnormal band is noted on alkaline Hb electrophoresis. The same Hb variants separated on agarose electrophoresis at pH 6.4 and stained with acid violet are shown in the *right panel* of Fig. 28.6. The order of migration (cathode to anode, fastest to slowest) is Hb F, Hb A, Hb S, and Hb C. Hb D, Hb G, Hb I, Hb J, Hb O, Hb A_2, and Hb E comigrate with Hb A.

Specific types of electrophoresis that are used for Hb analysis include isoelectric focusing, electrophoresis, and capillary electrophoresis.

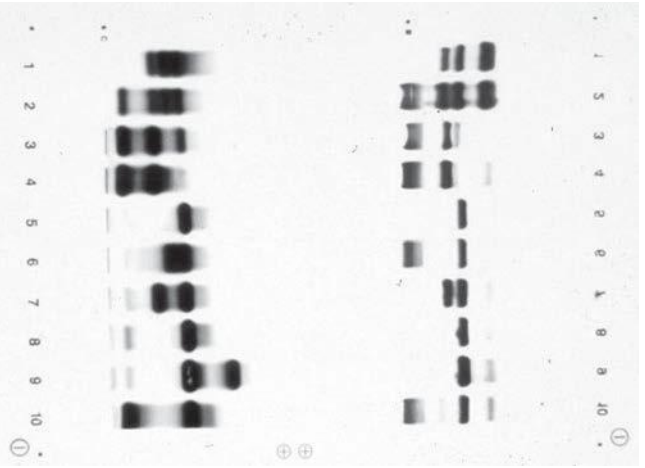


Fig. 28.6 Alkaline *(left)* and acid *(right)* electrophoresis of various hemoglobinopathies. Lane 1, hemoglobin *(Hb)* S, Hb FA control. Lane 2, HB S, Hb F, HbCA control. Lane 3, transfused sickle cell *(SC)* disease. Lane 4, SC disease. Lane 5, Hb A (normal). Lane 6, Hb Presbyterian. Lane 7, Hb S. Lane 8, raised Hb A2 (β-thalassemia trait). Lane 9, Hb J Baltimore. Lane 10, Hb C.

Isoelectric focusing electrophoresis. Isoelectric-focusing electrophoresis (IEF) has greater resolving power than conventional electrophoresis, but it is more expensive, time-consuming, and technique-dependent to perform. Commercial IEF gels are made of cellulose acetate or polyacrylamide with the pH gradient produced by the inclusion of amphoteric materials of different pHs in bands in the gel. Locations of the Hb bands are identified using stains similar to those used in conventional electrophoresis. The bands or zones produced by IEF (Fig. 28.7) are more clearly defined than those seen with conventional electrophoresis, and reliable quantification of separated Hbs may be made at high concentrations using densitometry.

Capillary isoelectric focusing electrophoresis. Capillary IEF combines the detection sensitivity of capillary electrophoresis with the resolution qualities and existing extensive data on Hb variant separation by immunoelectrophoresis and the automated sampling and digital data acquisition techniques developed for chromatography. With this approach, the hemolysate is introduced into the capillary chamber using low-pressure injection and then is focused at high voltage (typically ≈30 kV and 0.5 to 1.5 μA), during which it is essential to maintain adequate cooling. The separated Hbs are then eluted, using low-pressure and simultaneous voltage, past a single-wavelength spectrophotometric detector set to read at 415 nm or a dual-wavelength detector set at 415 and 450 nm.

Capillary electrophoresis. The introduction of commercial capillary electrophoresis instrumentation for the separation of Hbs has made this technique available to clinical laboratories. Separation in an alkaline buffer at a specific pH using high voltages is based on charge difference, electrolyte pH, and electro-osmotic flow. Hb measurement is commonly performed at a wavelength of 415 nm, and identification is based on retention time.

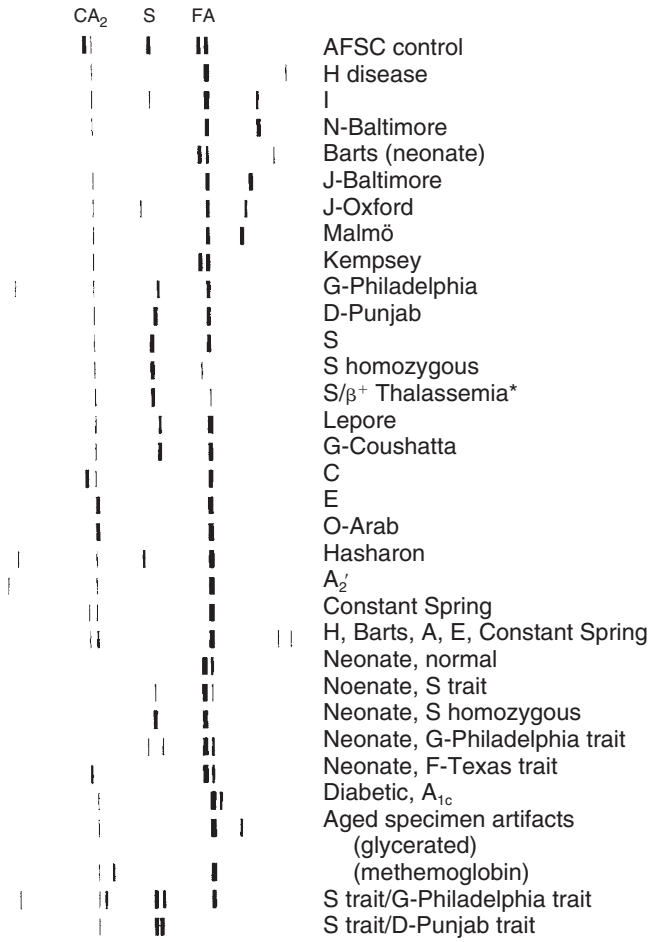


Fig. 28.7 A diagram of isoelectric focusing patterns for a variety of hemoglobin *(Hb)* variants. The conditions shown represent heterozygotes (traits), unless otherwise indicated. The width of the bars approximates the relative density of the bands observed. The acid anodic (pH 6) side is to the *right*, and the alkaline cathodic (pH 8) side is to the *left*. The same pattern is observed in homozygous patients with HbS disease who have received HbA by transfusion.

High-performance liquid chromatography. HPLC using a column packed with cation-exchange resin provides—in a single analytical protocol—the quantification of Hb F and Hb A_2. With it, the initial identification of an Hb variant on the basis of elution time may be made. It has also been used to detect α-thalassemia phenotypes.

After injection and subsequent adsorption onto the particles of a cation-exchange resin, molecules of Hb are eluted using gradient elution. The detection of eluted Hbs is achieved by monitoring the effluent solvent stream using a dual-wavelength photometer (usually set to measure at wavelengths of 415 and 690 nm) (Fig. 28.8).

Electrospray Mass Spectroscopy

Electrospray mass spectrometry is becoming an accepted method for the characterization of Hb variants involving substitution of one or more amino acids in the globin protein chains. Tandem mass spectrometry (MSMS) has been used for newborn screening for sickle cell disease and thalassemias.

Several mass spectrometry-based approaches, sometimes using laborious preparative steps, have been described. (1) Measurement of the mass of intact globin chains and the relative proportion of β and δ globin chains. Alpha or β variants are followed up by peptide-based analysis, based on tryptic digestion of whole blood, using electrospray MSMS in multiple reaction monitoring (MRM) mode to confirm identity. (2) Ultra-performance liquid chromatography coupled with MSMS using predetermined tryptic peptides to screen for clinically important structural variants. (3) Mass measurement of intact globin chains without complex pre-analytical sample digestion/separation procedures followed by top-down electron transfer dissociation (ETD) fragmentation of selected globin chains for the confirmation of variants. These techniques offer a faster and more specific detection and sequence confirmation of clinically significant Hb variants than conventional methods, but they require special instrumentations and skills for interpretation.

DNA Analysis

DNA analysis is used in the investigation of thalassemias and hemoglobinopathies to identify specific individuals at risk and those who may benefit from genetic counseling in populations with a known high incidence of disease. For example, DNA analysis has been used to do the following:

1. Diagnose α^0- and α^+-thalassemia.
2. Investigate potentially life-threatening disorders of Hb synthesis in the fetus; it is performed at less than 10 weeks' gestation on chorionic villus samples.
3. Characterize the β-thalassemia genotype.
4. Screen at-risk populations for clinically significant Hb variants.
5. Distinguish between conditions that have similar laboratory and clinical presentations but are the result of different genetic conditions.

Specific Tests

Tests that are used to measure Hbs and related analytes include those for Hb H, Hb S, unstable Hbs, and globin chains.

Determining hemoglobin H. Hb H, an insoluble tetramer consisting of four β-globin chains due to decreased production of α-globin chains. If the tetramers are oxidized, precipitation occurs, which may be viewed microscopically. In the laboratory, this oxidation is achieved by staining unfixed cells with freshly prepared methylene blue or brilliant cresyl blue at 37°C.

Sickling and hemoglobin S solubility tests. Sickling tests are useful in confirming the presence of Hb S in a sample after initial electrophoresis at alkaline pH. When Hb S is oxygenated, it is fully soluble. When Hb S is deoxygenated, polymerization occurs, forming deformed red cells with a characteristic rigid sickle shape. In the laboratory, deoxygenation and lysis of RBCs is achieved using a solution of sodium metabisulfite in a phosphate buffer. The addition of the sodium metabisulfite reagent to an Hb S-containing blood sample induces the typical sickle shape of RBCs (which is the basis of the sickling test by microscopic investigation

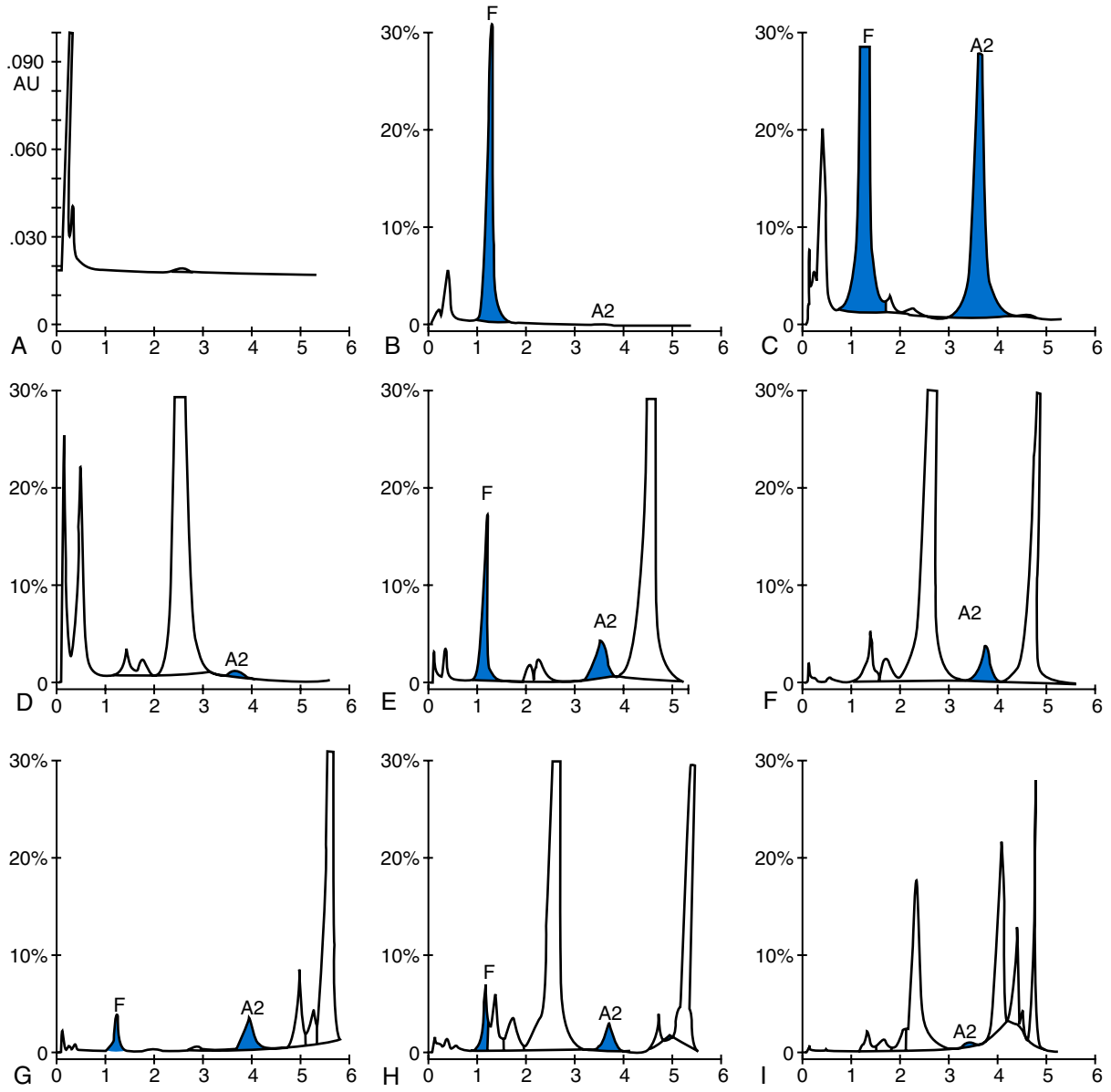


Fig. 28.8 High-performance liquid chromatography chromatograms obtained on the bio-rad variant β-thal short program for (A) hemoglobin *(Hb)* Bart's; (B) β0-thalassemia major; (C) B+-thalassemia homozygous E; (D) HbH; (E) homozygous S; (F) S trait; (G) homozygous C; (H) C trait; and (I) HbS-HbG Philadelphia. (From Clarke, G. M., Trefor, N., & Higgins, T. N. [2000]. Laboratory investigation of hemoglobinopathies and thalassemias: review and update. *Clinical Chemistry, 46*, 1284–1290.)

of the blood film preparation), and it also causes turbidity (which is the basis of the solubility test). In the Hb S solubility test, this turbidity is visualized by holding a lined card or a card with writing on it behind the reaction test tube (Fig. 28.9). A positive sample is not transparent, whereas a negative sample is. False negative results are obtained on anemic patients (Hb <8 g/dL; or <80 g/L) or on samples with hematocrit less than 15% (0.15). False positive results are found in samples with (1) Heinz bodies, (2) high concentrations of monoclonal protein, (3) hyperlipidemia, or (4) cold agglutinins. The test is subjective, and the combination of two identification techniques, such as HPLC and alkaline electrophoresis, may eliminate the necessity to perform this test on a routine basis.

Tests for unstable hemoglobins. These tests use heat or isopropanol to precipitate the unstable Hb and must be performed on fresh blood. More than 100 unstable Hbs are mainly the result of the interchange of nonpolar amino acid residues for polar amino acid residues in positions in the α- or β-globin chain associated with the heme cleft.

IRON

Iron is involved in the function of all cells. It can accept and donate electrons, depending on its oxidation state: ferrous iron (Fe[II]) or ferric iron (Fe[III]). Iron is mostly locked into Fe protoporphyrin (heme) and Fe–sulfur clusters, which serve as enzyme cofactors. Hemoproteins are involved in numerous

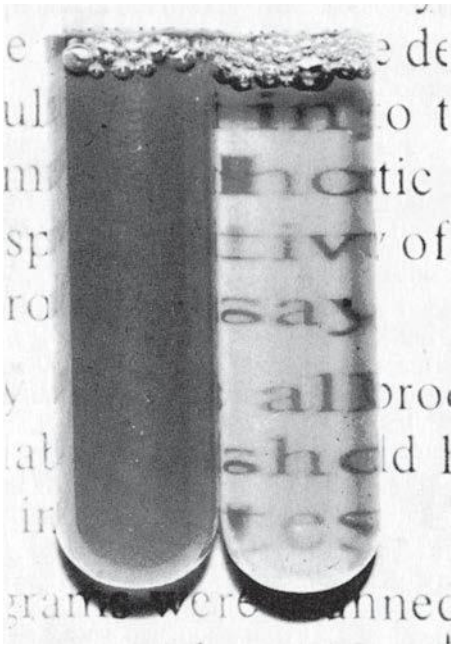


Fig. 28.9 Solubility test for hemoglobin *(Hb)* S. Deoxyhemoglobin S *(left tube)* is insoluble in 2.3 mol/L phosphate buffer. In contrast, normal hemolysate *(right tube)* is sufficiently transparent that print can easily be read through it.

biological functions, such as oxygen binding and transport (Hbs), oxygen metabolism (catalases, peroxidases), cellular respiration, and electron transport (cytochromes). Proteins containing nonheme Fe are important for fundamental cellular processes such as DNA synthesis, cell proliferation and differentiation (ribonucleotide reductase), gene regulation, drug metabolism, and steroid synthesis. However, Fe can also cause damage, because Fe(II) catalyzes the generation of highly reactive hydroxyl radicals (·OH) from hydrogen peroxide (H_2O_2) ($Fe^{2+} + H_2O_2 \rightarrow Fe^{3+} + HO^- + HO\cdot$), which is called the "Fenton reaction." These hydroxyl radicals damage cellular membranes, proteins, and DNA. A large number of scavenger molecules protect cells against Fe-mediated tissue damage. Proteins sequester Fe to reduce this threat. Iron circulates in a form bound to plasma **transferrin**, which is needed to offer the highly insoluble Fe(III) to cells via the transferrin receptor. Iron can safely be stored within cells in the form of ferritin and hemosiderin. Under normal circumstances, only small amounts of Fe exist outside this physiologic sink, although stored Fe can be mobilized for reuse. Many diseases arise from imbalances in Fe homeostasis. Too much Fe accumulates in **hereditary hemochromatosis** (HH) and in the Fe-loading anemias, which are often aggravated by multiple transfusions. In iron-deficiency anemia (IDA), insufficient amounts of Fe are available for heme synthesis. In **anemia of chronic disease** (ACD), Fe is redistributed to macrophages to promote resistance to infections.

Iron Metabolism

Systemic and Cellular Iron Regulation

The control of Fe homeostasis acts at both the cellular and the systemic level, and involves a complex system of different cell types, transporters, and signals. To maintain systemic Fe homeostasis, communication between cells that absorb Fe from the diet (duodenal enterocytes), consume Fe (mainly erythroid precursors), and store Fe (hepatocyte and tissue macrophages) must be tightly regulated. Each of these cell types plays an essential role in the homeostatic Fe cycle. Hepcidin controls Fe absorption and macrophage Fe release. Hepcidin is synthesized in the liver upon changes in body Fe needs, anemia, hypoxia, and inflammation, and is secreted in the circulation. It counteracts the function of ferroportin, a major cellular Fe exporter protein in the membrane of macrophages and the basolateral site of enterocytes, by inducing its internalization and degradation Fig. 28.10.

Body Iron Distribution

See Table 28.3 for average distribution of body iron.

Ferritin. **Ferritin** consists of a protein shell surrounding an iron core. Ferritin consists of an apoferritin shell composed of 24 subunits that are either L (light) or H (heavy) ferritin chains, and an interior ferric oxyhydroxide (FeOOH) crystalline core. Ferritin occurs in nearly all body cells and stores iron in a form that is shielded from body fluids and is therefore unable to cause oxidative damage. Ferritin in hepatocytes and macrophages provides a reserve of iron that is available for synthesis of Hb and other heme proteins. In healthy men, ≈1000 mg of storage iron is present, mostly as ferritin. In healthy premenopausal women, storage iron ranges up to ≈300 mg. Relative to tissues, minute quantities of relatively iron-poor ferritin are also present in serum in concentrations roughly proportional to total body iron stores. However, when elevated, serum ferritin does not always correlate with body iron content. Relatively large amounts of ferritin may be released into plasma as the result of diverse pathological conditions among which are infections, chronic inflammatory diseases, chronic liver diseases, the metabolic syndrome, and nonalcoholic steatohepatitis (NASH).

Hemosiderin. **Hemosiderin**, a terminus of the intracellular storage iron pathway, is formed when ferritin is aggregated and partially deproteinized in secondary lysosomes. Like ferritin, hemosiderin is usually found predominantly in cells of the (1) liver, (2) spleen, and (3) bone marrow. Unlike ferritin, hemosiderin is insoluble in aqueous solutions.

Tissue iron. Numerous cellular enzymes and coenzymes require iron as an integral part of the molecule or as a cofactor. These include *peroxidases* and *cytochromes,* all of which are heme proteins, like Hb. Other enzymes, such as *aconitase* and *ferredoxin,* contain iron that is coordinated with sulfur in a so-called iron–sulfur cluster. These enzymes and coenzymes occur in all nucleated cells of the body and form the *tissue iron compartment,* normally ≈150 mg of iron (see Table 28.3).

Myoglobin. **Myoglobin** very closely resembles a single Hb subunit and contains a single heme per molecule.

The labile iron pool. A fraction of the circulating non-transferrin-bound iron (NTBI) is redox-active and

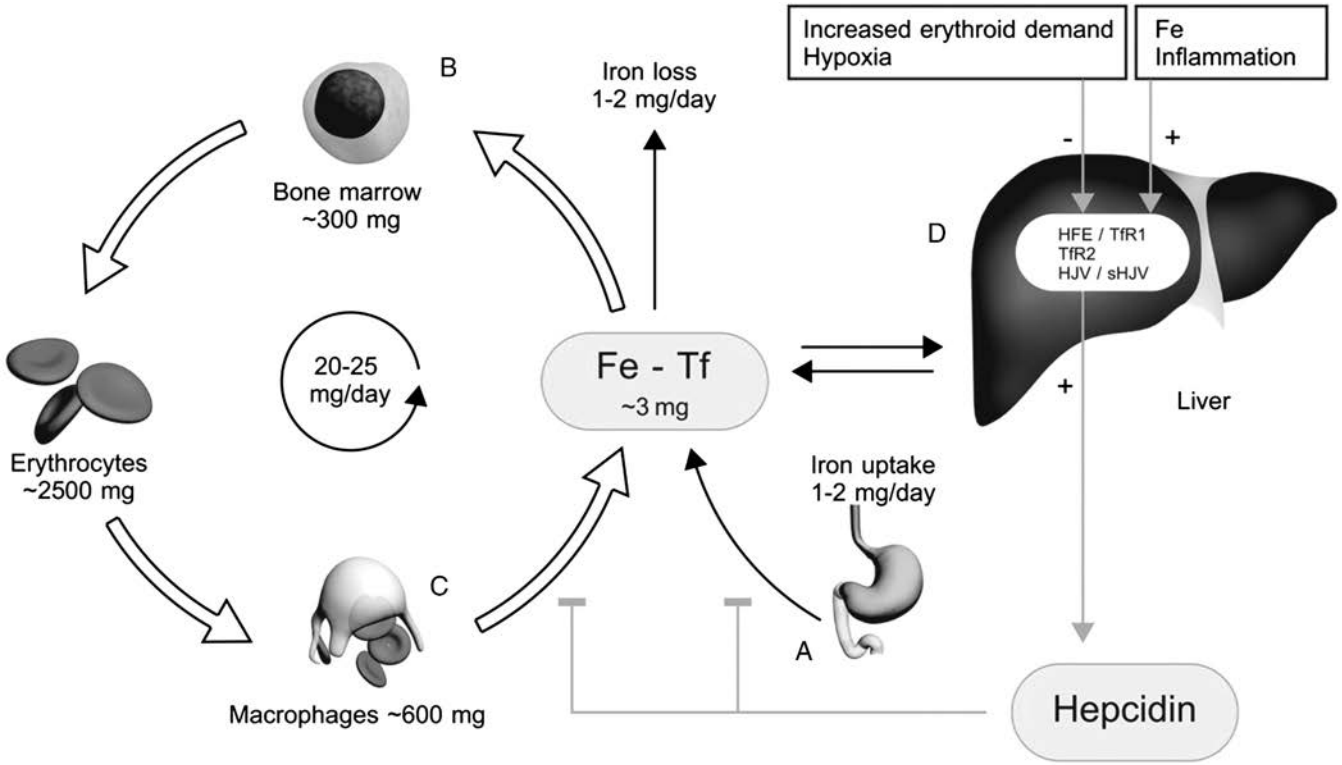


Fig. 28.10 Systemic iron (Fe) homeostasis. The largest flux of Fe takes place in the recycling of Fe from senescent erythrocytes out of (A) macrophages to incorporation in (B) erythroid precursors in the bone marrow. Values for the different tissues and fluxes are approximate. The (D) liver and (C) reticuloendothelial macrophages function as major iron stores. Only 1 to 2 mg Fe is absorbed and lost every day. Importantly, the total amount of Fe in the body can only be regulated by absorption, whereas Fe loss occurs only passively from sloughing of skin and mucosal cells, as well as from blood loss. Hepcidin, a recently identified antimicrobial, β-defensin-like peptide secreted by the liver, decreases the plasma Fe concentration by inhibiting Fe export by ferroportin from (A) duodenal enterocytes and (C) reticuloendothelial macrophages. As a consequence, an increase in hepcidin production results in a decrease in plasma Fe concentrations. Hepcidin expression of the hepatocyte is regulated by circulating and stored Fe, inflammatory stimuli, the erythroid Fe demand, and hypoxia, by pathways involving expression of *HFE, TFR2,* hemojuvelin *(HJV)*, and *TMPRSS6* genes. In HFE-, transferrin receptor (TfR)-2, and HJV-related hereditary hemochromatosis, hepcidin production is low despite increased liver Fe, which results in inappropriately increased Fe absorption. Defects in the *TMPRSS6* gene, encoding for matriptase-2, result in Fe refractory iron-deficiency anemia. The letters *A*, *B*, *C*, and *D* refer to sites with special functions in iron metabolism. (From Swinkels, D. W., Janssen, M. C., Bergmans, J., & Marx, J. J. [2006]. Hereditary hemochromatosis: genetic complexity and new diagnostic approaches. *Clinical Chemistry, 52*, 950–968.)

TABLE 28.3 Iron Content of Compartments in an Average Male

Compartment	Iron Content (mg)
Hemoglobin	2500
Storage iron[a]	1000
Myoglobin	130
Other cellular iron-containing proteins	150
Transport (transferrin)	3
Total	3000–5000

[a]Sum of parenchymal and reticuloendothelial iron stored in ferritin or hemosiderin.

designated labile plasma Fe. Although there are methods for measuring serum NTBI and labile plasma Fe, insufficient standardization and clinical correlation currently limit clinical use of these measurements.

Iron transport. Iron is transported from one organ to another in the plasma by the transport protein *apotransferrin.* The apotransferrin/Fe^{3+} complex is called mono-ferric or diferric (=holo) transferrin. When transferrin binds to the *transferrin receptor* of cells, the transferrin/receptor complex is internalized into an endosome that becomes acidified, releasing the iron from transferrin. Iron is then converted to ferrous iron (Fe^{2+}) by a ferroreductase, and transported to the cytosol via the divalent metal transporter DMT1. The apotransferrin is then transported back to the cell surface, ready to transport another transferrin molecule to the interior of the cell. This series of reactions has been designated the *transferrin cycle.*

Regulation of iron homeostasis

Systemic iron regulation. The total body iron content can only be regulated by absorption from the upper intestinal tract, whereas iron loss occurs only passively from sloughing

of skin and mucosal cells, as well as from blood loss. **Hepcidin**, a peptide hormone produced by the liver, regulates systemic iron. It decreases the iron absorption from the diet as well as iron export from macrophages. As a consequence, elevated hepcidin levels result in a decrease of plasma iron concentration.

Normal iron balance. The average daily American diet contains 10 to 15 mg of iron, much of it in the form of (1) heme proteins, (2) Hb, and (3) myoglobin in meat. Normally, ≈1 to 2 mg of iron is absorbed each day, mostly by the duodenum. To be absorbed, inorganic iron must be in the ferrous state (Fe^{2+}). Dietary heme is taken up via an unknown transport system.

Proteins that affect iron homeostasis. Many proteins influence iron homeostasis. Mutations in genes that encode these proteins may cause several inherited iron disorders, among which are iron overload or iron deficiency. Table 28.4 summarizes the effects of some of these proteins on iron homeostasis.

Clinical Significance

Iron deficiency, Fe overload, and ACD are the most prevalent disorders of Fe metabolism. Iron disorders may be classified according to (1) pathophysiology (see Table 28.4), (2) heritability (genetic or acquired), or (3) as being "primary," (e.g., resulting from a defect in an Fe metabolism-related protein) or "secondary" (e.g., the consequences of defects in other proteins).

Iron Deficiency

Iron deficiency and IDA—defined when anemia and Fe deficiency coexist—are global health problems and common medical conditions seen in everyday clinical practice.

Iron deficiency is particularly a disorder of children and premenopausal women in low- and middle-income countries, but it can also occur in men, and in developed countries in people of all ages. In children, Fe deficiency is frequently caused by the increased physiologic needs for dietary Fe for growth and development. In adults, and especially premenopausal women, Fe deficiency is almost always the result of chronic blood loss or childbearing.

Iron-deficiency anemia is associated with impaired quality of life, work productivity, aerobic exercise capacity, and fatigue, and there is also a relationship between Fe status and depression and cognitive functioning. Iron deficiency

TABLE 28.4 Features of Iron Metabolism Disorders

Disorder	Gene	Inheritance	Age at Presentation	Systemic Iron Overload	Anemia	Ferritin	TSAT
Impaired Hepcidin-Ferroportin Axis							
HH type 1	*HFE*	AR	Adult	Variable	No	Variable	High
HH type II A	*HAMP*	AR	Child	Yes	No	High	High
HH type II B	*HJV*	AR	Child to young adult	Yes	No	High	High
HH type III	*TfR2*	AR	Young adult	Yes	No	High	High
HH type IVA	*SLC40A1* (LOF[a])	AD	Adult	Yes	Sporadically	High	Normal
HH type IV B	*SLC40A1* (GOF)	AD	Adult	Yes	No	High	High
Iron transport disorder							
Low iron availability for erythropoiesis							
IRIDA	*TMPRSS6*	AR/AD		No	Yes	Low–normal	Low
Aceruloplasminemia	*CP*	AR		Yes	Yes	High	Normal/low
Anemia of chronic disease	NA	NA	Variable	No[a]	Yes	Normal–high	Low
Defect in iron acquisition of erythroid progenitor cells							
Hypotransferrinemia	*Tf*	AR		Yes	Yes	High	100%
Microcytic anemia with iron loading	*DMT1*	AR		Yes	Yes	Variable	High
Erythroid dysmaturation							
Nontransfusion-dependent thalassemias (β-thalassemia intermedia; Hb H disease)	Globin	AR	Variable	Yes	Yes	High	High

Continued

TABLE 28.4 Features of Iron Metabolism Disorders—cont'd

Disorder	Gene	Inheritance	Age at Presentation	Systemic Iron Overload	Anemia	Ferritin	TSAT
Sideroblastic anemia congenital							
X-linked sideroblastic anemia	*ALAS-2*	XL	Child	Variable	Mild/no anemia	Variable	Variable
Sideroblastic anemia	*SLC25A38*	AR	Child	Yes	Severe	High	High
Sideroblastic anemia	*GLRx5*	AR	Adult	Yes	Mild	High	High
Sideroblastic anemia with ataxia	*ABCB7*	XL	Child	No	Mild	Normal	Normal
Myelodysplastic syndrome (RA, RARS)	NA	NA	Adult	Variable	Variable	Variable	Variable
Congenital dyserythropoietic anemia Type Ia, Ib, II, III, IV	*CDAN1, C15ORF41, SEC23B, KIF23, KLF1*	Variable	Child	Variable	Yes	Variable	High
Other iron disorders							
Localized iron overload							
Neurodegeneration with brain iron accumulation							
Neonatal hemochromatosis	NA	NA	Neonate	Yes	No	Yes	Yes
Iron overload in sub-Saharan Africa	Unclear	Unclear	Adult	Yes	No	Yes	Yes
Hyperferritinemia-cataract syndrome	*FTL*-IRE	AD	Variable	No	No	High	Normal
Chronic liver disease Metabolic syndrome Steatohepatitis Cirrhosis	NA	NA	Variable	No/mild	Variable	Variable/high	Variable
Parenteral iron loading/chronic erythrocyte transfusion	NA	NA	Variable	Yes	No	High	High
Iron deficiency (anemia)	NA	NA	Variable	No	Variable	Low	Low

[a]Often referred to as "ferroportin disease."
GOF, Gain of function mutation; *HH*, hereditary hemochromatosis; *IRE*, iron response element; *IRIDA*, iron refractory iron-deficiency anemia; *LOF*, loss of function mutation; *RA*, refractory anemia; *RARS*, refractory anemia with ringed sideroblasts; *TSAT*, transferrin saturation.

also affects immune function and the susceptibility to infections. Furthermore, Fe deficiency is a predictor of mortality in patients with chronic kidney disease and heart failure; treating Fe deficiency in patients with heart failure improves quality of life.

Iron-deficiency anemia is generally an acquired condition due to blood losses, malabsorption, insufficient Fe intake, or a combination of these. Some patients with chronic disease are especially vulnerable to develop absolute Fe deficiency (e.g., patients with gastrointestinal tumors because of blood losses). Patients with advanced chronic kidney disease also have a negative Fe balance as a result of reduced dietary intake, impaired absorption from the gut, and increased Fe losses. In addition, patients with chronic kidney disease also often have a functional Fe deficiency because the available circulating Fe cannot keep pace with the requirements of erythropoiesis.

In patients with iron refractory iron-deficiency anemia (IRIDA), an inherited form of IDA due to a defect in the *TMPRSS6* gene encoding for the protein matriptase-2, the iron deficiency does not improve with oral Fe supplementation. This disorder is an indication for intravenous iron treatment.

Many different measurements have been advocated for the diagnosis of Fe deficiency. Overall, a low ferritin level is a sensitive and specific indicator of Fe deficiency uncomplicated by other diseases. A **transferrin saturation** (<15% to 20%) indicates an Fe supply that is insufficient to support normal erythropoiesis (Tables 28.5 and 28.6). Hb below normal combined with iron deficiency defines IDA. **Soluble transferrin receptor (sTfR)** is helpful in the investigation of the pathophysiology of anemia, to distinguish IDA from ACD and to detect functional iron deficiency. In the different stages of iron deficiency, the various biochemical and hematological tests efficiently complement each other and help characterize the severity in the individual patient (Fig. 28.11).

TABLE 28.5 Laboratory Tests for Iron Status in Adults (Standard International Units)

Iron Status Test	Iron Deficiency	Functional Iron Deficiency	Iron-Deficiency Anemia	IRIDA	Anemia of Chronic Disease	Iron-Deficiency Anemia and Anemia of Chronic Disease	Iron Overload in Hereditary Hemochromatosis	Reference Interval
Current								
Iron (μmol/L)	Low	Low–normal	Low	Low	Low	Low	High	13–36[a]
Transferrin (g/L)/TIBC (μmol/L)	High	Normal	High	High	Low	Variable	Low	2.0–3.6 45–78[a]
Transferrin saturation (%)	15–20[a]	15–20[a]	<15[a]	<10[b]	<15–20[a]	<15–20[a]	>45[a,c]	15–45[a,c]
Ferritin (μg/L)	<12–30[a,d]	<100–200[a]	<12[a]	Variable[e]	<100[a]	<100[a]	Men: >300 Women: >200	Men: 40–300[a,f] Women: 20–200[a,f]
Hb (mmol/L)	Normal	Normal	Low	Low	Low	Low	High–normal	Men: >8.1[f] Women: >7.4[f]
MCV (fL)	Normal	Normal	<80	Low	Low–normal	Low	High-normal	80–95[f]
MCH (fmol)	Normal	Normal	<27	Low	Low–normal	Low	High-normal	1.67–2.11[f]
Proposed								
Hepcidin	Low	Low	Very low	High for TSAT[g]	High	Normal–high	Low for ferritin[h]	Varies[a]
sTfR (mg/L)	High	High	High	High	Low–normal	Variable	Low–normal	Varies[a]
sTfR/log ferritin	NA	NA	>2[a,i]	NA	<1[a,i]	>2[a,i]	NA	Varies[a]
ZnPP mmol/mol heme	High	High	High	High	High	High	Low–normal	<40–80[a]
Reticulocyte hemoglobin content (fmol)	<1.74[a,j]	<1.80[a,k]	Low	Low	Low–normal	Low	High–normal	1.87–2.23[a,l]
Bone marrow iron[m]	Negative	Variable	Negative	Positive	Positive	Positive	Positive	Positive

[a]Results vary with the methodology used.
[b]In more severe cases, less than 5%.
[c]Some guidelines propose a threshold of 50% for men.
[d]In populations exposed to infections and in patients with renal failure, inflammatory bowel disease, chronic heart failure, or other (low grade) inflammatory diseases, threshold values indicating iron deficiency are generally considered to be higher than in those without these diseases (see the following). In these situations, levels more than 100 μg/L generally excludes absolute iron deficiency and for levels between 30 and 100 μg/L, other parameters are needed to diagnose iron deficiency.
[e]Mostly low-normal in untreated patients.
[f]Based on Camaschella, C. (2015). Iron-deficiency anemia. *New England Journal of Medicine, 372*, 1832–1843.
[g]Absolute levels vary, but in the absence of inflammation. levels are high for circulating iron levels.
[h]Absolute levels vary, but levels are low for serum ferritin levels.
[i]Value from (a) Weiss, G., & Goodnough, L. T. (2005). Anemia of chronic disease. *New England Journal of Medicine, 352*, 1011–1023; and (b) Punnonen, K., Irjala, K., & Rajamaki, A. (1997). Serum transferrin receptor and its ratio to serum ferritin in the diagnosis of iron deficiency. *Blood, 89*, 1052–1057.
[j]For CHr, based on Mast, A. E., Blinder, M. A., Lu, Q., et al. (2002). Clinical utility of the reticulocyte hemoglobin content in the diagnosis of iron deficiency. *Blood, 99*, 1489–1491.
[k]For CHr, from Thomas, D. W., Hinchliffe, R. F., Briggs, C., et al. (2013). Guideline for the laboratory diagnosis of functional iron deficiency. *British Journal of Haematology, 161*, 639–648.
[l]For CHr, from Piva, E., Brugnara, C., Spolaore, F., et al. (2015). Clinical utility of reticulocyte parameters. *Clinics in Laboratory Medicine, 35*, 133–163.
[m]By Perls staining of bone marrow for iron in reticulocytes according to Gale and colleagues.[319]
Hb, Hemoglobin; *IRIDA*, iron refractory iron-deficiency anemia; *MCH*, mean corpuscular Hb; *MCV*, mean corpuscular volume; *sTfR*, soluble transferrin receptor; *TIBC*, total iron-binding capacity; *TSAT*, transferrin saturation; *ZnPP*, zinc protoporphyrin.

TABLE 28.6 Laboratory Tests for Iron Status in Adults (Traditional Units)

Iron Status Test	Iron Deficiency	Functional Iron Deficiency	Iron-Deficiency Anemia	IRIDA	Anemia of Chronic Disease	Iron-Deficiency Anemia and Anemia of Chronic Disease	Iron Overload in Hereditary Hemochromatosis	Reference Interval
Current								
Iron (μg/dL)	Low	Low-normal	Low	Low	Low	Low	High	70–200[a]
Transferrin (g/L)	High	Normal	High	High	Low	Variable	Low	2.0–3.6
TIBC (μg/dL)								251–436[a]
Transferrin saturation (%)	15–20[a]	15–20[a]	<15[a]	<10[b]	<15–20[a]	<15–20[a]	>45[c]	15–45[a,c]
Ferritin (μg/L)	<12–30[d]	<100–200[a]	<12[a]	Variable[e]	<100[a]	<100[a]	Men: >300 Women: >200	Men: 40–300[a,f] Women: 20–200[a,f]
Hb (g/dL)	Normal	Normal	Low	Low	Low	Low	High–normal	Men: >13[f] Women: >12[f]
MCV (fL)	Normal	Normal	<80	Low	Low–normal	Low	High–normal	80–95[f]
MCH (pg)	Normal	Normal	<27	Low	Low–normal	Low	High–normal	27–34[f]
Proposed								
Hepcidin	Low	Low	Very low	High for TSAT[g]	High	Normal–high	Low for ferritin[h]	Varies[a]
sTfR (mg/L)	High	High	High	High	Low–normal	Variable	Low–normal	Varies[a]
sTfR/log ferritin	NA	NA	>2[a,i]	NA	<1[a,i]	>2[a,i]	NA	Varies[a]
ZnPP mmol/mol heme	High	High	High	High	High	High	Low–normal	<40–80[a]
Reticulocyte hemoglobin content (pg)	<28[a,j]	<29[a,k]	Low	Low	Low–normal	Low	High–normal	30.2–35.9[a,l]
Bone marrow iron[m]	Negative	Variable	Negative	Positive	Positive	Positive	Positive	Positive

[a]Results vary with the methodology used.
[b]In more severe cases, less than 5%.
[c]Some guidelines propose a threshold of 50% for men.
[d]In populations exposed to infections and in patients with renal failure, inflammatory bowel disease, chronic heart failure or other (low grade) inflammatory diseases, threshold values indicating iron deficiency are generally considered to be higher than in those without these diseases (see the following). In these situations, levels more than 100 μg/L generally exclude absolute iron deficiency and for levels between 30 and 100 μg/L, other parameters are needed to diagnose iron deficiency.
[e]Mostly low-normal in untreated patients.
[f]Based on Camaschella, C. (2015). Iron-deficiency anemia. *New England Journal of Medicine, 372*, 1832–1843.
[g]Absolute levels vary, but in the absence of inflammation. levels are high for circulating iron levels.
[h]Absolute levels vary, but levels are low for serum ferritin levels.
[i]Value from (a) Weiss, G., Goodnough, L. T. (1997). Anemia of chronic disease. *New England Journal of Medicine, 352*, 1011–1023; and (b) Punnonen, K., Irjala, K., Rajamaki, A. (1997). Serum transferrin receptor and its ratio to serum ferritin in the diagnosis of iron deficiency. *Blood, 89*, 1052–1057.
[j]For CHr, based on Mast, A. E., Blinder, M. A., Lu, Q., et al. (2002). Clinical utility of the reticulocyte hemoglobin content in the diagnosis of iron deficiency. *Blood, 99*, 1489–1491.
[k]For CHr, from Thomas, D. W., Hinchliffe, R. F., Briggs, C., et al. (2013). Guideline for the laboratory diagnosis of functional iron deficiency. *British Journal of Haematology, 161*, 639–648.
[l]For CHr, from Piva, E., Brugnara, C., Spolaore, F., et al. Clinical utility of reticulocyte parameters. *Clinics in Laboratory Medicine, 35*, 133–163.
[m]By Perls staining of bone marrow for iron in reticulocytes according to Gale and colleagues.[319]
Hb, Hemoglobin; *IRIDA*, iron refractory iron-deficiency anemia; *MCH*, mean corpuscular Hb; *MCV*, mean corpuscular volume; *sTfR*, soluble transferrin receptor; *TIBC*, total iron-binding capacity; *TSAT*, transferrin saturation; *ZnPP*, zinc protoporphyrin.

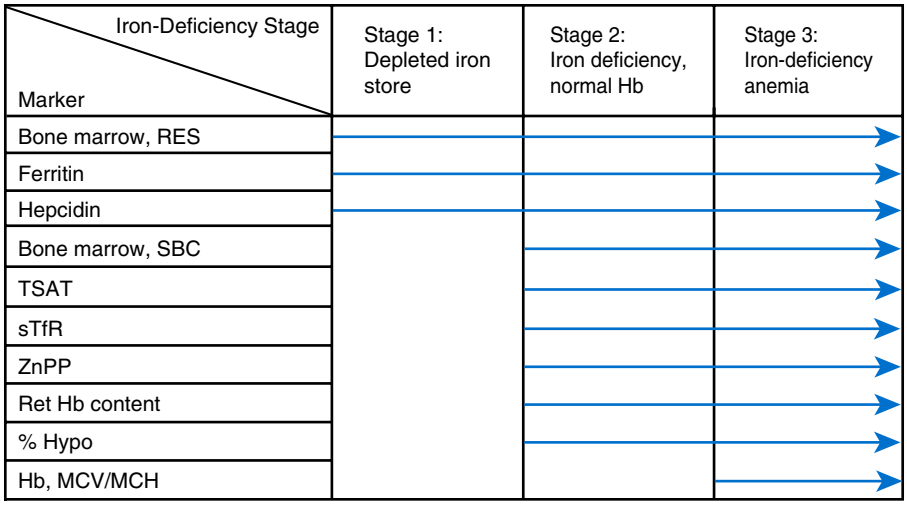


Fig. 28.11 Alterations in biochemical and hematological parameters at different stages of iron deficiency. Iron-deficiency stage 1: bone marrow hemosiderin within reticuloendothelial system *(RES)* cells and ferritin reflect the iron stores and allow the diagnosis of iron depletion. In this stage, hepcidin is low (mostly undetectable) to increase maximal iron uptake from the diet and release from the RES stores. Iron-deficiency stage 2: pathological values of transferrin saturation *(TSAT)*, bone marrow sideroblast count *(SBC)*, soluble transferrin receptor *(sTfR)*, zinc protoporphyrin *(ZnPP)*, reticulocyte hemoglobin *(Ret-Hb)* content, percentage of hypochromic cells *(% Hypo)* indicate that iron-deficient erythropoiesis has already developed. Iron-deficiency stage 3: Hb below reference values defines iron-deficiency anemia, which is accompanied by low red cell indexes. *MCH*, Mean corpuscular hemoglobin; *MCV*, mean corpuscular volume. (Modified from Hastka, J., Lasserre, J. J., Schwarzbeck, A., Reiter, A., & Hehlmann, R. [1996]. Laboratory tests of iron status: correlation or common sense? *Clinical Chemistry, 42*, 718–724.)

Anemia of Chronic Disease

ACD, also named anemia of inflammation, is a common disorder. It is often observed in patients with infectious and inflammatory diseases, among whom are patients with chronic kidney disease, inflammatory bowel disease, chronic heart failure, malignancies, and hepatic diseases. It is an acquired Fe distribution disorder with relatively low circulating Fe levels, despite the presence of adequate total body Fe stores. The pathogenesis of ACD includes shortened erythrocyte survival, impaired marrow function, and disturbances of Fe metabolism. In ACD, the release of Fe from cells to the circulation is inadequate to support heme formation, and Fe is sequestered in reticuloendothelial macrophages. The ACD is typically normocytic and normochromic. Patients with ACD have anemia, low Fe, low transferrin, low transferrin saturation (TSAT), normal or moderately elevated ferritin, elevated hepcidin, and mostly normal sTfR concentrations.

Iron Overload

Iron overload disorders are typically insidious, causing progressive and sometimes irreversible tissue damage before clinical symptoms develop. Fe overload disorders can be categorized according to whether the underlying pathophysiological defect is in the hepcidin-ferroportin axis, erythroid maturation, or in Fe transport. There are also some less common disorders that do not fit into one of these categories (see Table 28.4). Iron overload may also develop as a consequence of multiple RBC transfusions and parenteral Fe supplementation.

Disorders of the *hepcidin-ferroportin* axis. Each of the six disorders in this group has a primary form of Fe overload and is a subtype of HH (primary) (see Table 28.4). Five of the six different HH disorders may lead to a classical HH phenotype: normal Hb, elevated ferritin and TSAT, and tissue Fe overload (see Tables 28.5 and 28.6). The pathophysiology of these five conditions is similar, with an increased Fe absorption exceeding the needs of the body, which leads to increased Fe stores due to inadequate hepcidin-mediated down-regulation of ferroportin. This leads to Fe deposition in parenchymal organs (e.g., the liver and the pancreas).

Initial clinical symptoms of tissue Fe overload are often nonspecific and vague: fatigue and joint pain. In later stages, disease manifestations may include diabetes mellitus, hypogonadism and other endocrinopathies, liver cirrhosis, cardiomyopathy, skin pigmentation, and in cirrhotic patients, increased susceptibility of liver cancer. Early diagnosis and therapeutic phlebotomy can prevent the development of tissue damage, reduce morbidity and mortality, and provide long-term survival similar to the general population. Phlebotomy is the treatment for HH.

The most common form of HH is attributed to variants in the *HFE*-gene. The best described variants in HFE-hemochromatosis are (1) the variant in the cDNA of *HFE* at position 845, a G to A nucleotide change resulting in a cysteine to tyrosine replacement (p.Cys282Tyr or C282Y); and (2) the variant at *HFE* cDNA position 187, a C to G nucleotide change resulting in a histidine to asparagine replacement in the protein (p.His63Asp or H63D). Of the Caucasian population, approximately 1 in 200 are *HFE* C282Y homozygotes,

and approximately 1 in 10 carries the variant. In patients with Fe overload and of European ancestry, nearly 80% have been found to be homozygous for the C282Y variant in *HFE*. A smaller proportion (5%) are compound heterozygous for the C282Y/H63D variants. Although 38% to 76% of C282Y homozygous people have been found to develop raised serum Fe parameters, such as ferritin and TSAT, disease penetrance is 2% to 38% in men and 1% to 10% in women, presumably because menstrual blood loss and childbearing protects them from Fe overload. Also, both disease severity and clinical expression correlate well with the degree of Fe overload, but are heterogeneous among patients, depending not only on the age of diagnosis but also on other genetic and dietary modifiers that remain still largely unknown. Among C282Y/H63D compound heterozygotes the risk of disease progression is low, and documented Fe overload disease is rare. Furthermore, C282Y/H63D compound heterozygous hemochromatosis patients with clinical disease expression, frequently have additional risk factors for Fe overload or liver disease. Homozygosity for H63D is also rarely associated with **hemochromatosis** and is not considered a disease-associated genotype. Most patients with HFE-associated HH do not present until middle age (and women not until after menopause).

Disorders of erythroid maturation. This class of disorders represents forms of secondary Fe overload, and includes the Fe-loading anemias, among which are thalassemia syndromes (especially the β-thalassemias), sideroblastic anemias, and the congenital dyserythropoietic anemias. These diseases are characterized by ineffective erythropoiesis (i.e., by apoptosis of erythroid precursors), failure of erythroid maturation, and consequent expansion of the number of erythroid precursor cells in the bone marrow.

Disorders of iron transport and iron acquisition. The common pathophysiologic feature of these disorders is insufficient delivery and acquisition of transferrin-bound Fe to the RBC precursors in the bone marrow for heme synthesis, despite Fe stores. The resulting Fe-restrictive erythropoiesis, anemia, and hypoxia all contribute to low hepcidin-induced Fe overload.

Other forms of iron overload. Chronic erythrocyte transfusion is the most common cause of this group of disorders. **Transfusion-induced iron overload** is common in patients with severe β-TM and sickle cell disease. Transfusion-induced iron overload also occurs in persons with (1) severe aplastic anemia, (2) Blackfan-Diamond syndrome, (3) Fanconi anemia, (4) acute leukemia, (5) autoimmune hemolytic anemias, and (6) myelodysplasia with refractory anemia.

Rarely, iron overload develops in persons who ingest large quantities of supplemental iron. Iron overload also occurs in persons with renal insufficiency as a result of excessive intravenous supplements.

Analytical Methods

Several methods are used to measure iron and related analytes. These include methods for (1) serum iron, (2) iron-binding capacity, (3) transferrin saturation, (4) serum ferritin, (5) hepcidin, (6) sTfR, and (7) red cell indices.

Methods for Serum Iron, Total Iron-Binding Capacity, Transferrin, and Transferrin Saturation

Except when atomic absorption spectroscopy is used, hemolysis has very little effect on serum iron assay results because Hb iron is not released from heme by acid treatment. Markedly hemolyzed specimens should be rejected for analysis. Because iron is ubiquitous, scrupulous care is necessary to ensure that glassware, water, and reagents do not become contaminated with extraneous iron.

Iron is released from transferrin by decreasing the pH of the serum. It is reduced from Fe^{3+} to Fe^{2+} and then is complexed with a chromogen such as bathophenanthroline or ferrozine.

Such iron/chromogen complexes have extremely high absorbance at the applicable wavelength, which is proportional to the iron concentration.

Transferrin can be measured directly by immunological techniques. Alternatively, in mostly Anglo-Saxon countries, transferrin is quantified in terms of the amount of iron it will bind, a measure called total iron binding capacity (TIBC).

Serum unsaturated iron-binding capacity (UIBC) and TIBC are determined by the addition of sufficient Fe^{3+} to saturate iron-binding sites on transferrin. Excess Fe^{3+} is removed, e.g., by adsorption with (1) light magnesium carbonate ($MgCO_3$) powder, (2) a silica column, or (3) anion exchange resin, and the assay for iron content is then repeated. From this second measurement, the TIBC is obtained.

Many automated chemistry analyzers now measure UIBC and calculate TIBC, rather than measuring it directly. UIBC is measured by adding a known excess concentration of iron to serum. By leaving the pH near neutral, only the iron that did not bind to transferrin is measured upon the addition of an iron-binding chromogen. TIBC is calculated by adding the UIBC to the serum iron level.

The combination of total serum Fe and either TIBC or transferrin measurement allows the calculation of the saturation of transferrin with Fe:

$$\text{Transferrin saturation } (\%) = (100 \times \text{serum iron})/\text{TIBC}$$

Transferrin concentration can be used to derive TIBC, and indirectly, TSAT. Because 1 mol of transferrin (average molecular mass 79,570 Da) has the capacity to bind two atoms of Fe (atomic mass 55.84 Da), the theoretically derived formulas for calculating transferrin (and TSAT) from TIBC and vice versa are:

$$\text{Transferrin}\left(\frac{g}{L}\right) = 0.007 \times \text{TIBC}\left(\frac{\mu g}{L}\right)$$

$$\text{TIBC}\left(\frac{\mu g}{dL}\right) = \text{transferrin}\left(\frac{mg}{dL}\right) \times 1.41$$

$$\text{TIBC}\,\frac{\mu mol}{L} = \text{ltransferrin}\left(\frac{g}{L}\right) \times 25.2$$

$$\left(\begin{array}{l} \text{TSAT}(\%) = \dfrac{([\text{SerumFe}(\mu g/dL)])/}{([\text{transferrin}(mg/dL)])} \times 70.9 \\[2ex] \text{TSAT}(\%) = \dfrac{([\text{SerumFe}(\mu mol/L)])/}{([\text{transferrin}(mg/dL)])} \times 398 \end{array}\right)$$

These mathematical derivations have some weaknesses. The results of immunological measurements of transferrin concentration correlate with those of the TIBC assay. Note that the relationship is not entirely linear as a small portion of iron in serum is bound to proteins other than transferrin. Therefore calculated TIBC values are slightly greater than the amount of transferrin-bound iron. However, these small differences are of no practical or clinical consequence.

Reference intervals. Reference intervals for serum iron differ by as much as 35% between commercial methods. Therefore, a generic reference interval cannot be applied, and each laboratory should independently define its own reference intervals that are appropriate for the laboratory's method and local population. The reference value used for TSAT also differs between laboratories. The approximate and most used reference interval is 15% to 45%. There is day-to-day variation of serum iron and TSAT in many healthy people.

Clinical relevance. The serum iron concentration refers to the Fe^{3+} bound to serum transferrin and does not include the iron contained in serum as free Hb and ferritin. In clinical situations, serum Fe measurements are often combined with TIBC or transferrin, to calculate the TSAT that is considered more useful for clinical interpretation.

The serum iron concentration or TSAT is decreased in many, but not all, patients with IDA and with chronic inflammatory disorders (see Table 28.4). Elevated serum iron or TSAT occur (1) in iron overload such as HH, (2) in patients with aplastic anemia, (3) in children with acute iron poisoning, (4) after oral or parenteral iron use, and (5) as the result of acute liver injury.

Because only about one-third of the iron-binding sites of transferrin are occupied by Fe^{3+} in healthy individuals, serum transferrin has much reserve iron-binding capacity (UIBC). TIBC is a measurement of the maximum concentration of iron that transferrin binds. TIBC or transferrin is increased in subjects with iron deficiency and is decreased in those with chronic inflammatory disorders, chronic liver diseases, and malnutrition. In many persons with untreated HH, TIBC is slightly decreased.

Methods for Serum Ferritin

Serum ferritin assay may be performed by several commercially available methods, including: (1) immunoradiometric assay, (2) enzyme linked immunosorbent assay (ELISA), (3) immunochemiluminescence assays, and (4) immunofluorometric methods.

Reference intervals. Reference intervals for serum ferritin concentrations are summarized in Tables 28.5 and 28.6. Reference intervals and diagnostic cutoffs differ distinctly between laboratories. This can be attributed to variation in definition of reference population, variation in assay techniques and limited use of reference materials for calibration.

Clinical relevance. Under normal conditions serum ferritin concentrations are usually proportional to the body's iron content. The circulating protein is largely apoferritin. The plasma ferritin concentration declines very early in the development of iron deficiency—long before changes are observed in (1) blood Hb concentration, (2) RBC size, or (3) serum iron concentration. Thus, the measurement of serum ferritin concentration is used as a highly sensitive indicator of iron deficiency that is uncomplicated by other concurrent disease. On the other hand, many chronic diseases cause increased serum ferritin levels (see Table 28.4). Elevated ferritin has a low specificity for parenchymal iron overload and most subjects with elevated serum ferritin levels do not have HH or other iron overload disorders.

Since an elevated TSAT facilitates parenchymal iron deposition, the combination of high TSAT with hyperferritinemia can be observed in HH and transfusion-induced iron overload.

In subjects with proven hemochromatosis or iron overload, serum ferritin measurements are useful gauges to the progress of phlebotomy therapy in achieving iron depletion.

Methods for Hepcidin

Currently, several in-house and commercial immunochemical assays and laboratory-developed mass spectrometry assays have been shown to provide reliable hepcidin results. Some of the mass spectrometry assays have been thoroughly validated and are used in patient care, but most assays are labeled for "research use only."

Reference interval. Because hepcidin assays are not harmonized and standardized, there are no universal reference intervals or decision limits.

Clinical relevance. The measurement of hepcidin has provided important insights into the pathophysiology of Fe disorders, among which are HH, anemia of inflammation, and IRIDA. Proof-of-concept studies highlight hepcidin as a promising tool in the diagnosis of IRIDA, and the guidance of iron supplementation therapy. The recent identification of commutable reference material, allows future harmonization.

Method for Soluble Serum Transferrin Receptor

Soluble or serum transferrin receptor (sTfR1 or sTfR)—a truncated form of the transferrin receptor—is a single polypeptide chain with a molecular mass of 84.9 kDa that can be measured by a variety of standard immunoassay techniques. The automated homogeneous immunoturbidimetric assay is characterized by a high sample throughput and excellent precision. In addition, sTfR concentrations can be assessed by fluoroimmunoassays and immunofluorometric assays, which are based on immunoreactants labeled with fluorescent probes. The high sensitivity of fluorescence measurement combined with the sensitivity of the probe to changes in its environment offered the possibility of monitoring the concentration of sTfR directly in the reaction mixture.

Reference intervals. Current commercial methods for the assay of sTfR are characterized by different reference intervals.

Clinical relevance. Soluble TfR is useful in the investigation of the pathophysiology of anemia, to distinguish IDA from ACD and to detect functional Fe deficiency. Cell membranes of developing erythroid cells, especially the erythroblasts, in bone marrow are rich in transferrin receptors (TfRs) to which the Fe-transferrin complex binds. The number of TfRs increases in the presence of high erythroid proliferation

rates and low Fe supplies, and decreases in bone marrow hypoplasia and suppression. These variations in the quantity of TfRs in erythropoietic tissue are also reflected in changes in sTfR.

Red Blood Cell Indices

MCV, MCH, and MCH concentration (MCHC) are indices that are included in the CBC test. One of their uses is to characterize the RBCs of patients with anemia. Low values of MCV and MCH are indicators of IDA. MCV, MCH, and MCHC are slightly higher for untreated HH patients compared to healthy individuals.

Gold Standard Assessment of Body Iron

Bone marrow examination. The absence of stainable bone marrow Fe is the most commonly accepted reference criterion of Fe deficiency. One way of evaluating Fe stores is by examining a bone marrow aspirate for hemosiderin within reticuloendothelial cells. In Fe deficiency, marrow hemosiderin is absent; in the anemia of chronic disorders, Fe is always present. Fe stores are greatly increased in TM and in sideroblastic anemias. However, this method for quantification of Fe stores may be misleading when Fe stores are not normally distributed between reticuloendothelial and parenchymal tissues as occurs in ACD. Another way of evaluating bone marrow Fe includes evaluation of Fe in the erythroblasts, which represents the useable and functional Fe stores. Differentiation between functional Fe deficiency and quantitative deficiency of Fe stores is important, especially in areas of high infection pressure. Bone marrow examination for the sole purpose of assessing Fe stores is rarely justifiable. It may be helpful in the diagnosis of unexplained forms of anemia and if there are concerns that a high serum ferritin value is not a true reflection of body Fe stores.

BILIRUBIN

Hb, iron, and **bilirubin** are analytes that collectively may be viewed in terms of a manufacturing process in which the raw material (iron) is incorporated with other raw materials in a multistage complex process leading to a finished product (Hb). This finished product has a limited life span, after which degradation into the waste product (bilirubin) occurs. Within this process, many exquisite mechanisms are found: (1) biochemical control, (2) conservation, and (3) synthesis. For example, this process may be disrupted by (1) a deficiency in the supply of raw material, (2) a lack of control or synthesis mechanisms, (3) excessive loss of finished product, or (4) excessive conversion to or deficiency in the elimination of waste products. These disruptions lead to the clinical disorders of (1) hemoglobinopathies, (2) thalassemias, (3) IDAs, (4) liver disease, and (5) various genetic diseases.

Bilirubin is the orange-yellow pigment derived mainly from senescent *(aging)* RBCs. It is extracted and metabolized mainly in the liver and is excreted in bile and urine. The chemistry, biochemistry, clinical significance, and analytical methods for bilirubin are briefly reviewed in this section.

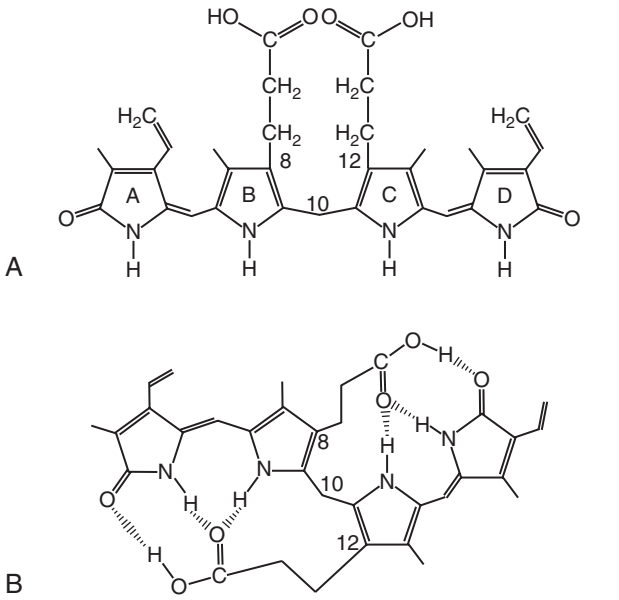


Fig. 28.12 Bilirubin IXα structure. (A) A linear molecular representation of unconjugated bilirubin. (B) The preferred structure of unconjugated bilirubin IXα, Z-Z configuration. The structure is stabilized by hydrogen bonding. The "ridge" involves carbon atoms 8 through 12.

Chemistry

The proposed structure of bilirubin is shown in the upper portion of Fig. 28.12. However, the insolubility of the bilirubin molecule in water and its solubility in a variety of nonpolar, lipid solvents are not predicted from this linear tetrapyrrole structure. X-ray crystallography showed bilirubin assuming a ridge-tiled configuration, stabilized by six intramolecular hydrogen bonds. This Z-Z *(trans)* conformation (see Fig. 28.12) prevents the interaction of bilirubin with polar groups in aqueous media. When exposed to light, the Z-Z configuration is converted by rupture of the intramolecular double bonds to the E-E *(cis)* conformation, which is more water soluble than the Z-Z conformation. Thus light-exposed forms of bilirubin are more water soluble and are readily excreted in the bile. This is the rationale for irradiating jaundiced newborns with 450 nm light.

Biochemistry

Bilirubin IXα is produced from the catabolism of heme. For each mole of heme, one mole each of carbon monoxide, bilirubin, and ferric iron is produced. Daily bilirubin production from all sources in humans averages from 250 to 300 mg. Approximately 85% of the total bilirubin produced is derived from the heme moiety of Hb released from senescent erythrocytes that are destroyed in the reticuloendothelial system. The remaining 15% is produced from RBC precursors destroyed in the bone marrow and from the catabolism of other heme-containing proteins.

In blood, bilirubin is bound to albumin and is transported to the liver. It dissociates from albumin at the membrane of the hepatocyte and is transported across the membrane into the liver (Fig. 28.13). Inside the liver cells, bilirubin is

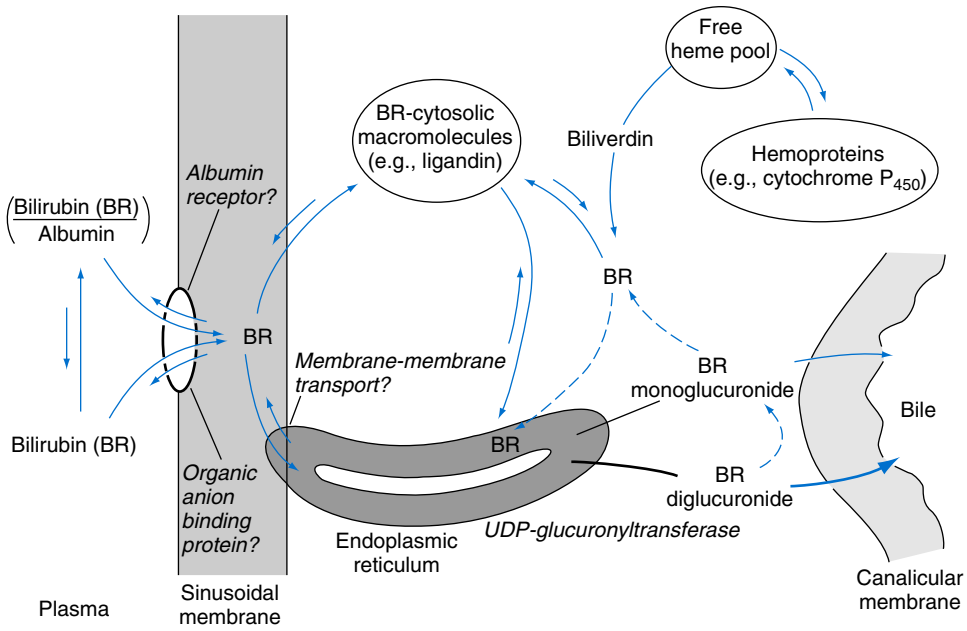


Fig. 28.13 At a glance. Bilirubin uptake, metabolism, and transport in the hepatocyte. (From Gollan, J. L., & Schmid, R. [1982]. Bilirubin update: formation, transport, and metabolism. In H. Popper, & F. Schaffner [Eds.], *Progress in liver diseases*. Vol 7. Chapter 15. Philadelphia: WB Saunders.)

reversibly bound to soluble proteins known as **ligandins** or *protein Y*. Inside the hepatocytes, bilirubin is rapidly conjugated with glucuronic acid to produce bilirubin monoglucuronide and diglucuronide, which then are excreted into bile (see Fig. 28.13). The microsomal enzyme bilirubin uridine diphosphate (UDP)–glucuronyltransferase (EC 2.4.1.17) catalyzes the formation of bilirubin monoglucuronide. In adults, virtually all bilirubin excreted in bile is in the form of glycosidic conjugates.

In the intestine, bilirubin glucuronides are hydrolyzed by the catalytic action of β-glucuronidase from the (1) liver, (2) intestinal epithelial cells, and (3) bacteria. The **unconjugated bilirubin** is then reduced by anaerobic intestinal microbial flora to form a group of three colorless tetrapyrroles collectively called **urobilinogens**. In the lower intestinal tract, the three urobilinogens are spontaneously oxidized to produce the bile pigments, which are orange–brown and the major pigments of stool.

Clinical Significance

Jaundice is a condition characterized by **hyperbilirubinemia** and deposition of bile pigment in the (1) skin, (2) mucous membranes, and (3) sclera, with a resulting yellow appearance of the patient; it is also called **icterus**. Defects in bilirubin metabolism resulting in jaundice occur at each step of the metabolic pathway (see Fig. 28.13). Bilirubin disorders are characterized by elevations in conjugated or unconjugated bilirubin in the absence of other abnormal liver tests. For more information on disorders of bilirubin metabolism and other conditions associated with increased bilirubin, readers are referred to the 6th edition of the *Tietz Textbook of Clinical Chemistry and Molecular Diagnostics*.

Jaundice in the Neonate

Disorders that cause jaundice in the neonate are classified as unconjugated or conjugated hyperbilirubinemia (Box 28.1).

Unconjugated hyperbilirubinemia. Unconjugated hyperbilirubinemia poses a risk for development of **kernicterus** (acute bilirubin encephalopathy), especially in low-birth-weight infants. It is possible to prevent this syndrome by phototherapy and exchange transfusion in infants with elevated unconjugated bilirubin concentrations. Causes of unconjugated hyperbilirubinemia in the neonate are (1) physiological jaundice of the newborn, (2) hemolytic disease owing to Rh or ABO incompatibility, and (3) breast milk hyperbilirubinemia.

Physiological jaundice of the newborn. Babies frequently become jaundiced within a few days of birth; this condition is known as *physiological jaundice of the newborn*. Unconjugated bilirubin concentrations reach a peak within 3 to 5 days of birth and remain elevated for less than 2 weeks. Factors contributing to physiological jaundice include (1) increased bilirubin load in the newborn caused by the shortened life span of RBCs; (2) decreased conjugation of bilirubin resulting from relative lack of glucuronyltransferase (conjugating enzyme) in the first few days after birth; and (3) exposure of breast-feeding infants to pregnanediol, nonesterified fatty acids, and other inhibitors of bilirubin conjugation present in breast milk.

Physiological jaundice of the newborn is treated with phototherapy; exposure to light of approximately 450 nm disrupts intramolecular hydrogen bonds in the bilirubin molecule and yields several photoisomers that are more water soluble than the Z-Z-isomer and thus are excreted in the bile. Exchange transfusions are rarely necessary.

Guidelines for managing newborns with hyperbilirubinemia were introduced in 2004. These guidelines were developed to

BOX 28.1 Physiologic Classification of Jaundice

Unconjugated Hyperbilirubinemia

Increased Production of Unconjugated Bilirubin From Heme

Hemolysis
- Hereditary
- Acquired

Ineffective erythropoiesis
Rapid turnover of increased red blood cell mass (in the neonate)

Decreased Delivery of Unconjugated Bilirubin (in Plasma) to Hepatocyte

Right-sided congestive heart failure
Portocaval shunt

Decreased Uptake of Unconjugated Bilirubin Across Hepatocyte Membrane

Competitive inhibition
- Drugs
- Others

Gilbert syndrome
Sepsis, fasting

Decreased Storage of Unconjugated Bilirubin in Cytosol (Decreased Y and Z Proteins)

Competitive inhibition
Fever

Decreased Biotransformation (Conjugation)

Neonatal jaundice (physiologic)
Inhibition (drugs)
Hereditary (Crigler-Najjar)
- Type I (complete enzyme deficiency)
- Type II (partial deficiency)

Hepatocellular dysfunction
Gilbert syndrome

Conjugated Hyperbilirubinemia (Cholestasis)

Decreased Secretion of Conjugated Bilirubin Into Canaliculi

Hepatocellular disease
- Hepatitis
- Cholestasis (intrahepatic)

Dubin-Johnson syndrome and Rotor syndrome
Drugs (estradiol)

Decreased Drainage

Extrahepatic obstruction
- Stones
- Carcinoma
- Stricture
- Atresia

Sclerosing cholangitis
Intrahepatic obstruction
- Drugs
- Granulomas
- Primary biliary cirrhosis
- Bile duct paucity
- Tumors

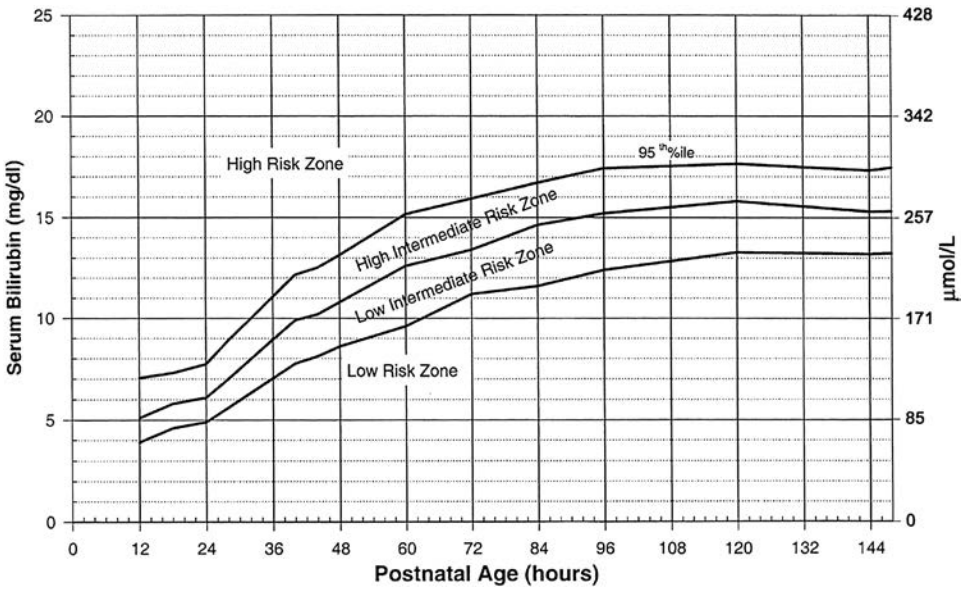


Fig. 28.14 Bhutani nomogram to assess risk of hyperbilirubinemia in newborns. (From Subcommittee on Hyperbilirubinemia. [2004]. Management of hyperbilirubinemia in the newborn infant 35 or more weeks of gestation. *Pediatrics, 114*, 297.)

assess the newborn's risk for hyperbilirubinemia prior to their discharge. The increase in bilirubin concentration from the time of birth prior to discharge from the hospital is monitored using a nomogram similar to the Bhutani nomogram (Fig. 28.14). The total bilirubin concentration is plotted versus the time (hours) to determine the newborns risk for hyperbilirubinemia.

Conjugated hyperbilirubinemias

Biliary atresia. **Biliary atresia** is a condition in which the bile ducts either fail to develop or develop abnormally. It occurs in one in 10,000 births, and females are more frequently affected. Having no exit to the intestine, bile accumulates inside the liver and eventually escapes into the blood, causing mixed hyperbilirubinemia. Other possible causes include (1) certain viral infections, (2) α_1-antitrypsin deficiency, and (3) trisomy 17 or 18.

Extrahepatic biliary atresia, which is more common than the intrahepatic type, may involve all or part of the extrahepatic biliary tree. If jaundice persists beyond 14 days of age, a direct or **conjugated bilirubin** measurement must be performed to exclude biliary atresia. If bilirubin is elevated, the urine should be tested for bile. Early identification of this condition is essential if these infants are to benefit from the operation of portoenterostomy, which should be performed no later than 60 days after birth. If portoenterostomy is not successful, liver transplantation is the treatment of choice. Children rarely live beyond 3 years unless the lesion is surgically correctable.

Analytical Methods

Serum bilirubin is measured in body fluids by spectrophotometric or chromatographic methods. For a description of additional methods and materials, the reader is referred to the 6th edition of the *Tietz Textbook of Clinical Chemistry and Molecular Diagnostics*.

Diazo Methods

The reaction of bilirubin with diazotized sulfanilic acid, known as the *diazo reaction,* is the basis of the most widely used methods for measuring bilirubin. The observation that the reaction was slow in sera from jaundiced infants and required an accelerator (ethanol) to proceed, and that it was rapid in bile and in adult sera with the addition of ethanol led to the terms **indirect bilirubin** and **direct bilirubin**, respectively.

The reaction of the coupling of bilirubin with a diazo compound is shown in Fig. 28.15. In this example, diazotized sulfanilic acid reacts with bilirubin to produce two azodipyrroles, which are reddish purple at neutral pH and blue at low or high pH values. Numerous variations of this method have been developed. All use one of a variety of "accelerators," which facilitate the reaction of unconjugated (indirect) bilirubin with the diazo reagent. The most commonly used accelerators are (1) caffeine, (2) dyphylline, and (3) several surface-active agents (surfactants).

Total Bilirubin

The diazo method to quantitate total bilirubin in serum was described by Jendrassik and Grof, and later modified by Doumas and colleagues.

In this procedure, serum is added to an aqueous solution of caffeine and sodium benzoate (accelerator), which displaces unconjugated bilirubin from its association sites on albumin. Formation of hydrogen bonds with caffeine

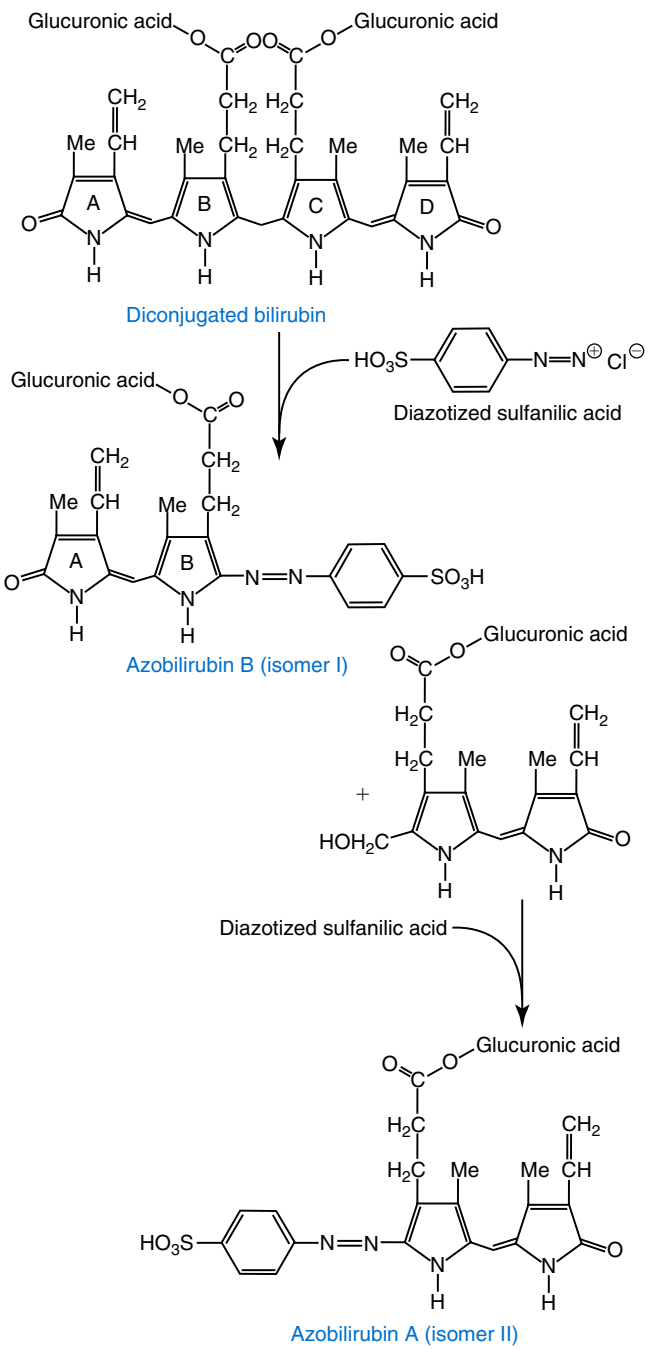


Fig. 28.15 The reaction of bilirubin glucuronide with diazotized sulfanilic acid to produce isomers I and II of azobilirubin B. Unconjugated bilirubin reacts in the same way to produce isomers I and II of azobilirubin A.

renders the bilirubin water soluble and facilitates its reaction with diazotized sulfanilic acid. After a 10-minute incubation period at room temperature, alkaline tartrate is added, and absorbance of the alkaline azobilirubin is measured at 598 nm. This method is currently the reference method for measuring total bilirubin in serum.

Direct bilirubin. Bilirubin monoconjugates and diconjugates (mainly glucuronides), and δ-bilirubin, because they are water soluble, react with the diazo reagents in the absence of accelerators. A reliable method for direct bilirubin should not

measure any unconjugated bilirubin. To prevent the unconjugated bilirubin from reacting, it is necessary to keep the pH of the reaction mixture near 1.0. A preferred manual method for direct bilirubin is available. Ditaurobilirubin (bilirubin conjugated with taurine and available as the disodium salt)—a water-soluble synthetic material—is used by instrument manufacturers for calibrating direct bilirubin methods; it is also present in quality control and proficiency-testing materials.

REVIEW QUESTIONS

1. Heme is a:
   a. Chelate of iron with the four pyrrole groups of a porphyrin.
   b. Conjugated protein and an oxygen-carrying pigment of the erythrocytes.
   c. Protein found in red skeletal muscle that releases oxygen.
   d. Colorless compound formed in the intestines by the reduction of bilirubin.
2. The role of hemoglobin is to:
   a. Transport iron between organs.
   b. Store iron and readily release it when body iron stores are low.
   c. Reversibly bind oxygen.
   d. Conjugate bilirubin in the liver.
3. What is bilirubin conjugated to in a hepatocyte to form conjugated bilirubin?
   a. Cholic acid
   b. Glycogen
   c. Bile acid
   d. Glucuronic acid
4. In an individual with suspected β-thalassemia, which of the following laboratory results would correctly indicate the presence of this disease?
   a. Increased hemoglobin concentration, mean corpuscular volume (MCV) and mean corpuscular hemoglobin concentration (MCHC) with the peripheral blood smear showing increased macrocytes and Howell–Jolly bodies
   b. Decreased hemoglobin concentration, MCV and MCHC with the peripheral blood smear indicating microcytosis, target cells, and polychromasia
   c. Decreased hemoglobin concentration, increased MCV and MCHC and persistence of hemoglobin F with the peripheral blood smear indicating spherocytosis and nucleated red blood cells
   d. Increased hemoglobin concentration and normal MCV and MCHC with a normal peripheral blood smear
5. Which of the following is associated with low serum iron and high total iron binding capacity (TIBC) or transferrin?
   a. Hemochromatosis
   b. Iron-deficiency anemia
   c. Iron intoxication
   d. Anemia of chronic disease
6. This readily soluble iron/protein complex is the form in which iron is stored in tissues:
   a. Ferritin
   b. Transferrin
   c. Hemosiderin
   d. Hemoglobin
7. The correct formula for determining the percent of transferrin saturation % is:
   a. MCV <70 mL = increased transferrin
   b. Total iron binding capacity ÷ serum iron
   c. Total iron binding capacity × 100
   d. [Serum iron: TIBC] × 100
8. The major difference between thalassemia and hemoglobinopathy is that:
   a. In thalassemia the globin chains of hemoglobin are structurally altered.
   b. In thalassemia the serum level of conjugated bilirubin is dramatically increased.
   c. In a hemoglobinopathy the globin chains of hemoglobin are structurally altered.
   d. In a hemoglobinopathy the globin chains of hemoglobin are insufficiently produced.
9. Anemia of chronic disease is characterized by:
   a. Elevated transferrin saturation (TSAT) and elevated or normal ferritin
   b. Elevated TSAT and low or normal ferritin
   c. Low TSAT and low or normal ferritin
   d. Low TSAT and elevated or normal ferritin
10. Symptomatic hereditary hemochromatosis is best characterized by:
    a. Elevated TSAT and elevated ferritin
    b. Normal TSAT and elevated ferritin
    c. Elevated TSAT and normal ferritin
    d. Normal TSAT and normal ferritin
11. Hemolytic disease in a newborn results from maternal-fetal Rh incompatibility. This produces physiological jaundice that is characterized by:
    a. Unconjugated hyperbilirubinemia.
    b. Conjugated bilirubinemia.
    c. Hemosiderosis.
    d. Hemoglobinopathy.
12. Extrahepatic biliary atresia:
    a. Is characterized by the lack of bile ducts within the liver.
    b. Is less common than intrahepatic biliary atresia.
    c. Is characterized by a mixed hyperbilirubinemia.
    d. Produces jaundice that can be observed within the first few days of life.
13. Bilirubin:
    a. In the conjugated form is highly toxic to the nervous system when elevated.
    b. In the plasma is bound to albumin before conjugation in hepatocytes.
    c. Becomes conjugated to cholic acid in the hepatocyte.
    d. All of the above.

SUGGESTED READINGS

Beguin, Y. (2003). Soluble transferrin receptor for the evaluation of erythropoiesis and iron status. *Clinica Chimica Acta*, *329*, 9–22.

Bhutani, V. K., Johnson, L., & Sivieri, E. M. (1999). Predictive ability of predischarge hour-specific serum bilirubin for subsequent significant hyperbilirubinemia in healthy term and near-term newborns. *Pediatrics*, *103*, 6–14.

Camaschella, C. (2015). Iron-deficiency anemia. *The New England Journal of Medicine*, *372*, 1832–1843.

Doumas, B. T., Poon, P. K. C., Perry, B. W., et al. (1985). Candidate reference method for determination of total bilirubin in serum: development and validation. *Clinical Chemistry*, *31*, 1779–1789.

Fleming, R. E., & Ponka, P. (2012). Iron overload in human disease. *The New England Journal of Medicine*, *366*, 348–359.

Girelli, D., Nemeth, E., & Swinkels, D. W. (2016). Hepcidin in the diagnostics of iron disorders. *Blood*, *127*, 2809–2813.

Musallam, K. M., Rivella, S., Vichinsky, E., et al. (2013). Non-transfusion dependent thalassemia. *Haematologica*, *98*, 833–844.

Nicolle, S., Jackson, N., Radi, K., et al. (2013). Evaluation of mass spectrometry based approaches for the diagnosis of hemoglobinopathies. *Blood*, *122*, 4679.

Piel, F. B., & Weatherall, D. J. (2014). The α-thalassemias. *The New England Journal of Medicine*, *371*, 1908–1916.

Schoorl, M., Schoorl, M., van Pelt, J., et al. (2015). Application of innovative hemocytometric parameters and algorithms for improvement of microcytic anemia discrimination. *Hematology Reports*, *7*, 5843.

Weiss, G., & Goodnough, L. T. (2005). Anemia of chronic disease. *The New England Journal of Medicine*, *352*, 1011–1023.

29

Porphyrins and Porphyrias

Michael N. Badminton, Sharon D. Whatley, Aasne K. Aarsand

OBJECTIVES

1. Define the following terms:
   - 5-Aminolevulinic
   - Heme
   - Porphyrin
   - Porphobilinogen (PBG)
   - Porphyria
   - Zinc protoporphyrin
2. Summarize the biosynthetic pathway of heme, including intracellular sites of each synthetic step, rate-limiting step, and enzyme, and incorporation into hemoglobin.
3. List six functions of heme.
4. Compare and contrast the acute and nonacute porphyrias, including causes, symptoms, and laboratory analyses; describe the screening tests that differentiate between the two categories of porphyria.
5. List and categorize seven porphyrias; state the major elevated intermediates involved in each.
6. List the abnormalities of porphyrin metabolism that are unrelated to porphyria.
7. Explain the effects of lead toxicity on the heme biosynthetic pathway and on iron status.
8. List the specimen collection and storage requirements for samples undergoing porphyrin analysis.
9. Describe the laboratory methods of analysis for the following:
   - Porphobilinogen
   - 5-Aminolevulinic acid
   - Porphyrins in urine, stool, blood, and plasma
   - DNA assessments
10. Analyze and solve case studies related to abnormal porphyrin metabolism and lead toxicity.

KEY WORDS AND DEFINITIONS

Acute intermittent porphyria (AIP) An autosomal dominant hepatic porphyria caused by mutation in the *HMBS* gene (locus: 11q23.3), which encodes hydroxymethylbilane synthase.

Acute porphyrias Inherited disorders of heme biosynthesis, characterized by acute attacks of neurovisceral symptoms; potentially life threatening; diagnosed by elevated urine porphobilinogen (PBG).

5-Aminolevulinic acid (ALA) Immediate precursor of porphobilinogen; two molecules of ALA combine to form one molecule of porphobilinogen.

Congenital erythropoietic porphyria (CEP) An autosomal recessive porphyria due to mutation in the *UROS* gene (locus: 10q25.2–q26.3), which encodes uroporphyrinogen-III synthase.

Coproporphyrin A porphyrin with four methyl and four propionic acid side chains attached to the tetrapyrrole backbone.

Cutaneous porphyrias Disorders of heme biosynthesis in which accumulations of porphyrins in the skin cause skin damage on exposure to sunlight.

Erythropoietic protoporphyria (EPP) An autosomal recessive disorder due to mutation in the *FECH* gene (locus: 18q21.3), which encodes ferrochelatase, causing a partial deficiency of the enzyme.

Ferrochelatase (FECH) A mitochondrial enzyme of the lyase class that catalyzes the insertion of ferrous iron into protoporphyrin IX to form heme.

Hereditary coproporphyria (HCP) An autosomal dominant hepatic porphyria caused by mutations of the *CPOX* gene (locus: 3q12) that result in partial deficiency of coproporphyrinogen oxidase activity.

Porphobilinogen (PBG) Immediate precursor of the porphyrins; a pyrrole ring with acetyl, propionyl, vinyl, and aminomethyl side chains; four molecules of PBG condense to form one molecule of 1-hydroxymethylbilane, which is then converted successively to uroporphyrinogen-III, coproporphyrinogen-III, protoporphyrinogen-IX, protoporphyrin-IX, and heme.

Porphyrins Any of a group of compounds containing the porphyrin structure; four pyrrole rings connected by methylene bridges in a cyclical configuration, to which a variety of side chains are attached.

Porphyria cutanea tarda (PCT) The most common form of porphyria, characterized by cutaneous photosensitivity.

Porphyrias A group of mainly inherited metabolic disorders that result from partial deficiencies of the enzymes of heme biosynthesis, which cause increased formation and excretion of porphyrins, their precursors, or both.

Porphyrin precursors ALA and PBG, the biosynthetic intermediates, are metabolized to porphyrinogens and porphyrins.

Protoporphyrin A porphyrin with four methyl, two vinyl, and two propionic acid side chains attached to the tetrapyrrole backbone.

Upregulation A process by which the cell increases the quantity of a cellular component such as a specific protein or RNA.

Uroporphyrin A porphyrin with four acetic acid and four propionic acid side chains attached to the tetrapyrrole backbone.

Variegate porphyria (VP) An autosomal dominant hepatic porphyria due to mutation in the *PPOX* gene (locus: 1q22), which encodes protoporphyrinogen oxidase.

X-linked erythropoietic protoporphyria An erythropoietic porphyria caused by mutations in the ALAS2 gene (locus Xp11.21) that result in increased ALA synthase activity.

Zinc protoporphyrin (ZPP) A normal but minor by-product of heme biosynthesis found in the red blood cell; when insufficient Fe^{2+} is available for heme biosynthesis, increased ZPP is formed.

The **porphyrias** are a group of uncommon, inherited diseases resulting from either partial deficiency in one of the enzymes of heme biosynthesis or, in one disorder, increased activity of the rate-controlling enzyme of erythroid heme synthesis. Each disorder is associated with a specific pattern of overproduction of pathway intermediates, which are excreted in excessive amounts in urine, feces, or both. The clinical consequences depend on the nature of the heme precursors that accumulate. In the **acute porphyrias**, excess **porphyrin precursors** (5-aminolaevulinic acid [ALA] and **porphobilinogen [PBG]**) are associated with potentially fatal acute neurovisceral attacks that are often provoked by various commonly prescribed drugs or hormonal factors. In the nonacute porphyrias, and in those acute porphyrias in which skin lesions occur, accumulation of **porphyrins** results in photosensitization of sun-exposed skin. Diagnosis depends on laboratory investigations to demonstrate the pattern of heme precursor accumulation and excretion specific for each type of porphyria through examination of appropriate specimens using adequately sensitive and specific methods. DNA analysis is rarely necessary for diagnosis of symptomatic cases, but is the method of choice when investigating healthy at-risk relatives. Abnormalities of porphyrin accumulation and excretion also occur in a wide variety of other disorders that are collectively more common than the porphyrias. Recognition of secondary porphyrin disorders is important to avoid diagnostic errors.

PORPHYRIN CHEMISTRY

In this section, porphyrin structure, nomenclature, and chemical characteristics are reviewed.

Structure and Nomenclature

The basic porphyrin structure consists of four monopyrrole rings connected by methene bridges to form a tetrapyrrole ring (Fig. 29.1). The porphyrin compounds of relevance to the porphyrias differ in the substituents occupying peripheral positions 1 through 8 (Fig. 29.2). Variation in the distribution of the same substituents around the peripheral positions of the tetrapyrrole ring gives rise to porphyrin isomers, which are depicted by Roman numerals (e.g., I, II, III). The reduced form of a porphyrin, known as a *porphyrinogen* (see Fig. 29.1), differs by the absence of six hydrogens (four from the methylene bridges and two from ring nitrogens). Porphyrinogens are unstable in vitro and spontaneously oxidize to the corresponding porphyrins. Under the lower cellular oxygen tension, porphyrinogens are sufficiently stable to act as intermediates of the heme biosynthetic pathway; aromatization to **protoporphyrin** at the penultimate step requires protoporphyrinogen oxidase.

Chelation of Metals

The arrangement of four nitrogen atoms in the center of the porphyrin ring enables porphyrins to chelate various metal ions. Protoporphyrin that contains iron is known as heme; ferroheme refers specifically to the Fe^{2+} complex and ferriheme to Fe^{3+}. Ferriheme associated with a chloride counter ion is known as hemin, or hematin when the counter ion is hydroxide.

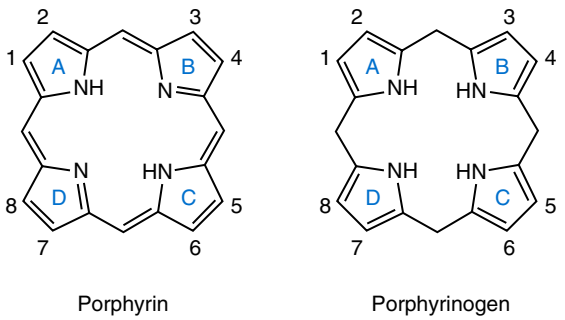


Fig. 29.1 Porphyrin and porphyrinogen structures: numbers 1 to 8 represent various substituents, the nature and order of which determine the type of porphyrin or porphyrinogen. Numbering system and ring designations are based on the Fischer system.

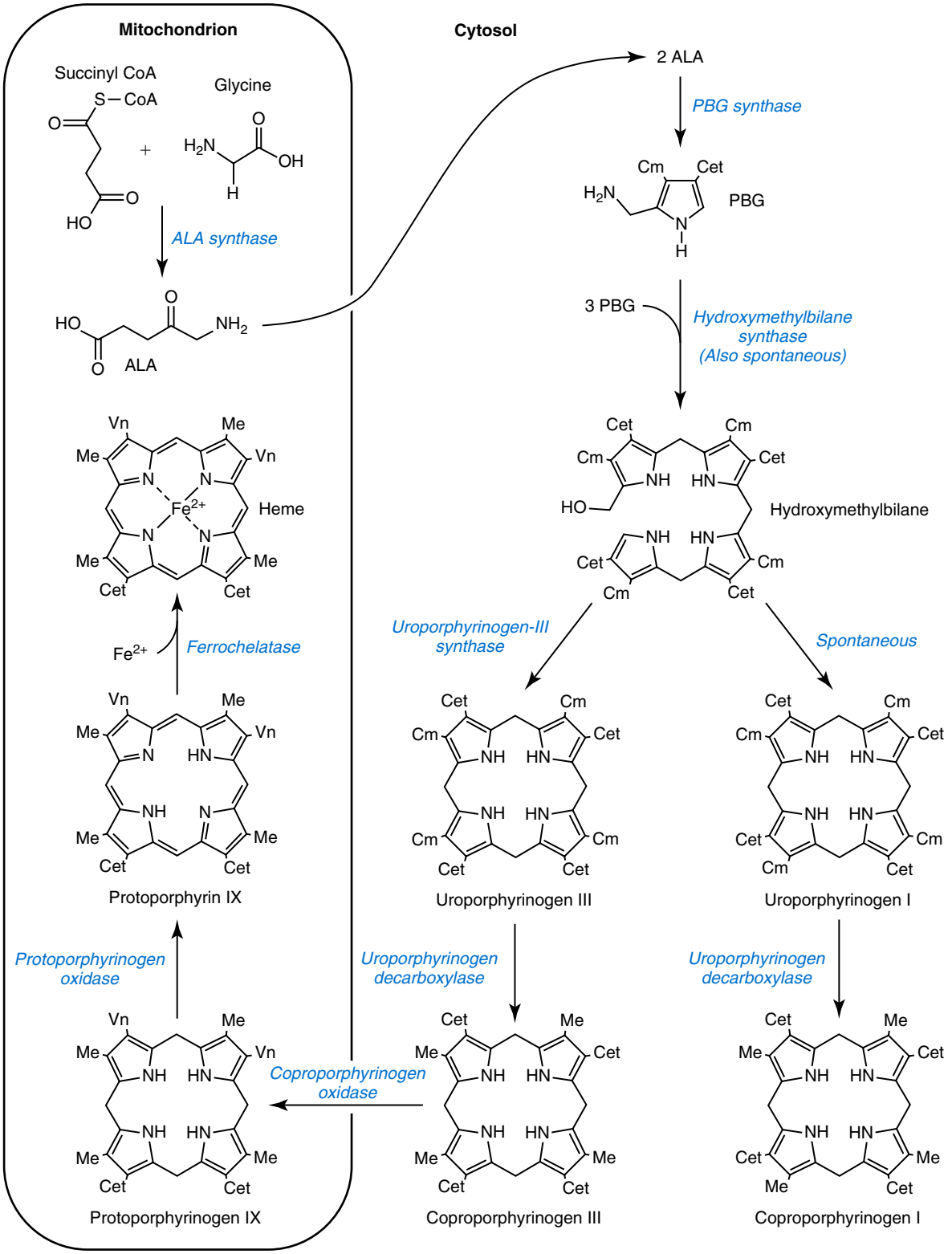


Fig. 29.2 Biosynthetic pathway of porphyrins and heme. C_{et},—CH_2CH_2COOH; C_m, —CH_2COOH; *Me*, —CH_3; *Vn*, —$CH{=}CH_2$. *ALA*, 5-Aminolevulinic acid; *CoA*, coenzyme A; *PBG*, porphobilinogen.

Spectral Properties

Porphyrins, named from the Greek root for "purple" ("porphyra"), owe their color to the conjugated double-bond structure of the tetrapyrrole ring. Porphyrinogens have no conjugated double bonds and are therefore colorless. Porphyrins show particularly strong absorbance near 400 nm, often called the Soret band, and as a result of this exposure, display a characteristic orange–red fluorescence in the range of 550 to 650 nm. Absorbance and fluorescence are altered by substituents around the porphyrin ring and by metal binding. Zinc chelation shifts the fluorescence emission peak of protoporphyrin to shorter wavelengths and reduces the fluorescence intensity. The strong binding of iron alters the character of protoporphyrin to the extent that heme lacks significant fluorescence.

HEME BIOSYNTHESIS

The complex tetrapyrrole ring structure of heme is built up in a stepwise fashion from precursors succinyl–coenzyme A (CoA) and glycine (see Fig. 29.2). The pathway is present in all nucleated cells and it has been estimated that daily synthesis of heme in humans is 5 to 8 mmol/kg body weight. The pathway is compartmentalized, with some steps occurring in the mitochondrion and others in the cytoplasm.

Enzymes of Heme Biosynthesis

The genes for all enzymes of human heme biosynthesis have been characterized (Table 29.1) and enzyme structures determined by x-ray crystallography.

5-Aminolevulinate Synthase (EC 2.3.1.37)

5-Aminolevulinate synthase (ALAS), the initial enzyme of the pathway is mitochondrial and has a housekeeping (ALAS1) and erythroid (ALAS2) isoenzyme. It catalyzes the formation of ALA (sometimes referred to as aminolevulinate to emphasize its ionic nature at physiological pH) from succinyl-CoA and glycine and requires a pyridoxal phosphate cofactor, which forms a Schiff base with the amino group of glycine at the enzyme surface. The carbanion of the Schiff base displaces CoA from succinyl-CoA with the formation of α-amino-β-ketoadipic acid, which is then decarboxylated to ALA. The activity of ALAS is rate limiting as long as the catalytic capacities of other enzymes in the pathway are normal.

5-Aminolevulinic Acid Dehydratase (EC 4.2.1.24)

Aminolevulinic acid dehydratase (ALAD) (also known as PBG synthase) is a cytoplasmic enzyme that catalyzes the formation of the monopyrrole PBG from two molecules of ALA with the elimination of two molecules of water. The enzyme requires zinc as a cofactor and reduced sulfhydryl groups at the active site, and is therefore susceptible to inhibition by lead.

Hydroxymethylbilane Synthase (EC 2.5.1.61)

Hydroxymethylbilane synthase (HMBS) (also known as PBG deaminase) is a cytoplasmic enzyme that catalyzes the formation of one molecule of the linear tetrapyrrole 1-hydroxymethylbilane (HMB; also known as preuroporphyrinogen) from four molecules of PBG with the release of four molecules of ammonia. The enzyme is susceptible to allosteric inhibition by intermediates farther down the heme biosynthetic pathway, notably coproporphyrinogen-III and protoporphyrinogen-IX.

Uroporphyrinogen-III Synthase (EC 4.2.1.75)

Uroporphyrinogen synthase (UROS) is a cytoplasmic enzyme that rearranges and cyclizes HMB to form uroporphyrinogen-III. Each pyrrole ring of HMB contains a methylcarboxylate and an ethylcarboxylate substituent, which are in the same orientation. By the rotation of zero, one, or two alternate or two adjacent pyrrole rings, it is possible to arrive at four different isomers (I to IV). Apart from closing the ring structure, the enzyme rotates the D-ring via a spirane intermediate, producing type III, the only isomer which can contribute to

TABLE 29.1 Human Enzymes and Genes of Heme Biosynthesis

Enzyme	Monomer Mol Mass (kDa)[ab]	Chromosomal Location of Gene	Gene Size (kb)	Expression
ALAS1	70.6	3p21.2	17	Ubiquitous
ALAS2	64.6	Xp11.21	22	Erythroid cells
ALAD	36.3	9q32	13	Ubiquitous and erythroid-specific mRNAs
HMBS	37.0	11q23.3	10	Ubiquitous and erythroid-specific isoforms
UROS	29.5	10q26.1–q26.2	34	Ubiquitous and erythroid-specific mRNAs
UROD	40.8	1p34.1	3	Ubiquitous
CPOX	40.3	3q11.2–q12.1	14	Ubiquitous
PPOX	50.8	1q23.3	5	Ubiquitous
FECH	47.8	18q21.3	45	Ubiquitous

[a]ALAD is a homo-octamer, and HMBS and UROS are monomers; all other enzymes are homodimers.
[b]Molecular masses for ALAS1, ALAS2, CPOX, and FECH include presequences that are cleaved during mitochondrial import.
ALAD, Aminolevulinic acid dehydratase; *ALAS*, aminolevulinate synthase; *CPOX*, coproporphyrinogen oxidase; *FECH*, ferrochelatase; *HMBS*, hydroxymethylbilane synthase; *mRNA*, messenger ribonucleic acid; *PPOX*, protoporphyrinogen oxidase; *UROD*, uroporphyrinogen decarboxylase; *UROS*, uroporphyrinogen.

heme biosynthesis. HMB is unstable, and in those porphyrias in which excess HMB accumulates, cyclization occurs nonenzymatically forming the type I isomer. Normally, only minimum amounts of uroporphyrinogen-I are produced.

Uroporphyrinogen Decarboxylase (EC 4.1.1.37)

The last cytoplasmic enzyme in the pathway catalyzes the decarboxylation of all four carboxymethyl groups to form the tetracarboxylic coproporphyrinogen. The enzyme can use I and III isomers of uroporphyrinogen as substrate. Decarboxylation commences on ring D and proceeds stepwise through rings A, B, and C with formation of heptacarboxylate, hexacarboxylate, and pentacarboxylate intermediates at a single active site. Decreased uroporphyrinogen decarboxylase (UROD) activity causes accumulation of these intermediates in addition to its substrate, uroporphyrinogen. At high substrate concentrations, decarboxylation occurs by a random mechanism.

Coproporphyrinogen Oxidase (EC 1.3.3.3)

Coproporphyrinogen oxidase (CPOX), which is located in the mitochondrial intermembrane space, catalyzes the sequential oxidative decarboxylation of the 2- and 4-carboxyethyl groups to produce the more lipophilic protoporphyrinogen-IX, with formation of a tricarboxylic intermediate, harderoporphyrinogen. Oxygen is required as the oxidant and the enzyme requires sulfhydryl groups for activity, making it a target for inhibition by metals. The enzyme is specific for the type III isomer only. The product of the enzyme differs from the substrate in that replacement of two of the carboxyethyl groups by vinyl groups introduces a third substituent into the molecule. The number of possible isomeric forms is therefore increased, and conventionally the numbering system changes, so that the III isomer becomes the IX isomer. In UROD-deficient states, one of the ethylcarboxylate groups of the accumulated pentacarboxylate porphyrinogen is decarboxylated by CPOX to form the isocoproporphyrin series of porphyrins.

Protoporphyrinogen Oxidase (EC 1.3.3.4)

Protoporphyrinogen oxidase (PPOX), a flavoprotein located in the inner mitochondrial membrane, catalyzes the removal of six hydrogens (four from methylene bridges and two from ring nitrogens) to form protoporphyrin-IX. Although nonenzymatic oxidation also occurs in vitro, under the low cellular oxygen tension, PPOX is essential for oxidation to occur. The protoporphyrin produced is the only porphyrin that functions in the heme pathway. Other porphyrins are produced by nonenzymatic oxidation and represent porphyrinogens that have irreversibly escaped from the pathway.

Ferrochelatase (EC 4.99.1.1)

Ferrochelatase (FECH) (also known as heme synthase) is an iron-sulfur protein located in the inner mitochondrial membrane. This enzyme inserts ferrous iron into protoporphyrin to form heme. During this process, two hydrogens are displaced from the ring nitrogens. Other metals in the divalent state also act as substrates, yielding the corresponding chelate (e.g., incorporation of Zn^{2+} into protoporphyrin to yield **zinc protoporphyrin [ZPP]**). In iron-deficient states, Zn^{2+} successfully competes with Fe^{2+} in developing red cells, so that the concentration of ZPP in erythrocytes increases. In lead poisoning, impaired iron delivery or utilization within the mitochondrion has the similar effect of increasing erythrocyte ZPP. Other dicarboxylic porphyrins also serve as substrates (e.g., mesoporphyrin).

REGULATION OF HEME BIOSYNTHESIS

Heme supply in all tissues is controlled by the activity of mitochondrial ALAS. Two isoforms of ALAS are known. The ubiquitous isoform, ALAS1, is encoded by a gene on chromosome 3p21 and is expressed in all tissues. Because it has a half-life of approximately 1 hour, changes in its rate of synthesis produce short-term alterations in enzyme concentration and cellular ALAS activity. Synthesis of ALAS1 is under negative feedback control by heme. In the liver, ALAS1 is induced by a wide variety of drugs and chemicals that induce microsomal cytochrome P450–dependent oxidases (CYPs). This effect is thought to be mediated mainly by direct transcriptional activation of drug-responsive nuclear receptors, rather than occurring secondary to depletion of an intracellular regulatory heme pool. The induction of ALAS1 is prevented by heme, which acts by destabilizing messenger ribonucleic acid (mRNA) for ALAS1, by blocking mitochondrial import of pre-ALAS1, by increased proteolysis and possibly by inhibiting transcription. In addition, ALAS1 activity is regulated by a transcriptional co-activator, PGC-1α; an effect that links the rate of hepatic heme synthesis with nutritional status.

The erythroid isoform, ALAS2, is encoded by a gene on chromosome Xp11.21 and is expressed only in erythroid cells. Its activity is regulated by two distinct mechanisms. Transcription is enhanced during erythroid differentiation by the action of erythroid-specific transcription factors, and mRNA concentrations are regulated by iron. Iron deficiency in erythroid cells promotes specific binding of iron regulatory proteins to an iron-responsive element in the 5′ untranslated region (UTR) of ALAS2 mRNA with consequent inhibition of translation.

Function of Heme

Heme functions as a prosthetic group in various proteins in which, depending on the function of the protein, the iron shifts freely between the 2^+ and 3^+ valency states. Heme synthesis (70% to 80%) occurs in the bone marrow and approximately a further 15% in the liver. Heme-containing proteins participate in a variety of redox reactions, including:

1. Oxygen transport (by hemoglobin in the blood) and storage (by myoglobin in muscle)
2. Mitochondrial respiration (by cytochromes b_1, c_1, and a_3)
3. Enzymatic destruction of peroxides (by catalase and peroxidase)
4. Drug metabolism (by microsomal cytochrome P-450 mixed function oxidases)

TABLE 29.2 Adult Reference Intervals

Specimen	Analyte	Reference Interval (SI Units)	Reference Interval (Traditional Units)
Urine	Porphobilinogen	<1.5 μmol/mmol creatinine	<3.0 mg/g creatinine
		<10 μmol/L	<0.23 mg/dL
	5-aminolevulinic acid	<3.8 μmol/mmol creatinine	<4.4 mg/g creatinine
		<50 μmol/L	< 0.66 mg/dL
	Total porphyrin	<35 nmol/mmol creatinine	<216 μg/g creatinine
		20–320 nmol/L	14–224 μg/L
	Uroporphyrin	0.8–3.1 nmol/mmol creatinine	5.9–22.8 μg/g creatinine
	Heptacarboxylate porphyrin	<0.9 nmol/mmol creatinine	<6.3 μg/g creatinine
	Coproporphyrin-I	1.2–5.7 nmol/mmol creatinine	6.9–33.0 μg/g creatinine
	Coproporphyrin-III	4.8–23.8 nmol/mmol creatinine	27.8–137.7 μg/g creatinine
	% Coproporphyrin-III[a]	68–86	68–86
Feces	Total porphyrin	10–200 nmol/g dry wt	6–117 μg/g dry wt
	Coproporphyrin-I	1.1–5.5 nmol/g feces	0.7–3.6 μg/g feces
	Coproporphyrin-III	0.2–2.5 nmol/g feces	0.1–1.6 μg/g feces
	Coproporphyrin-III/I ratio	0.3–1.4	0.3–1.4
	Total dicarboxylate porphyrin	0.5–12.8 nmol/g feces	0.3–7.2 μg/g feces
Erythrocytes	Total porphyrin	0.4–1.7 μmol/L erythrocytes	25–106 μg/dL erythrocytes

[a]Percentage of total coproporphyrin.
Laboratories should verify that these ranges are appropriate for use. Further guidance is provided in Chapter 5.
wt, Weight.

5. Desaturation of fatty acids (by microsomal cytochrome b_5)
6. Tryptophan metabolism (by tryptophan oxygenase)

Reactions of nitric oxide (NO) are often mediated by the reaction of heme with NO in control enzymes such as guanylate cyclase.

Other naturally occurring tetrapyrrole derivatives include vitamin B_{12} and chlorophyll, each of which contains an atom of chelated cobalt and magnesium, respectively.

SOLUBILITY AND EXCRETION OF HEME PRECURSORS

Typically, only minute quantities of heme precursors accumulate in the body. The route of excretion largely depends on solubility. The porphyrin precursors ALA and PBG are water soluble and are excreted almost exclusively in urine as is uroporphyrinogen. PBG polymerizes readily, particularly at high concentrations in acid solution to form primarily the I-isomer of uroporphyrin. The last intermediate of the pathway, protoporphyrin (and also protoporphyrinogen), is insoluble in water and is excreted in the feces via the biliary tract. The other porphyrins are of intermediate solubility and appear in both urine and feces. Coproporphyrinogen-I is taken up and excreted by the liver in preference to the III isomer, so that coproporphyrinogen-I predominates in feces and coproporphyrinogen-III in urine. All porphyrinogens in the urine or feces are slowly oxidized to the corresponding porphyrins.

Once in the gut, porphyrins are susceptible to modification by gut flora. The two vinyl groups of protoporphyrin are reduced to ethyl groups, hydrated to hydroxyethyl groups, or removed, giving rise to a variety of secondary porphyrins. Gut flora can also metabolize heme (from dietary, gut cell lining or gastrointestinal bleeding) to produce a variety of dicarboxylic porphyrins. In addition, some bacteria are capable of de novo synthesis of porphyrins.

The differing solubility of individual porphyrins is of importance not only in determining the route of excretion from the body, but also in the design of analytical methods for their extraction and fractionation. At pH 7, the carboxyl groups are ionized, and the molecule has a net negative charge. Below pH 2, the pyrrole nitrogens and the carboxyl groups become protonated so that the molecule has a net positive charge. At physiologic pH, the solubility of a given porphyrin is determined by the number of substituent carboxyl groups. **Uroporphyrin** has eight carboxylate groups and is the most soluble porphyrin in aqueous media. Protoporphyrin has only two carboxylate groups, is insoluble in water, but dissolves readily in lipid environments, and binds readily to the hydrophobic regions of proteins such as albumin. **Coproporphyrin**, with four carboxylate groups, has intermediate solubility. Reference ranges for porphyrins and porphyrin precursors in blood, urine and feces are reported in Table 29.2

THE PORPHYRIAS

The porphyrias are a group of metabolic disorders associated with a disturbance of heme biosynthesis (Table 29.3). All are inherited in monogenic patterns, apart from some forms of **porphyria cutanea tarda (PCT)** and rare erythropoietic porphyrias associated with malignant myeloid disorders. Each type of porphyria is defined by the association of characteristic clinical features with a specific pattern of accumulation of heme precursors that reflects increased formation of the substrate of the enzyme that is partially deficient or becomes secondarily rate-limiting in that type of porphyria (Table 29.4). The porphyrias are characterized clinically by two main features; skin

TABLE 29.3 Main Types of Human Porphyria

Disorder	Defective Enzyme	Prevalence[a] (Per Million)	Neurovisceral Crises	Skin Lesions	Inheritance
Acute Porphyrias					
ADP	ALAD	–	+	–	AR
AIP	HMBS	5.9	+	–	AD
HCP	CPOX	0.9	+	+[b,c]	AD
VP	PPOX	3.2	+	+[b,c]	AD
Nonacute Porphyrias					
CEP	UROS	0.3	–	+[c]	AR
PCT	UROD	40	–	+[c]	Complex (20% AD)
EPP	FECH	9.2	–	+[d]	AR
XLEPP	ALAS2	0.15	–	+[d]	XL

[a]Estimated prevalence of clinically overt disease.
[b]Skin lesions and neurovisceral crises may occur alone or together.
[c]Fragile skin, bullae.
[d]Acute photosensitivity without fragile skin, bullae.
AD, Autosomal dominant; *ADP*, ALA dehydratase deficiency porphyria; *AIP*; acute intermittent porphyria; *ALAD*, aminolevulinic acid dehydratase; *ALAS*, Aminolevulinate synthase; *AR,* autosomal recessive; *CEP*, congenital erythropoietic porphyria; *CPOX*, coproporphyrinogen oxidase; *EPP*, erythropoietic protoporphyria; *FECH*, ferrochelatase; *HCP*, hereditary coproporphyria; *HMBS*, hydroxymethylbilane synthase; *PCT*, porphyria cutanea tarda; *PPOX*, protoporphyrinogen oxidase; *UROD*, uroporphyrinogen decarboxylase; *UROS*, uroporphyrinogen; *VP,* variegate porphyria; *XL,* X-linked; *XLEPP*, X-linked erythropoietic protoporphyria.

TABLE 29.4 Porphyrias: Patterns of Overproduction of Heme Precursors During Clinically Overt Phase of Disease

Porphyria	Urine PBG/ALA	Urine Porphyrins	Fecal Porphyrins	Erythrocyte Porphyrins	Plasma Fluorescence Emission Peak
ADP	ALA	Copro-III	Not increased	ZPP	–
AIP	PBG>ALA	Mainly uroporphyrin from PBG	Normal or increased[a] Copro-III/I ratio normal	Not increased	615–622 nm[b]
CEP	Not increased	Uro-I, Copro-I	Copro-I	ZPP, Proto, Copro-I, Uro-I	615–620 nm
PCT	Not increased	Uro, Hepta[c]	Isocopro, Hepta, Penta	Not increased	615–622 nm
HCP	PBG>ALA[d]	Copro-III, uroporphyrin from PBG	Copro-III, Copro-III/I ratio increased	Not increased	615–622 nm[b]
VP	PBG>ALA[d]	Copro-III, uroporphyrin from PBG	Proto IX>Copro-III[e] X-porphyrin Copro-III/I ratio increased	Not increased	624–628 nm
EPP	Not increased	Not increased	±Proto[f]	Proto	626–634 nm[g]
XLEPP	Not increased	Not increased	±Proto	Proto, ZPP[h]	626–634 nm[g]

[a]Slight increase only unless uroporphyrin is present.
[b]Not always increased.
[c]Other methylcarboxylate-substituted porphyrins are increased to a smaller extent; uroporphyrin is a mixture of type I and III isomers; heptacarboxylate porphyrin is mainly type III.
[d]PBG and ALA may be normal when only skin lesions are present.
[e]Coproporphyrin-III/I ratio increased, but usually less than in overt HCP.
[f]Not increased in about 40% of patients.
[g]Protoporphyrin (Proto) bound to globin. If hemolysis is present the peak shifts to the left (i.e., at 626–628 nm or is quenched).
[h]Zn-protoporphyrin (ZPP) 20%–60% of total erythrocyte protoporphyrin.
ADP, ALA dehydratase deficiency porphyria; *AIP*; acute intermittent porphyria; *ALA*, aminolevulinic acid; *CEP*, congenital erythropoietic porphyria; *EPP*, erythropoietic protoporphyria; *HCP*, hereditary coproporphyria; *PBG*, porphobilinogen; *PCT*, porphyria cutanea tarda; *VP,* variegate porphyria; *XLEPP*, X-linked erythropoietic protoporphyria.

lesions in sun-exposed areas and acute neurovisceral attacks, typically comprising of (1) abdominal pain, (2) peripheral neuropathy, and (3) mental disturbance. The skin lesions are caused by porphyrin-catalyzed photodamage, of which singlet oxygen is the main mediator. Acute attacks are associated with increased formation of ALA and consequently PBG from induced activity of hepatic ALAS1 and partial hepatic heme deficiency, often in response to induction of hepatic CYPs by drugs and other factors, such as hormones. The relationship of these biochemical changes to the neuronal dysfunction that

underlines all clinical features of the acute attack is uncertain. The leading hypothesis for pathophysiology is that ALA, or a product of ALA or PBG, is neurotoxic, possibly through interference with the gamma amino butyric acid (GABA) receptor pathway. As listed in Table 29.3, the porphyrias are classified as acute (in which acute neurovisceral attacks occur), or nonacute.

Acute Porphyrias

Acute porphyrias include (1) ALA dehydratase deficiency porphyria (ADP), (2) **acute intermittent porphyria (AIP)**, (3) **variegate porphyria (VP)**, and (4) **hereditary coproporphyria (HCP)**. These disorders are autosomal dominant except for the very rare disorder, ADP, which is autosomal recessive.

The inherited defect in each of the autosomal dominant acute porphyrias (see Table 29.3) is a mutation leading to complete or near complete inactivation of one of the pairs of allelic genes that encode the enzyme whose partial deficiency causes the disorder. Heme supply is maintained at normal or near normal concentration by **upregulation** of ALAS1, with a consequent increase in the substrate concentration for the defective enzyme. These compensatory changes vary between tissues; they are most prominent in the liver and are undetectable in most other organs. They also vary between individuals with some showing no evidence of overproduction of heme precursors, and others having biochemically manifest disease with or without clinical symptoms.

Low clinical penetrance is a prominent feature of all the autosomal dominant acute porphyrias. Family studies indicate that many affected individuals are asymptomatic throughout life.

All the autosomal dominant acute porphyrias show extensive allelic heterogeneity but founder mutations are present in some populations and explain the high frequency of VP in South Africans of Dutch descent and of AIP in Sweden.

Clinical Features

The life-threatening, acute neurovisceral attacks that occur in AIP, VP, and HCP are clinically identical. Acute attacks are very rare before puberty, and usually occur first between the ages of 15 and 40, and are more common in women. Precipitating factors have been identified in about two-thirds of patients who present with acute attacks. The most important are drugs; alcohol, especially binge drinking; the menstrual cycle; calorie restriction; infection; and stress. Drugs known to provoke acute attacks include barbiturates, sulfonamides, progestogens, and some anticonvulsants, but many others have been implicated in the precipitation of acute attacks (www.porphyria-europe.org).

The main clinical features are summarized in Table 29.5. Acute attacks almost always start with abdominal pain that rapidly becomes very severe and may also be present in the back and thighs. Vomiting, constipation, tachycardia, and hypertension, are frequent with convulsions as a consequence of hyponatremia or secondary to central nervous system involvement. Pain may dissipate within a few days, but in severe cases a predominant motor neuropathy develops that may progress to flaccid quadriparesis. The acute phase may be accompanied by mental confusion with abrupt changes in mood, hallucinations, and other psychotic features. Persistent psychiatric illness is not a feature of the acute porphyrias, although mild anxiety or depression may be present in some patients. Abdominal pain usually resolves within 2 weeks, but recovery from neuropathy may take many months and is not always complete. Most patients have one or a few attacks followed by complete recovery and prolonged remission. About 5% have repeated acute attacks, which, in women, may be premenstrual.

TABLE 29.5 Clinical Features of an Acute Neurovisceral Attack of Porphyria

Symptom/Sign	Percentage of Acute Attacks
Abdominal pain	85–90
Nonabdominal pain	25–70
Vomiting and nausea	30–90
Constipation/diarrhea	50–80
Psychologic symptoms (insomnia, anxiety, depression, confusion, hallucinations)	20–85
Acute encephalopathy (headache, somnolence, seizures, altered consciousness)	2–20
Motor neuropathy (muscle weakness, pain, low/absent tendon reflexes)	10–90
Hemi/tetraparesis	30–40
Respiratory paralysis	10–55
Sensory neuropathy	10–40
Hypertension	40–75
Tachycardia (>80/min)	30–85
Hyponatremia (<135 nmol/L)	30–60

Data modified from Harper, P., & Sardh, E. (2014). Management of acute intermittent porphyria. *Expert Opinion on Orphan Drugs, 2*(4), 349–368.

Skin lesions similar to those of PCT and other bullous porphyrias are present in about 80% of patients with clinically manifest VP (see Table 29.3). About 60% of patients with this condition present with skin lesions alone. The skin is less commonly affected in HCP; skin lesions without an acute attack are uncommon and usually are provoked by intercurrent cholestasis.

Long-term complications of acute porphyria include chronic renal failure, hypertension, and primary hepatocellular carcinoma.

Management of Families

The diagnosis of autosomal dominant acute porphyria should be followed by an investigation of the patient's family to identify affected, often asymptomatic relatives, so they can be advised to avoid drugs and other factors known to provoke potentially fatal acute attacks.

POINTS TO REMEMBER

Acute porphyrias

- Present clinically after puberty.
- Low clinical penetrance means that there is usually no family history.
- Clinical effects are due to autonomic, central, and motor neuropathy.
- Common precipitants are prescribed drugs, hormonal fluctuations, excess alcohol consumption, starvation, infection, and stress.
- Diagnosed by demonstrating increased PBG excretion in a random urine sample.
- Mild hyponatremia is common.

Nonacute Porphyrias

The nonacute porphyrias fall into two categories depending on whether patients have bullous skin lesions (PCT, **congenital erythropoietic porphyria [CEP]**) or acute photosensitivity (**erythropoietic protoporphyria [EPP]**, **X-linked erythropoietic protoporphyria [XLEPP]**).

Porphyria Cutanea Tarda

PCT is by far the most common porphyria. The disease occurs at all ages in both sexes with onset usually during the fifth and sixth decades. Lesions on sun-exposed skin, particularly the backs of the hands, the forearm, and the face, are present in all patients. These lesions are identical to those seen in the other bullous porphyrias (see Table 29.3). Increased mechanical fragility of the skin, with trivial trauma leading to erosions, is present in virtually all patients.

A combination of skin lesions with liver damage is strongly associated with alcohol abuse, estrogens, infection with hepatotropic viruses, particularly hepatitis C (HCV), smoking, and mutations in the hemochromatosis *(HFE)* gene. PCT may also complicate HIV infection. Hepatic iron overload in combination with at least one of the other associated factors are present in almost all patients.

PCT results from decreased activity of UROD in the liver, which is understood to be secondary to a UROD inhibitor, porphomethene, generated by iron dependent partial oxidation of uroporphyrinogen. This inhibition leads to the overproduction of uroporphyrin and other carboxymethyl-substituted porphyrins, which accumulate in the liver and skin, where they act as photosensitizers. This iron dependence explains why patients with iron overload disorders, such as haemochromatosis, develop PCT. Two main types of PCT can be identified in most populations; about 80% of patients have the sporadic (type I) form of PCT, in which the enzyme defect is restricted to the liver and is a result of viral illness (e.g., HCV) or exogenous substances (e.g., hepatic iron overload, alcohol excess). The *UROD* gene appears to be normal. The rest have familial (type II) PCT, which is inherited in an autosomal dominant manner. In this form, the mutation of one *UROD* gene leads to half-normal UROD activity in all tissues. However, the hepatic enzyme activity has to be 20% or less of normal for porphyrins to accumulate and cutaneous symptoms to appear. This further reduction in activity is dependent on the additional triggering factors associated with liver damage outlined above and explains the low penetrance of familial PCT. PCT may also be caused by exposure to certain polyhalogenated aromatic hydrocarbons, such as hexachlorobenzene and 2,3,7,8-tetrachlorodibenzo-*p*-dioxin.

Congenital Erythropoietic Porphyria

CEP is the least common but most severe of the **cutaneous porphyrias**. The clinical features vary in severity from hydrops fetalis with death in utero—through the onset in infancy of severe skin lesions with transfusion-dependent hemolytic anemia—to mild skin lesions, resembling PCT.

With age, progressive scarring (particularly if erosions become infected), and atrophic changes lead to photomutilation with erosions of the terminal phalanges; destruction of ears, nose, and eyelids; and alopecia. Accumulation of porphyrin in bone is visible as erythrodontia—brownish-red teeth that fluoresce in UVA light. The skin changes are usually accompanied by hemolytic anemia and splenomegaly. Hemolysis may be fully compensated or mild, but in some patients, anemia is severe enough to require repeated transfusion.

CEP is an autosomal recessive disease resulting from decreased UROS activity (see Table 29.3). Patients are homoallelic or heteroallelic for mutations in the *UROS* gene or, rarely, the *GATA1* gene, which encodes an erythroid transcription factor.

Erythropoietic Protoporphyria and X-Linked Erythropoietic Protoporphyria

The protoporphyrias are characterized by life-long acute photosensitivity caused by accumulation of protoporphyrin-IX in the skin. In EPP, overproduction of protoporphyrin-IX results from decreased activity of FECH. EPP is an autosomal recessive disease where most individuals are compound heterozygotes for a *FECH* mutation that abolishes or severely decreases FECH activity and a hypomorphic *FECH* c.315-48C allele (rs2272783). In approximately 4% of EPP families, affected individuals are heteroallelic or homoallelic for *FECH* mutations. In XLEPP gain-of-function mutations in ALAS2, usually deletions within the C-terminal region lead to formation and accumulation of protoporphyrin in excess of the amount required for hemoglobinization. XLEPP is inherited in an X-linked pattern with expression of disease in males and most females. The symptoms of XLEPP are clinically indistinguishable from EPP.

Patients with either form of protoporphyria present with acute photosensitivity characterized by burning pain, which normally starts between birth and the age of 6 years. Onset after the age of 40 years is very rare; most cases are due to acquired FECH deficiency associated with myelodysplasia. Exposure to sun is followed by an intensely painful, burning,

prickling, itching sensation in the skin, usually within 5 to 30 minutes, which is most frequently on the face and the backs of the hands. Symptoms persist for several hours up to days and are not relieved by shielding the skin from light. Recurrent sun exposure leads to chronic skin changes, such as shallow linear scars over the bridge of the nose and elsewhere on the face, and the skin may become thickened and waxy, especially over the knuckles. Seasonal palmar keratoderma is present in about 3% of patients.

The most severe complication of protoporphyria is progressive hepatic failure, which is caused by the accumulation of protoporphyrin in the liver. About 15% of patients have abnormal biochemical tests of liver function, particularly increased aspartate aminotransferase; however, only about 5% of patients develop liver failure. Protoporphyria may also increase the risk of cholelithiasis, due to high concentrations of protoporphyrin in the bile.

POINTS TO REMEMBER

Cutaneous Porphyrias

- Porphyrins absorb light with wavelengths around 400 nm.
- The absorbed radiant energy is either
  - Released as red fluorescent light, or
  - Interacts with molecular oxygen to create reactive oxygen species.
- The skin lesions are therefore caused by porphyrin-catalyzed photodamage.
- Photosensitivity can occur in any porphyria where porphyrin ring structures accumulate; that is, all the porphyrias except AIP and ADP.
- Symptoms manifest clinically as an immediate or delayed photosensitivity dependent on the type of porphyrins that accumulate and their localization within the skin.
- Protection of the skin from sunlight is essential to reduce symptoms.

Abnormalities of Porphyrin Metabolism Not Caused by Porphyria

Abnormalities of porphyrin metabolism or excretion, or both, may occur in a number of conditions other than the porphyrias. A number of other conditions need to be considered when porphyrin results are interpreted from patients in whom porphyria is suspected. Many are more common than the porphyrias. These disorders include exposure to various toxins, for example lead, hereditary tyrosinemia, renal and hepatobiliary disorders, hematological disorders, and situations where there is increased heme in the gastrointestinal system (diet, bleeding, bacteria). These abnormalities and their causes are detailed in Table 29.6. Conversely, there are a number of conditions that can give rise to skin lesions identical to those that characterize bullous porphyrias, such as PCT, but in which no abnormality in accumulation or excretion of porphyrins can be demonstrated. This is usually described as "pseudoporphyria" and is mainly associated with photosensitization from prescribed drugs, sunbeds, and end-stage renal failure.

TABLE 29.6 Secondary Disorders of Porphyrin Metabolism

Sample	Increased Heme Precursors	Clinical Conditions
Urine	Coproporphyrin Aminolevulinic acid	Liver dysfunction, fever, drugs, alcohol Lead poisoning, hereditary tyrosinemia I
Feces	Protoporphyrin, other dicarboxylic porphyrins	Intestinal bleed, high protein intake
Plasma	Uroporphyrin coproporphyrin	Renal failure Liver dysfunction
Erythrocytes	Zinc protoporphyrin	Iron deficiency anemia Lead poisoning

LABORATORY DIAGNOSIS OF PORPHYRIA

Laboratory investigation is essential for the diagnosis and exclusion of the porphyrias in patients with symptoms of acute porphyria, typical cutaneous lesions, and in relatives of patients known to have porphyria because none of the clinical features are sufficiently distinctive to allow diagnosis on clinical grounds alone.

Patients With Symptoms of Porphyria

In patients with current symptoms caused by porphyria, the excessive production of heme precursors is always present. Diagnosis depends on demonstrating specific patterns of overproduction of heme precursors (see Table 29.4) and is usually straightforward, provided appropriate specimens are examined for the relevant intermediates using adequately sensitive techniques. Although the porphyrias are uncommon disorders, the variability of the associated clinical features may make it necessary to consider them as differential diagnoses in many clinical situations. Therefore it is important to apply diagnostic strategies that not only allow for the diagnosis of a porphyria, but also adequately exclude the diagnosis. The diagnostic strategy of choice depends on the mode of presentation.

Patient Presenting With Suspected Acute Attack

Failure to correctly diagnose an attack of acute porphyria not only delays appropriate life-saving treatment but may result in unnecessary surgery or administration of drugs that aggravate the attack with potentially fatal consequences. A false diagnosis of porphyria may be equally serious because of delayed vital surgery or other treatment, and may also lead to analgesic (including opiate) misuse and dependency. The essential first-line investigation is the measurement of urinary PBG by an adequately sensitive and specific method. Testing for ALA is often performed along with PBG. Based on the currently available evidence, PBG excretion is always increased during an acute attack of AIP, HCP, or VP, usually to a concentration that exceeds 10 times the upper reference limit. Normal PBG excretion,

at a time when symptoms are present, provides very strong evidence against any type of acute porphyria as their cause, except for the very rare ADP. In ADP, PBG is normal or only slightly elevated, while ALA is markedly increased. The same pattern may be observed in lead poisoning. Further investigation is usually indicated by the finding of increased urinary coproporphyrin, which is a biochemical feature of both conditions (Tables 29.4 and 29.6).

Differentiation between the acute porphyrias. Management of the attack is the same regardless of the type of porphyria, so further investigation is not a matter of urgency. However, differentiation between the acute porphyrias is necessary for deciding which gene to examine by DNA analysis and for the selection of appropriate tests for use in family studies, as well as assessing for risk of skin symptoms. The absence of skin lesions does not exclude VP or HCP (see Table 29.3).

Once the diagnosis of acute porphyria has been established, the most efficient diagnostic strategy uses plasma fluorescence scanning and fecal porphyrin fractionation with coproporphyrin III and I separation to differentiate between AIP, VP, and HCP (see Table 29.4).

Investigation of an asymptomatic patient with past symptoms consistent with an acute attack. Diagnosis may be less straightforward when a patient presents for investigation after all clinical symptoms have resolved. Appropriate investigations include those described earlier.

If all these tests are normal, the patient may have AIP in remission, but is most likely to have had a nonporphyric illness with symptoms mimicking those of acute porphyria. If clinical suspicion remains high, mutational analysis of *HMBS*, supplemented by assay of HMBS activity in mutation-negative patients, may help to distinguish these possibilities.

Patients With Cutaneous Symptoms

Active skin lesions of cutaneous porphyrias require increased circulating porphyrins to reach the skin. Thus, analysis of plasma porphyrins by fluorescence emission spectroscopy is usually the first step when diagnosing cutaneous porphyrias (Fig. 29.3). Urine ALA and PBG have no role in the diagnosis of cutaneous porphyrias. The route of further investigation to identify the specific porphyria is dictated by the clinical presentation.

Patients with acute photosensitivity without skin fragility or bullae. When suspecting EPP/XLEPP, the essential investigation consists of the measurement of erythrocyte protoporphyrin concentration using a sensitive fluorometric method. Screening tests using solvent extraction of blood or fluorescence microscopy of erythrocytes are unreliable. If the erythrocyte (proto)porphyrin concentration is within the reference interval, EPP/XLEPP is excluded. If the concentration is increased, it is important to determine whether the increase is caused mainly by free protoporphyrin, as in EPP, both free and ZPP as in XLEPP, or primarily by ZPP, as in iron deficiency and lead toxicity (Tables 29.4 and 29.6).

DNA analysis is required to confirm the diagnosis of XLEPP in patients with an increased percentage of ZPP or in patients with a family history suggestive of XLEPP.

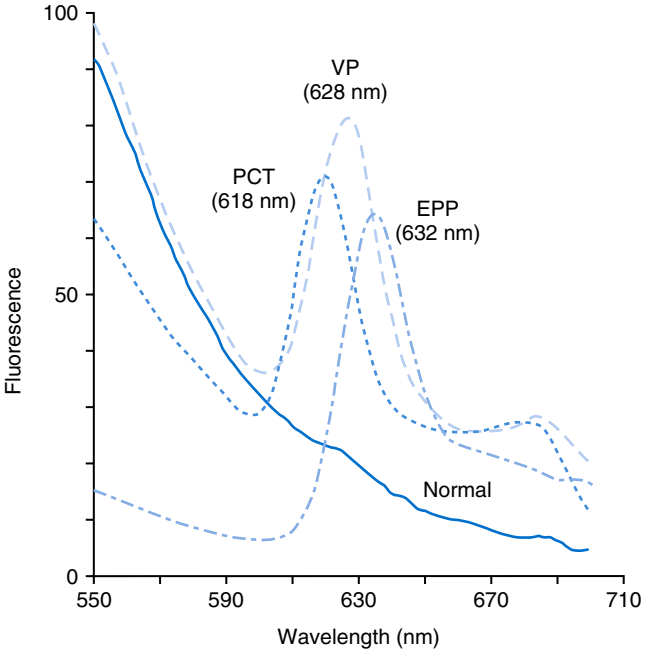


Fig. 29.3 Fluorescence emission spectra (excitation, 405 nm) of dilutions in phosphate-buffered saline (PBS) of plasma from a normal individual and from patients with various porphyrias. *EPP*, Erythropoietic protoporphyria; *PCT*, porphyria cutanea tarda; *VP*, variegate porphyria.

Patients with bullae/fragility/scarring. Clinically indistinguishable skin lesions occur in PCT, VP, HCP, CEP (see Table 29.3), hepatoerythropoietic porphyria (HEP), and other rare variants. In addition, identical skin lesions characterize pseudoporphyria, in which porphyrin metabolism is normal. Initial investigations should include the fluorescence emission spectroscopy of plasma. If this is normal, then porphyria is excluded as the cause of any active skin lesions.

A positive plasma porphyrin fluorescence scan with any increase in total urinary porphyrin should be further investigated by the determination of individual porphyrins in urine and feces using a technique capable of resolving all porphyrins of clinical interest, including isomers. The pattern observed in each of these porphyrias is unique (see Table 29.4).

Patients may present for investigation after their skin lesions have healed. In PCT, excretion and plasma porphyrin concentrations return to normal during remission, but the proportions of individual porphyrins in urine and feces may remain abnormal for longer than the total porphyrin concentrations, and the determination of individual porphyrins may suggest the diagnosis. The plasma fluorescence scan in VP and fecal coproporphyrin-III excretion in HCP remain abnormal during clinical remission.

Asymptomatic Relatives of Patients With Porphyria

Depending on the type of porphyria, investigating healthy at-risk relatives may be beneficial.

Presymptomatic Diagnosis of Autosomal Dominant Acute Porphyrias

The screening of family members to identify asymptomatic individuals who have inherited AIP, VP, or HCP, and therefore are at risk for acute attacks, is an essential part of the management of families with these disorders. Screening may be carried out by metabolite measurement, enzyme assay, and/or DNA analysis, although the use of metabolite and enzyme assays are limited by the low negative predictive value of these tests. Mutation detection by DNA analysis is the method of choice for family studies unless a mutation cannot be identified in the proband, as it allows asymptomatic disease to be excluded with certainty.

Erythropoietic Protoporphyria

In EPP caused by mutations in the *FECH* gene, testing the unaffected parent for the presence of the hypomorphic *FECH* c.315-48C allele is helpful in assessing the risk that a future child will have clinically overt disease; the presence of this allele increases the risk from about 1 in 100 to 1 in 4. In XLEPP, mutational analysis of the *ALAS2* gene may be required to identify female carriers, who show phenotypic and biochemical heterogeneity, reflecting the degree of X-chromosomal inactivation of the mutant gene.

Other Porphyrias

Family investigation has a more limited role in the clinical management of other porphyrias. In PCT, the autosomal dominant familial form has been identified by erythrocyte UROD assay or mutational analysis, but as yet little evidence suggests that family studies are necessary unless requested by anxious relatives, as the identification of a mutation does not alter the management of the patient or family.

For severe homozygous or compound heterozygous porphyrias, such as CEP, ADP, HEP, and homozygous dominant acute porphyrias, mutation analysis may be required if prenatal diagnosis in subsequent pregnancies is indicated.

LABORATORY ANALYSIS

The analytical methods used to diagnose and monitor porphyria are briefly described here. Full descriptions can be found in *Tietz Textbook of Clinical Chemistry and Molecular Diagnostics,* 6th edition.

Preanalytical Aspects

Specimen Collection and Stability

For porphyrin analysis, all samples must be protected from light, as urinary porphyrin concentrations have been observed to decrease by up to 50% if kept in the light for 24 hours. Urinary porphyrins and PBG are best analyzed in fresh, random (10 to 20 mL) samples collected without preservative. Very dilute urine (creatinine <2 mmol/L) is unsuitable for analysis and it may be helpful to request an early morning urine sample to ensure it is appropriately concentrated.

Urine often is of normal color in the nonacute porphyrias, except in CEP, when it is usually red, occasionally to such an extent that it is mistaken for hematuria. During an acute attack, urine may be a red–brown color because of the presence of uroporphyrin and other pigments formed by the nonenzymatic polymerization of PBG. This color may be evident immediately, or develop progressively during handling and storage.

About 5 to 10 g wet weight of feces is adequate for porphyrin measurements. Diagnostically, important changes in concentration are unlikely to occur within 36 hours at room temperature, and samples are stable for many months at −20°C.

Blood, anticoagulated with ethylenediaminetetraacetic acid (EDTA), shows no loss of protoporphyrin for up to 8 days at room temperature, and for at least 8 weeks at 4°C in the dark.

Quality Management and External Quality Assessment

Participation in external quality assessment (EQA) schemes, preferably assessing the total testing process, and the use of internal quality control (IQC), is particularly important for porphyria-related analyses. IQC is available commercially for urine analytes (porphyrin, ALA and PBG) but not fecal, plasma, or erythrocyte porphyrins. The European Porphyria Network (EPNET) operates a clinical EQA scheme, assessing the total testing process for porphyria-related analyses in all sample types. A specialized analytical EQA scheme covering all sample types is operated by the Royal College of Pathologists of Australasia with the European Molecular Quality Network (EMQN) providing a quality scheme for the molecular testing of the porphyria genes.

Methods for Porphyrins and Porphyrin Precursors in Urine and Feces

The water-soluble metabolites, PBG and ALA, are excreted by the kidneys and usually are measured in the urine in the clinical laboratory.

Porphobilinogen

Most methods for PBG are based on the reaction of Ehrlich's reagent (dimethylaminobenzaldehyde in acidic solution) with the α-methene carbon of the pyrrole ring to form a colored product variously described as "rose red" or "magenta," which has a characteristic absorption spectrum with a peak at 553 nm and a shoulder at 540 nm. Porphyrins do not contain any α-methene hydrogens and so do not react. Some other substances in urine may react with the reagent to give red products, notably urobilinogen, may inhibit the reaction, or are pigmented themselves and so mask the red chromogen. All need to be removed. This is best achieved by ion-exchange chromatography (first described by Mauzerall and Granick), but methods for accurate quantification of PBG based on this procedure are time-consuming. Methods based on high-performance liquid chromatography (HPLC) and tandem-mass spectrometry that are sufficiently sensitive to measure PBG in plasma have also been described.

There are a number of qualitative methods. The Watson-Schwartz and Hoesch methods have been criticized for poor detection limits and interferences whereas modified Mauzerall-Granick methods that use ion exchange resin reduced the time taken to perform the test to 10 minutes and produced a semiquantitative result. If a qualitative or semiquantitative screening

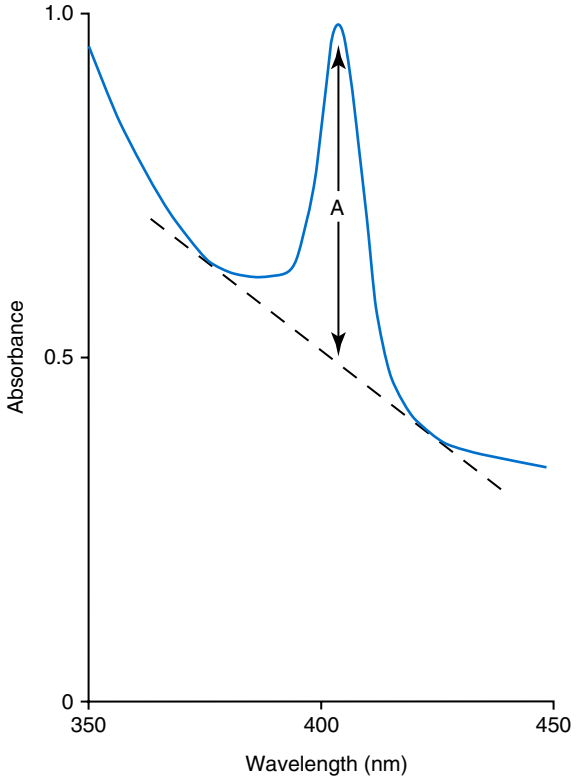


Fig. 29.4 Absorption spectrum of acidified urine showing the procedure for measurement of corrected absorbance *(A)* of the porphyrin peak.

test is used, it is essential to include appropriate controls and to confirm all positive test results using a specific quantitative method. Commercial methods for the measurement of both PBG and ALA, based on that of Mauzerall and Granick, are available (Bio-Rad Laboratories, Hercules, CA, United States; Recipe Clinical Diagnostics, Munich, Germany).

5-Aminolevulinic Acid

It is possible to measure **5-aminolevulinic acid (ALA)** directly but it is more generally converted into an Ehrlich-reacting pyrrole through condensation with a reagent, such as acetylacetone, after separation from PBG by two-stage anion-exchange chromatography (see above).

Analysis of Porphyrins in Urine and Feces

The methods for porphyrin fractionation are complex and time-consuming and are not available in every laboratory. Consequently, simple qualitative screening tests are often used to differentiate the majority of specimens that do not require further investigation from the few that justify fractionation of the individual porphyrins. Screening tests in which extracts of urine or feces are examined visually for typical red–pink fluorescence of porphyrins are not sensitive analytically and should not be used. Methods based on spectrophotometric scanning of acidified urine (Fig. 29.4) or fecal extracts for the presence of the Soret band are recommended and yield semiquantitative information. Quantitative fluorometric methods are also available.

All methods for the fractionation of porphyrins are based on varying solubilities of individual porphyrins caused by their different β-substituents and, to a lesser extent, on the substituent order around the macrocycle. Methods include (1) differential extraction with solvents, (2) paper and thin-layer chromatography, and (3) HPLC. Solvent extraction methods should not be used because they yield only limited and sometimes misleading information. Reversed-phase HPLC with fluorometric detection, the current method of choice, separates all porphyrins of clinical interest, including isomers and metal chelates, without the need for prior methylation. Representative HPLC chromatograms associated with porphyrin excretion patterns in several different porphyrias are shown in Fig. 29.5

Analysis of Blood Porphyrins

The methods described here require a spectrofluorometer with a red-sensitive photomultiplier. If such equipment is not available locally, samples must be referred to a specialized laboratory.

The most widely used method for erythrocyte total porphyrin is based on double extraction and fluorometry. If erythrocyte (proto)porphyrin concentration is increased, it is important to differentiate between free protoporphyrin, as in EPP, or ZPP, as in iron deficiency and lead toxicity (Fig. 29.6). This requires extraction with a neutral solvent, such as ethanol or acetone, to prevent the demetalation caused by strong acids, followed by fluorescence spectroscopy or HPLC to distinguish free protoporphyrin from ZPP (fluorescence emission maxima 630 nm and 587 nm, respectively).

Analysis of Plasma Porphyrins

Plasma porphyrins may be determined by fluorescence emission spectroscopy of saline-diluted plasma or HPLC analysis of deproteinized extracts. The fluorescence emission method offers the advantages of simplicity and inclusion of porphyrins that are bound covalently to plasma proteins and HPLC analysis cannot, from a diagnostic point of view, replace that method. Porphyrins at neutral pH fluoresce in the 610 to 640 nm region; the wavelength of maximum emission depends primarily on the porphyrin structure, but it is also influenced by the nature of the porphyrin-protein complex (see Fig. 29.3). Plasma fluorescence emission spectroscopy is an essential front-line investigation for suspected cutaneous porphyria.

POINTS TO REMEMBER

Diagnostics

- Identification of individual porphyrias usually requires analysis of urine, feces, and blood samples.
- All specimens must be protected from light.
- Symptomatic porphyria is always associated with increased production of porphyrins and porphyrin precursors.
  - Active acute autosomal dominant porphyria can be diagnosed by measuring increased PBG in a random urine sample.
  - Active cutaneous porphyrias can be diagnosed as a first step by finding a positive plasma porphyrin fluorescence emission screen in a freshly separated plasma sample.
- EPP cannot be diagnosed by urine porphyrin analysis, a blood sample for erythrocyte protoporphyrin analysis is essential.
- Fluorescent porphyrin analysis requires a red-sensitive photomultiplier for detection.
- Laboratories specializing in porphyria diagnosis must register with an appropriate EQA Scheme.

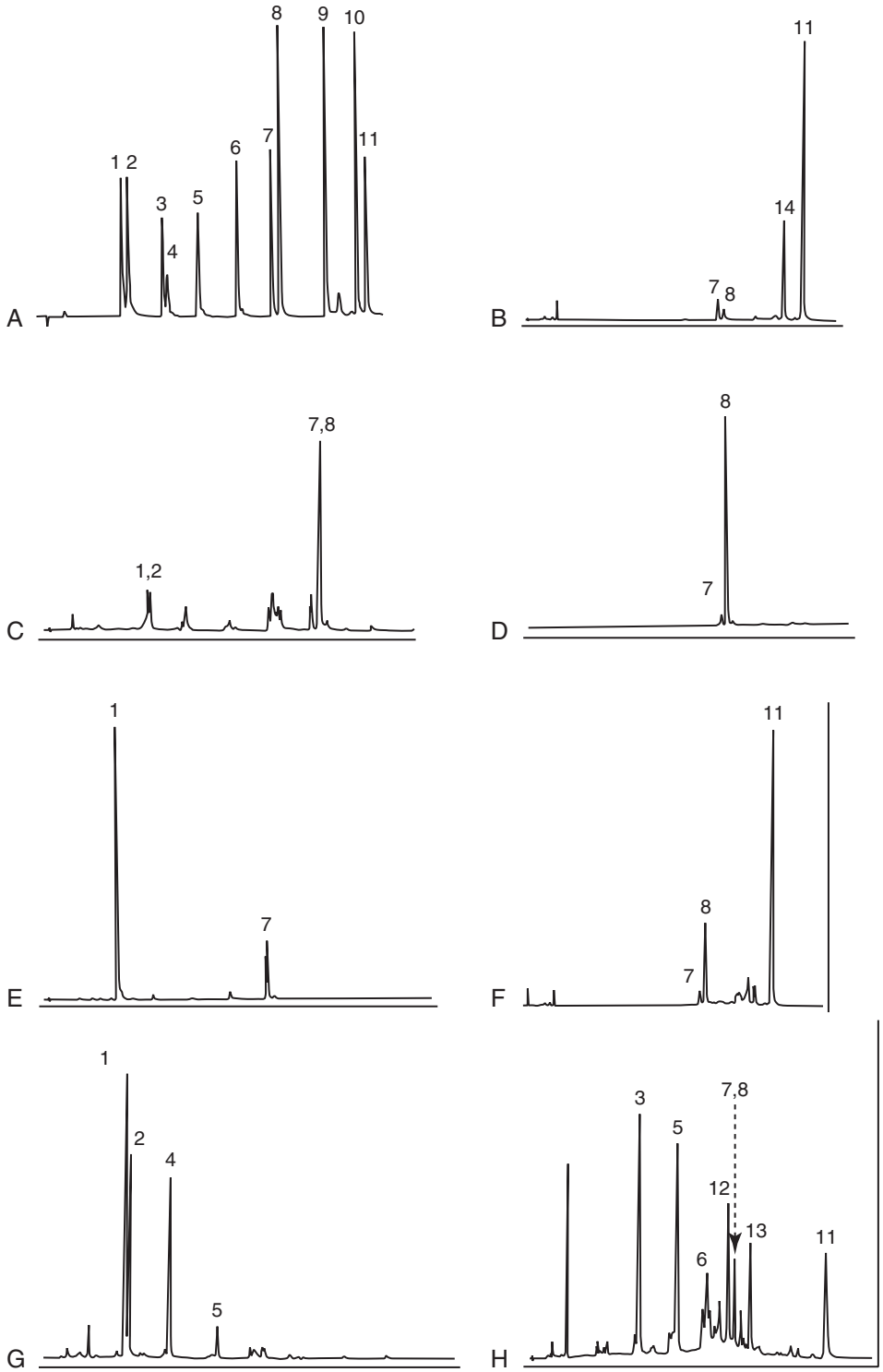


Fig. 29.5 Representative high-performance liquid chromatography chromatograms for (A) working standard, (B) normal feces, (C) normal urine, (D) feces—hereditary coproporphyria, (E) urine—congenital erythropoietic porphyria, (F) feces—variegate porphyria, (G) urine—porphyria cutanea tarda, and (H) feces—porphyria cutanea tarda chromatographic conditions. Peaks are *(1)* uroporphyrin-I, *(2)* uroporphyrin-III, *(3)* heptacarboxylate porphyrin-I, *(4)* heptacarboxylate porphyrin-III, *(5)* hexacarboxylate porphyrin, *(6)* pentacarboxylate porphyrin, *(7)* coproporphyrin-I, *(8)* coproporphyrin-III, *(9)* deuteroporphyrin-IX, *(10)* mesoporphyrin-IX, *(11)* protoporphyrin-IX, *(12)* hydroxyisocoproporphyrin, and *(13)* isocoproporphyrin.

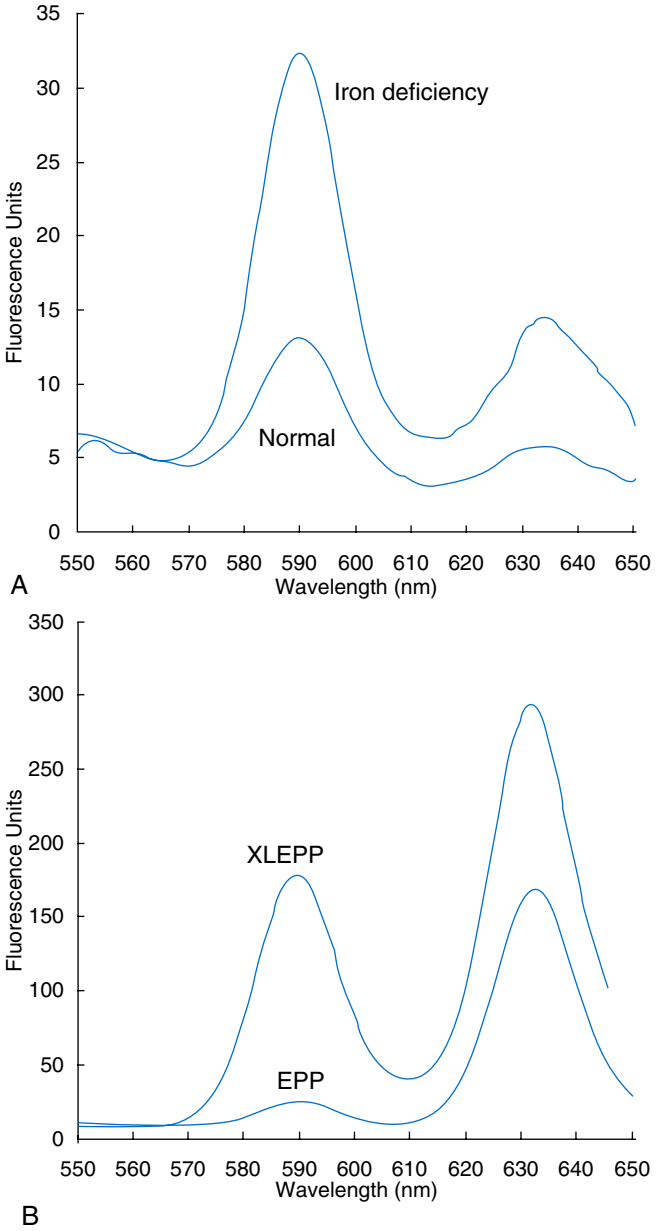


Fig. 29.6 Fluorescence emission spectra (excitation at 405 nm) of ethanolic extracts of erythrocytes from a normal individual and with iron deficiency (A) and from patients with erythropoietic protoporphyria *(EPP)* or X-linked protoporphyria *(XLEPP)* (B). Note that different scales are used.

Enzyme Measurements

Assay of individual enzymes of the heme biosynthetic pathway is rarely required for the assessment of patients with symptoms of porphyria. However, measurement of enzyme activities may be useful for family studies when it is not possible to identify the individual mutation, or when DNA analysis is not available. Erythrocytes are a convenient source of cytoplasmic enzymes (ALAD, HMBS, UROS, and UROD); however, assay of the mitochondrial enzymes (CPOX, PPOX, and FECH) requires nucleated cells, and sampling thus requires specialized techniques.

DNA Analysis

DNA analysis is required in the porphyrias mainly for family screening, for identifying the pattern of inheritance in families with EPP, and as an aid in assessing the prognosis in CEP. As most mutations are restricted to one or a few families, identification of a mutation in a new family almost always requires analysis of the entire gene.

Analysis of a gene for the presence of a mutation is most commonly carried out using direct sequencing and gene dosage analysis of genomic DNA. Unclassified variants continue to be identified in the porphyria genes, and these need to be assessed individually. Once the mutation that causes porphyria in a family has been identified, relatives are screened for its presence by direct sequencing of the region containing the mutation, or by some other mutation-specific method (see Chapter 48).

The clinical sensitivity of mutation detection in most porphyrias is above 90%, provided gene dosage analysis is included in the analysis.

POINTS TO REMEMBER

Genetic Testing in Acute Porphyria

- The main purpose is to identify family members who are at risk of an acute attack.
- The type of acute porphyria must be determined biochemically in the index case.
- The pathogenic mutation must be identified in the relevant gene.
- Family members should be appropriately counseled and offered testing for the presence of the familial mutation.
- Following appropriate counseling and consent, the testing of children is clinically justifiable.
- Patients with a predictive diagnosis of acute porphyria should be counseled on the avoidance of agents or behaviors that can trigger acute attacks.

REVIEW QUESTIONS

1. The precursors of the tetrapyrrole ring structure of porphyrin are:
   a. 5-aminolevulinic acid and iron.
   b. acetyl CoA and porphyrin.
   c. succinyl CoA and glycine.
   d. zinc and porphyrinogen.

2. The important initial screening test that allows for differentiation between the two categories of porphyria is:
   a. serum coproporphyrin.
   b. urine aminolevulinic acid (ALA) and porphobilinogen (PBG).
   c. serum zinc protophyrin.
   d. red blood cell zinc protoporphyrin.

3. Another name for protoporphyrin that contains iron is:
   a. heme.
   b. PBG.
   c. coproporphyrin.
   d. ferrochelatase.
4. The skin lesions and photosensitivity observed in patients with nonacute cutaneous porphyrias are the result of:
   a. autonomic neuropathy.
   b. excessive production of ALA.
   c. accumulation of porphyrins in the liver.
   d. excess presence of porphyrins in skin that generate oxygen radicals.
5. In lead toxicity, what replaces iron as a substrate for ferrochelatase to be incorporated into protoporphyrin?
   a. Carbon dioxide
   b. Zinc
   c. Copper
   d. Lead
6. The most common of all the porphyria disorders is:
   a. variegate porphyria.
   b. acute intermittent porphyria.
   c. porphyria cutanea tarda.
   d. congenital erythropoietic porphyria.
7. The initial steps of heme synthesis occur in the:
   a. mitochondrion.
   b. cytosol.
   c. endoplasmic reticulum.
   d. cell nucleus.
8. The best specimen type for analyzing porphobilinogen is:
   a. heparinized plasma separated and frozen immediately after collection.
   b. a 24-hour urine collection collected in a dark brown container and preserved with 0.1% HCl.
   c. a fresh early morning urine specimen collected without preservative and protected from light.
   d. blood anticoagulated with EDTA and protected from light.
9. Erythropoietic protoporphyria is characterized by which of the following symptoms?
   a. Flaccid quadriparesis caused by the accumulation of porphobilinogen in muscle
   b. Lifelong acute photosensitivity caused by protoporphyrin-IX accumulation in skin
   c. Liver damage and hepatic siderosis caused by iron accumulation in cells
   d. Severe abdominal pain and peripheral neuropathy caused by the induction of hepatic cytochromes
10. The reduced form of a porphyrin is known as a:
   a. protoporphyrin.
   b. heme molecule.
   c. pyrrole ring.
   d. porphyrinogen.

SUGGESTED READINGS

Aarsand, A. K., Villanger, J. H., Støle, E., Deybach, J. C., Marsden, J., To-Figueras, J., et al. (2001). European specialist porphyria laboratories: diagnostic strategies, analytical quality, clinical interpretation, and reporting as assessed by an external quality assurance program. *Clinical Chemistry*, *57*, 1514–1523.

Anderson, K. E., Sassa, S., Bishop, D. F., & Desnick, R. J. (2000). Disorders of heme biosynthesis: X-linked sideroblastic anemia and the porphyrias. In C. R. Scriver, A. L. Beaudet, W. S. Sly, & D. Valle (Eds.), *The Metabolic and Molecular Basis of Inherited Disease* (8th ed.) (pp. 2961–3062). New York: McGraw-Hill.

Danton, M., & Lim, C. K. (2006). Porphyrin profiles in blood, urine and feces by HPLC/electrospray ionization tandem mass spectrometry. *Biomedical Chromatography*, *20*(6–7), 612–621.

Desnick, R. J., & Aplin, K. H. (2002). Congenital erythropoietic porphyria: advances in pathogenesis and treatment. *British Journal of Haematology*, *117*, 779–795.

Elder, G. H. (2003). Porphyria cutanea tarda and related disorders. In K. M. Kadish, K. M. Smith, & R. Guilard (Eds.), *The Porphyrin Handbook Medical Aspects of Porphyrias: Vol. 14*. (pp. 67–92). Amsterdam: Academic Press.

Elder, G., Harper, P., Badminton, M., Sandberg, S., & Deybach, J. C. (2013). The incidence of inherited porphyrias in Europe. *Journal of Inherited Metabolic Disease*, *36*, 849–857.

Gouya, L., Puy, H., Robreau, A. M., Lamoril, J., Da Silva, V., Grandchamp, B., et al. (2002). The penetrance of autosomal dominant erythropoietic protoporphyria is modulated by expression of wild type FECH. *Nature Genetics*, *30*, 23–27.

Hift, R. J., & Meissner, P. N. (2005). An analysis of 112 acute porphyric attacks in Cape Town, South Africa. *Medicine*, *84*, 48–60.

Hindmarsh, J. T., Oliveras, L., & Greenway, D. C. (1999). Plasma porphyrins in the porphyrias. *Clinical Chemistry*, *45*, 1070–1076.

Holme, S. A., Anstey, A. V., Finlay, A. Y., Elder, G. H., & Badminton, M. N. (2006). Erythropoietic protoporphyria in the United Kingdom: clinical features and effect on quality of life. *The British journal of Dermatology*, *155*, 574–581.

Lim, C. K., & Peters, T. J. (1984). Urine and faecal porphyrin profiles by reversed-phase high-performance liquid chromatography in the porphyrias. *Clinica Chimica Acta*, *139*, 55–63.

Marsden, J. T., & Rees, D. C. (2014). Urinary excretion of porphyrins, porphobilinogen and δ-aminolaevulinic acid following an attack of acute intermittent porphyria. *Journal of Clinical Pathology*, *67*, 60–65.

Mauzerall, D., & Granick, S. (1956). The occurrence and determination of delta-aminolevulinic acid and porphobilinogen in urine. *The Journal of Biological Chemistry*, *219*, 435–446.

Meyer, U. A., Schuurmans, M. M., & Lindberg, R. L. (1998). Acute porphyrias: pathogenesis of neurological manifestations. *Seminars in Liver Disease*, *18*, 43–52.

Phillips, J. D., Bergonia, H. A., Reilly, C. A., Franklin, M. R., & Kushner, J. P. (2007). A porphomethene inhibitor of uroporphyrinogen decarboxylase causes porphyria cutanea tarda. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 5079–5084.

Puy, H., Gouya, L., & Deybach, J.-C. (2010). Porphyrias. *Lancet*, *375*, 924–937.

Stein, P., Badminton, M., Barth, J., Rees, D., & Stewart, M. F. (2013). Best practice guidelines on clinical management of acute attacks of porphyria and their complications. *Annals of Clinical Biochemistry*, *50*, 217–223.

Whatley, S. D., Ducamp, S., Gouya, L., Grandchamp, B., Beaumont, C., Badminton, M. N., et al. (2008). C-terminal deletions in the ALAS2 gene lead to gain of function and cause X-linked dominant protoporphyria without anemia or iron overload. *American Journal of Human Genetics*, *83*, 408–414.

Whatley, S. D., Mason, N. G., Woolf, J. R., Newcombe, R. G., Elder, G. H., & Badminton, M. N. (2009). Diagnostic strategies for autosomal dominant acute porphyrias: retrospective analysis of 467 unrelated patients referred for mutational analysis of the HMBS, CPOX, or PPOX gene. *Clinical Chemistry*, *55*, 1406–1414.

Woolf, J., Marsden, J. T., Degg, T., Whatley, S., Reed, P., Brazil, N., et al. (2017). Best practice guidelines on first-line laboratory testing for porphyria. *Annals of Clinical Biochemistry*, *54*, 188–198.

Further references with a list of useful online resources can be found in the *Tietz Textbook of Clinical Chemistry and Molecular Diagnostics,* 6th edition.

Therapeutic Drug Monitoring

*Michael C. Milone, Leslie M. Shaw**

OBJECTIVES

1. Review basic principles of applied pharmacokinetics and drug monitoring
2. Discuss the rationale for drug monitoring and some of the pitfalls of applying it in the clinical setting
3. List drugs that are commonly monitored during their use

KEY WORDS AND DEFINITIONS

Antiepileptic A substance to prevent or alleviate seizures.

Applied pharmacokinetics The use of pharmacokinetic principles to guide the clinical application of therapeutic drugs.

Beta-blocker A drug that induces adrenergic blockade at either Pl- or P2-adrenergic receptors or at both.

Bioavailability The fraction of a drug absorbed into the systemic circulation relative to the same drug administered intravenously.

Biotransformation The chemical alteration of a xenobiotic within the body that generally enhances the xenobiotic's aqueous solubility and excretion.

Dose-response relationship The relationship between the dose of an administered drug and the response of the organism to the drug.

Drug half-life The time required for one-half of an administered drug to be lost through metabolism and elimination.

Drug interactions The effects of one drug on the intestinal absorption, metabolism, or action of another drug.

Enzyme induction Increased synthesis of an enzyme in response to an inducer or other stimulus.

First-pass effect Extensive metabolism of a drug with a high hepatic extraction rate by the liver before it reaches the systemic circulation.

Generic drug A drug not protected by a trademark. Also, the scientific name as opposed to the proprietary, brand name.

Immunophilin A generic term for an intracellular protein that binds immunosuppressive drugs such as cyclosporine, FK 506, or rapamycin.

Immunosuppressant An agent capable of suppressing immune responses.

Pharmacodynamics The study of the biochemical and physiological effects of drugs and the mechanisms of their actions, including the correlation of actions and effects of drugs with their chemical structure; also, such effects on the actions of a particular drug or drugs.

Pharmacokinetics The activity or fate of drugs in the body over a period of time, including the processes of absorption, distribution, localization in tissues, biotransformation, and excretion.

Therapeutic drug monitoring The process of using drug concentration or other pharmacodynamics biomarkers to guide drug dosing.

Toxicology Subdiscipline of pharmacology concerned with adverse effects of chemicals on living systems.

Xenobiotics A chemical substance foreign to the biological system.

The ability of medicines to both heal and hurt has been recognized since ancient times. Immortalized by the writing of the Renaissance Swiss-German physician, Paracelsus over 500 years ago:

> *All things are poison and nothing is without poison, only the dose makes that a thing is not a poison.*

The challenge in medicine is therefore to determine the optimal dose of medicine that will help a patient with limited associated harm.

Studies in the 1990s identified adverse events associated with drug therapy as ranking within the top 10 causes of death in the United States. Although many severe or fatal events may not be preventable, more than one-third appear to be preventable. In addition to the tragic life-and-death consequences of adverse drug events (ADEs), failed drug therapy also has a significant economic impact. The estimated cost

*The authors gratefully acknowledge the contributions by Christine L.H. Snozek, Gwendolyn A. McMillan, and Thomas P. Moyer on which this chapter is based.

of ADEs ranges from $17 to $29 billion in the United States alone. Just examining hospital-associated ADEs, the average cost of treating each preventable event is estimated to be greater than $3000 in addition to the increased length of stay. The causes of preventable drug-related adverse events vary; however, inadequate monitoring of therapy represents major sources of preventable ADEs contributing to as much as 40% of these events. Improving monitoring strategies is therefore likely to have an important impact on both health and its associated cost.

There are many ways in which drug therapy can be monitored. Clinical signs and symptoms of toxicity are often an effective way to detect toxicity or treatment failure. **Beta-blockers** represent a typical example. Blood pressure and heart rate monitoring can be used to assess the efficacy and toxicity of these drugs. Both are also easily measurable in the clinical setting and even at home. As a result, there is little need to perform monitoring beyond these straightforward clinical assessments.

The efficacy and toxicity of some drugs, however, can be much more difficult to monitor based on clinical signs and symptoms alone. Insulin treatment in diabetes represents a case in point. The consequences of inappropriate insulin treatment are insidious, potentially life threatening, and very difficult to detect. Excessive insulin can lead to an acute decrease in blood glucose culminating in coma, permanent brain injury, and death. Inadequate insulin dosing, although not as acutely life threatening, leads over time to vascular disease, end-stage renal failure, blindness, and neuropathy, which result in significant morbidity and mortality. Laboratory testing of blood glucose, a direct biomarker of insulin's mechanism of action, provides an ideal means to monitor insulin therapy and prevent these complications. It is so important in diabetes therapy that it has driven the commercial development of simple, point-of-care (POC) devices that patients can use to routinely monitor their therapy at home. Biomarkers of drug efficacy or toxicity like blood glucose, although highly desirable, are not always available. In the absence of a useful drug effect biomarker, measuring the drug itself provides a potential surrogate.

Therapeutic drug monitoring (TDM) is the traditional term used for the activity of measuring drug concentrations to tailor the dose of the medication to an individual. There is an implicit assumption in TDM of a relationship between drug concentrations and efficacy or toxicity outcomes. The use of TDM and **applied pharmacokinetics** to guide drug therapy began in the 1960s, coincident with the development of robust analytical techniques. Since that time, TDM has become the standard of care for monitoring therapy with many drugs including antiepileptic drugs (AEDs), immunosuppressive drugs (ISDs), and antibiotics. TDM is a complex process that involves several members of the health-care team including pharmacists, laboratory professionals, and physicians. To justify the costs associated with TDM, it must improve clinical outcome and reduce the overall cost of drug therapy. Prospective, randomized, concentration-controlled trials (RCCTs) of TDM are limited; however, some of these RCCTs show that concentration control can improve efficacy and reduce drug therapy toxicity. The cost-effectiveness of TDM is lacking for most drugs, but TDM has been shown to improve the costs of aminoglycoside therapy. TDM may also aid in the detection of nonadherence to drug therapy, which represents a frequent and important cause of preventable ADEs. This benefit may even extend to drugs not routinely monitored by traditional blood concentration monitoring.

FUNDAMENTAL PRINCIPLES OF APPLIED PHARMACOKINETICS

Basic Concepts and Definitions

Pharmacology comprises that body of knowledge surrounding chemical agents and their effects on living processes. This is a broad field, and it has traditionally been confined to those drugs that are useful in the prevention, diagnosis, and treatment of disease. Pharmacotherapeutics is that part of pharmacology concerned primarily with the application or administration of drugs to patients for the purpose of prevention and treatment of disease. For this aspect of medical practice to be effective, the pharmacodynamic and pharmacokinetic properties of drugs should be understood. **Toxicology** is the subdiscipline of pharmacology that is concerned with adverse effects of chemicals on living systems. Toxic effects and mechanisms of action may be different from therapeutic effects and mechanisms for the same drug. Similarly, at the high dosage of drugs at which toxic effects may be produced, rate processes are frequently altered compared with those at therapeutic doses. For these reasons, the terms *toxicodynamics* and *toxicokinetics* are now applied to these special situations.

Pharmacodynamics

Pharmacodynamics encompasses the processes of interaction of pharmacologically active substances with target sites, and the biochemical and physiological consequences leading to therapeutic or adverse effects. For many drugs, the ultimate effect or *mechanism of action* at the molecular level is poorly understood, if at all. A pharmacological effect (i.e., the therapeutic or toxic response to a drug) may be elicited either by direct interaction of the drug with the receptor that controls a specific function or by a drug-mediated alteration of the physiological process regulating the function. For most drugs, the intensity and duration of the observed pharmacological effect are proportional to the concentration of the drug at the receptor. In a given tissue, the site at which a drug acts to initiate events leading to a specific biological effect is called the drug's *site of action*.

The *mechanism of action* of a drug is the biochemical or physical process that occurs at the *site of action*. Drug action is usually mediated through a receptor. Cellular enzymes and structural or transport proteins are important examples of drug receptors. Nonprotein macromolecules may also bind drugs, resulting in altered cellular functions controlled by membrane permeability or DNA transcription. Some drugs are chemically similar to important natural endogenous

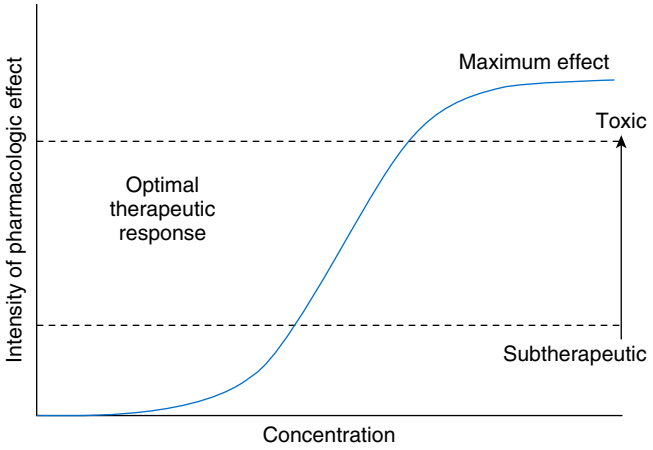


Fig. 30.1 The dose-effect relationship. The probability of increasing pharmacologic response and risk of toxicity parallels concentration for most drugs. The plateau (maximum effect) is likely due to saturation at the receptor.

substances and may compete for binding sites. In addition, some drugs may block formation, release, uptake, or transport of essential substances. Others may produce an effect by interacting with relatively small molecules to form complexes that actively bind to receptors. These and other examples of receptor binding are more completely discussed in pharmacology texts.

Although the exact molecular interactions that give rise to the mechanism of action for many drugs remain obscure, numerous theoretical models have been developed to explain drug action. One concept postulates that a drug binds to intracellular macromolecular receptors through ionic and hydrogen bonds and van der Waals forces. This theoretical model further postulates that if the drug-receptor complex is sufficiently stable and able to modify the target system, an observable pharmacological response will occur. As Fig. 30.1 illustrates, the response is concentration dependent until a maximal effect is reached. The plateau may be due to saturation at the receptor or overload of a transport process.

The utility of monitoring drug concentration is based on the premise that pharmacologic response correlates with the concentration of the drug at the site of action (receptor). Although attempts have been made to measure the concentration of drugs at the receptor site in a patient, in general, this approach is technically impractical, if not impossible for most drugs. Studies have shown that for many drugs, a strong correlation exists between the serum drug concentration and the observed pharmacological effect. In addition, years of relating blood concentrations to drug effects have demonstrated the clinical utility of drug concentration information. One must nevertheless always keep in mind that a serum drug concentration does not necessarily equal the concentration at the receptor; it merely reflects it.

Pharmacokinetics

Pharmacokinetics describes the processes of uptake of drugs by the body, the distribution of the drugs into tissue, the **biotransformations** (i.e., metabolism) the drugs undergo, and

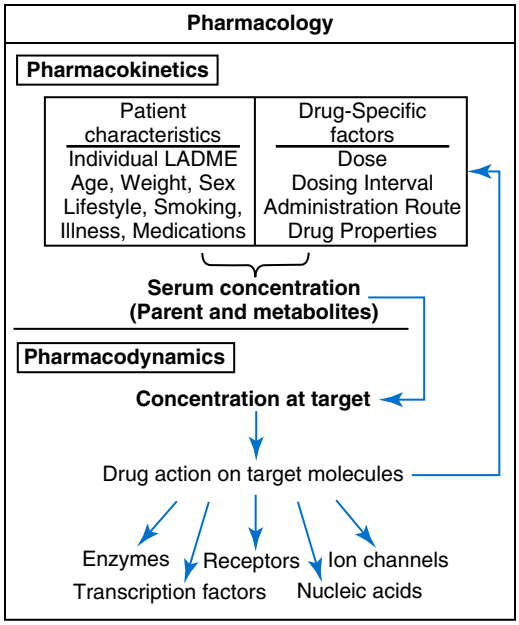


Fig. 30.2 Conceptual relationship between pharmacodynamics and pharmacokinetics. Many drug- and patient-specific factors determine both the serum concentration of a therapeutic compound and its metabolites. Serum concentration, in turn, is related to the amount of drug present at the target site, resulting in effects on a variety of molecules (several examples shown) to induce a pharmacologic response. The efficacy and degree of response provide feedback to allow optimization of the dosing regimen for an individual patient.

the elimination of the drugs and their metabolites from the body. Applied pharmacokinetics is the discipline that uses the principles of pharmacokinetics to enhance the safety and effectiveness of a drug in an individual patient. It is this aspect of pharmacology that most strongly influences the interpretation of TDM results and that is dealt with in more detail in this chapter. Fig. 30.2 illustrates the conceptual relationship between pharmacodynamics and pharmacokinetics, and the many factors affecting drug concentration and pharmacological response.

Drugs Constantly Undergo Change in the Body

Many factors are now recognized as having a profound influence on the pharmacokinetics of drugs and consequently on a patient's pharmacological response (Box 30.1). For example, the consideration of the patient's history, with particular emphasis on his or her pathophysiologic state and adjunct drug therapy, is essential at the initiation of drug therapy and TDM because these important factors may affect the absorption, distribution, metabolism, and excretion (ADME) of a drug.

Absorption

Most drugs administered chronically to patients are administered extravascularly. Although intramuscular and subcutaneous routes are used, the oral route accounts for most of the extravascular doses administered. The absorption process depends on the drug dissociating from its dosing form, dissolving in gastrointestinal fluids, and then diffusing across

BOX 30.1 Factors That Affect Drug Distribution in Humans

Demographic Factors
Age
Weight
Gender
Race
Genetics (e.g., metabolic enzyme polymorphisms)

Health-Related Factors
Liver disease (cirrhosis, hepatitis, cholestasis)
Kidney disease
Thyroid disease (hypothyroidism or hyperthyroidism)
Cardiovascular disease (arrhythmias, congestive heart failure)
Gastrointestinal disease (e.g., sprue or other malabsorption, peptic ulcer disease)
Cancer
Surgery
Burns
Volume status (e.g., dehydration)
Nutritional status (cachetic or anorexic)
Pregnancy or other factors affecting plasma proteins or body composition

Extracorporeal Factors
Hemodialysis
Peritoneal dialysis
Cardiopulmonary bypass
Hypothermia or hyperthermia

Chemical and Environmental Factors Influencing
Absorption of drug
Food or coadministered drug affecting extent or rate of absorption
Immediate- or extended-release formulation
Distribution of drug
Coadministered drugs affecting protein binding to plasma proteins or tissue
Metabolism of drug
Food, herbs, or drugs competing for metabolism
Coadministration of drugs that induce metabolic enzymes (e.g., phenobarbital)
Coadministration of drugs that inhibit metabolic enzymes (e.g., cimetidine)
Excretion of drug
Coadministration of a drug that competes for renal tubular secretory pathways (e.g., probenecid, penicillin)
Coadministration of drugs that enhance renal tubular reabsorption

biological membrane barriers into the bloodstream. The rate and extent of drug absorption may vary considerably depending on the nature of the drug itself (e.g., solubility, pK_a), on the matrix in which it is present, and on the physiological environment (e.g., pH, gastrointestinal motility, vascularity).

The fraction of a drug dose that is absorbed into the systemic circulation is referred to as its **bioavailability**. The bioavailability of a particular drug, if the drug is to be useful, must generally be great enough so that the active component can pass in sufficient amount and in a desirable time from the gut into the systemic circulation. Bioavailability of greater than 70% is most desirable for drugs to be orally useful. An exception would be a case in which the lumen of the gastrointestinal tract is the site of drug action (e.g., antibiotics used to sterilize the gut, such as oral vancomycin). Low bioavailability would then be considered advantageous.

Some drugs that are rapidly and completely absorbed nevertheless have low bioavailability to the systemic circulation. This is true of drugs with a high *hepatic extraction rate.* After oral administration, drugs that are absorbed in the lumen of the small intestine are carried by the portal vein directly to the liver. The liver may extensively metabolize a drug with a high hepatic extraction rate before it reaches the systemic circulation leading to low oral bioavailability. This phenomenon is called the **first-pass effect**.

In addition to the extent of absorption, the rate of absorption is also important. The absorption of a drug is generally considered a first-order process, and the absorption rate constant of a drug is usually much greater than its elimination rate constant. Efforts are now being made in the pharmaceutical industry to decrease the apparent rate of absorption of many drugs by manipulating their formulations to produce sustained (or extended) release products. Formulations that provide sustained release permit drugs taken orally to be taken at less frequent intervals. Conditions that may influence the extent or rate of drug absorption include abnormal gastrointestinal motility, diseases of the stomach and the small and large intestines, gastrointestinal infections, radiation, food, and interaction with other substances in the gastrointestinal tract.

Distribution

After a drug enters the vascular compartment, it interacts with various blood constituents and is carried by various transport processes to different body organs and tissues. The overall process is referred to as *distribution.* The factors determining the distribution pattern of a drug are binding of the drug to circulating blood components, binding to fixed receptors, passage of the drug through membrane barriers, and the ability to dissolve in structural or storage lipids. Molecular weight, pK_a, lipid solubility, and other physical and chemical properties of the drug are important determinants of distribution.

Once a drug enters the systemic circulation, it distributes and comes to equilibrium with many of the blood components, such as plasma proteins. An equilibrium exists between *free* and *protein-bound* drug fractions. It is generally believed that only the free fraction of the drug is available for distribution and elimination. In addition, only the free drug is available to transit across cellular membranes or to interact with the drug receptor to elicit a biological response. Therefore changes in the protein-binding characteristics of a drug can have a profound influence on the distribution and elimination of a drug, and on the manner in which total plasma or serum steady-state

concentrations are interpreted. Each drug has its own characteristic protein-binding pattern that depends on its physical and chemical properties. As a general rule, however, acidic drugs are bound primarily to albumin, and basic drugs primarily to globulins, particularly α_1-acid glycoprotein (AAG). Some drugs bind to both albumin and globulins.

Anything that alters the concentration of free drug in the plasma ultimately alters the amount of drug available to enter the tissues and interact with specific receptor systems. Disease states can alter free drug concentrations. For example, in uremia, the composition of plasma is altered by an increase in nonprotein nitrogen compounds, by acid-base and electrolyte imbalances, and often by a decrease in albumin; free drug concentrations are frequently increased. It is important to also recognize that even though the total drug concentration may remain unchanged, displacement of a drug from its plasma protein–binding sites increases free drug concentrations and can result in clinical toxicity. For example, phenytoin is 90% bound and 10% free in healthy subjects. In uremic patients, 20% to 30% of the total plasma concentration of phenytoin may be free. If one considers a healthy patient who has a total plasma phenytoin concentration of 15 μg/mL, the free phenytoin concentration is likely to be 1.5 μg/mL. If a uremic patient has a total concentration of 15 μg/mL, the free drug concentration may be 4.5 μg/mL. A free phenytoin concentration of 4.5 μg/mL is sufficient to precipitate severe phenytoin side effects, including lethargy and increased seizure frequency. In uremic patients, it is advisable to quantitate free phenytoin concentrations and adjust the drug dose to maintain free phenytoin concentration at approximately 2.0 μg/mL.

Equilibrium dialysis represents the gold-standard method for measuring the free, unbound concentration of a drug. However, this method typically requires 16 to 18 hours of incubation to achieve equilibrium, which severely limits the turnaround time for testing. Ultrafiltration techniques are a useful alternative that can usually be accomplished in a fraction of the time. In ultrafiltration, a sample of serum or plasma is forced through a filter membrane with a low molecular weight cutoff value, typically by centrifugation, to yield a protein-free sample. Provided this process is done rapidly and under appropriate temperature control, ultrafiltration can provide a useful estimate of the free drug concentration in circulating blood.

Metabolism

Biotransformation

The liver is the principal organ responsible for xenobiotic metabolism. One of its major roles is to convert lipophilic nonpolar molecules to more polar water-soluble forms. The drug molecule (a xenobiotic) can be modified by phase I reactions, which alter chemical structure by oxidation, reduction, or hydrolysis; or by phase II reactions, which conjugate the drug (glucuronidation or sulfation) to water-soluble forms. Typically, both phase I and phase II reactions occur. Most drug metabolism takes place in the microsomal fraction of the hepatocytes, where many environmental chemicals and endogenous biochemicals (**xenobiotics**) are also processed and by the same mechanisms.

Enzymes of the hepatic microsomal system can be induced or inhibited. Microsomal **enzyme induction** leads to an increase in the activity of the enzymes present, most commonly through increases in the number of oxidizing enzymes. The many isoenzymes of cytochrome P450 are affected variably by different enzyme-inducing drugs. Phenobarbital and theophylline represent prototypical enzyme inducers with broad induction effects; however, they induce different cytochrome P450 isoforms leading to effects on different drugs. Because the drug-metabolizing enzymes of the liver are also relatively nonspecific and interact with a wide variety of endogenous and exogenous substances, the presence of one drug may inhibit the metabolism of a second drug that is coadministered. When patients are on a drug with a narrow therapeutic index (TI), their dosing regimen would need to be adjusted should a known enzyme-inducing or inhibiting drug be added to or deleted from their therapy.

The role of TDM becomes particularly apparent for drugs that undergo hepatic metabolism. Wide variability in the rate of metabolism of any given drug exists not only in different patients in the general population, but also in the same patient at different times and in different circumstances. This variability is due to factors such as age, weight, gender, genetics, exposure to environmental substances, diet, coadministered drugs, and disease. Furthermore, unlike kidney function, where creatinine provides a useful biomarker of function, there is no acceptable, endogenous, biochemical marker by which hepatic function, and consequently hepatic capability for drug clearance, can be routinely assessed before drug therapy is initiated.

Excretion

Excretion of drugs or chemicals from the body can occur through biliary, intestinal, pulmonary, or renal routes. Although each of these represents a possible mechanism of drug elimination, renal excretion is a major pathway for the elimination of most water-soluble drugs or metabolites and is important in TDM. Alterations in renal function may have a profound effect on the clearance and apparent half-life of the parent compound or its active metabolite(s); decreased renal function causes increased serum drug concentrations and increases the pharmacological response.

Kidney function, in contrast to liver function, is readily and reliably evaluated by estimation of creatinine clearance. Renal clearance of creatinine at 120 mL/min approximates the glomerular filtration rate of 90 to 130 mL/min (see also Chapter 21). Therefore, measurement of creatinine clearance or estimated glomerular filtration rate (eGFR) on a routine basis provides an effective tool to evaluate kidney function. A strong correlation has been shown to exist between creatinine clearance and the total body clearance or elimination rate constant of those drugs primarily dependent on the kidneys for their elimination. Examples of drugs whose therapeutic use is adjusted to account for changes in creatinine clearance include the aminoglycoside antibiotics gentamicin, tobramycin, and amikacin.

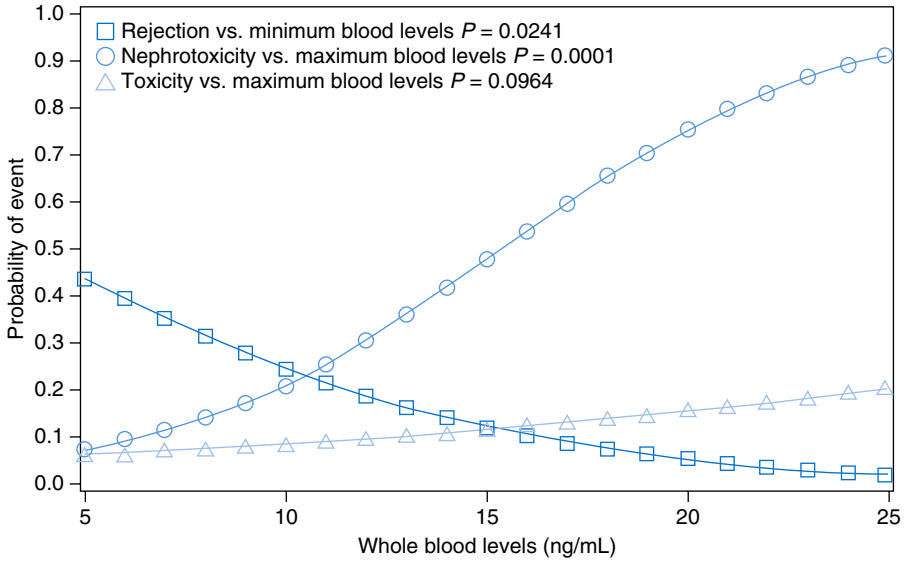


Fig. 30.3 The pharmacodynamics of tacrolimus in renal transplantation. (Reprinted with permission from Venkataramanan, R., Shaw, L. M., Sarkozi, L., Mullins, R., Pirsch, J., MacFarlane, G., Shaked, A. [2001]. Clinical utility of monitoring tacrolimus blood concentrations in liver transplant patients. *Journal of Clinical Pharmacology, 41*[5], 542–551.)

General Considerations for the Clinical Use of Therapeutic Drug Monitoring

There is no universal set of rules that determine whether a drug might benefit from TDM; however, numerous characteristics of a drug contribute to the need for TDM. TDM is most valuable when the drug in question is used chronically, has variable pharmacokinetics, and has a narrow TI. For drugs with a narrow TI, there is little if any window between blood (or serum for those drugs routinely measured in serum) concentrations associated with efficacy and those associated with toxic effects. The immunosuppressive calcineurin inhibitor drug, tacrolimus, provides a good example of a narrow TI drug with a wide variability in pharmacokinetics as shown in Fig. 30.3. The concentrations of tacrolimus that are associated with efficacy (i.e., freedom from graft rejection) and those associated with kidney toxicity overlap considerably. TDM helps navigate this "tightrope." By itself, however, narrow TI is not necessarily sufficient to warrant TDM. Many narrow TI drugs are routinely used without monitoring such as most cancer chemotherapy drugs. Additional factors therefore contribute to the need for TDM such as the severity of failed drug therapy (e.g., antimicrobials) or toxicity (e.g., ISDs). Difficulty in recognizing efficacy or toxicity with use of these drugs can be life threatening. Many patients, especially those with chronic disease, also require prolonged drug therapy. The problem of compliance is particularly evident with patients who are characteristically free of pain or unusual discomfort as with epilepsy, asthma, hypertension, mild heart disease, and transplantation. Patients may develop a sense that their disease has been cured and they no longer need the drug. The end result of noncompliance is exacerbation of the existing disorder and treatment failure. Drug concentration values therefore provide positive feedback to physicians regarding compliance.

Preanalytical Factors That Affect Therapeutic Drug Monitoring Results

To interpret a TDM result, it is critical that the dose of drug given to the patient is known. Dosage uniformity standards exist for US Food and Drug Administration (FDA)-approved drugs. Although most drug formulations are fairly consistent, the actual dose of a drug in a single tablet, suspension, or vial may vary. United States Pharmacopeia (USP) standards dictate that the mean dose contained in a tablet must be within ±10% of the product labeling and no tested sample may fall outside of 75% and 125% of the mean. As a result, the dose content for a single dose could theoretically vary by as much as 25% and still fall within acceptable USP content uniformity limits. This is further compounded by inaccuracies introduced by manipulation of the dose (e.g., splitting of scored tablets, failure to administer or consume the entire dose or from vomiting). Drugs may also degrade over time, which depends on storage, leading to further dose inaccuracy.

Beyond dose accuracy, pharmacokinetics is inherently time based. Thus, accurate timing of sample collection is often an important factor, especially when timed samples, such as peak or trough samples, are the primary means of monitoring. Vancomycin trough samples provide an example where collection immediately prior to the next dose is recommended due to the relatively short **half-life** of this drug (~6 to 12 hours). Premature sampling may lead to falsely increased concentrations that could alter decision making given the generally narrow therapeutic range for this drug. In contrast, drugs with slow distribution but long half-life, such as digoxin, are optimally sampled in the postdistributive phase of greater than 6 hours following dosing. Sampling during the steady state (generally >4 doses) is also an important assumption made for monitoring of many drugs such as vancomycin.

Variation in collection and handling of samples for TDM can also affect the quality of concentration data. Although serum is the preferred sample for monitoring of many drugs, different collection tubes are available for generating serum that include plain glass or plastic tubes without additive (e.g., red-stoppered BD Vacutainer tubes) and specialized serum separator tubes (e.g., yellow-stoppered BD Vacutainer tubes) that contain a gel that separates serum from the cellular components following centrifugation. The latter tube may affect drug concentration due to adsorption of the drug by the gel following prolonged contact. The adsorptive effects vary across drugs, and have been shown to significantly reduce the concentration of some drugs, such as phenytoin, particularly when the collection tubes are underfilled. The stability of drugs following collection may also vary with some drugs requiring prompt serum generation following by freezing for optimal stability (e.g., busulfan).

The importance of preanalytical factors in the accuracy and precision of TDM results should not be underestimated. TDM laboratories are part of a larger health-care system. It is imperative that laboratories engage and work with the many parts of this system including pharmacists, nurses, and the phlebotomy team. Each individual plays an important role in TDM, and the effective control of the many factors that affect TDM accuracy are key to the success of this complex feedback control system.

POINTS TO REMEMBER

Preanalytic factors affecting TDM results:
- Dose accuracy
- Appropriate sampling time
- Specimen collection and handling
- Physiologic changes in the patient (e.g., serum albumin concentration and phenytoin distribution)

Analytical Factors That Affect Monitoring Results

The laboratory, of course, plays a central role in the TDM process. Beyond the preanalytical factors affecting TDM results, analytical factors are important considerations in the interpretation of drug concentration data. Notable challenges in the analytical field include the availability of standardized reference material and the selectivity of methods for the drug of interest.

Determining whether a method gives accurate results over the long term and across laboratories is often a difficult task. This is generally accomplished by comparing the results of testing against a definitive analytical method or by using a certified reference material. Definitive methods generally employ very time-consuming and expensive techniques (e.g., isotope dilution-mass spectrometry) to establish material purity and quantity that are not practical for general use in most clinical laboratories. The availability of certified reference material (also frequently referred to as standardized reference material) to which methods may be compared is a critical factor in helping to control systematic error (refer to Chapter 7). Early studies of method comparability for AEDs demonstrated that concentrations varied greatly across laboratories as well as within laboratories over time. These data led to the development in the early 1980s by the National Institute for Standards and Technology (NIST) of a standard reference material of sufficient purity and accuracy to allow standardization of AED monitoring methods across different laboratories and manufacturers. Unfortunately, this type of standardized reference material is unavailable for the majority of monitored drugs.

The ability of a given method to detect a compound of interest among many potential substances present in a sample (referred to as method selectivity) represents another potential source of systematic error. Selectivity is particularly important for TDM tests because many drugs are structurally related, including to some endogenously produced compounds. Most drugs are also extensively metabolized through several different biotransformation pathways. These metabolites are often similar to the parent compound in structure, but pharmacologic activity is variable. Tacrolimus provides a useful example. Antibodies directed against tacrolimus that are used for immunoassays demonstrate similar reactivity toward tacrolimus along with two of its metabolites, 15-desmethyl-tacrolimus and 31-desmethyl-tacrolimus. Although both metabolites are detected equally, the 15-desmethyl metabolite exhibits little immunosuppressive activity, whereas the 31-desmethyl metabolite exhibits equal immunosuppressive activity when compared to that of tacrolimus. This metabolite interference with target drug detection can lead to inappropriate changes in drug dosing when substantial and not recognized. Beyond metabolites, interference can also occur from unrecognized substances within the sample matrix such as bilirubin. For immunoassay-based testing, heterophile antibodies, such as human-anti-mouse antibody (HAMA), are difficult to detect and can adversely affect testing in unpredictable ways (see Chapter 15).

Concept of a Therapeutic Range

The relationship between serum drug concentration and clinical outcome is the basis for TDM as depicted in Fig. 30.1. One of the most common interpretative errors encountered regarding TDM is the presumption that a concentration within a reported therapeutic range will ensure treatment success in the absence of toxicity, and concentrations outside that range will not. In reality, the probability of success and toxicity are a continuum across the concentration range. These probabilities are maximized and minimized, respectively, within the therapeutic range. Nevertheless, therapeutic success can certainly occur at concentrations below the therapeutic range. More importantly, success may also be achieved by concentrations above the therapeutic range without associated toxicity in some patients. Experienced users of TDM often recognize these relationships, and weigh the risks and potential benefits of drug concentrations in the context of the clinical situation using the therapeutic range as a guide.

Similar to a reference interval for any laboratory test, therapeutic ranges are based on population data from drug therapy trials, and in some instances, confirmed or established by randomized concentration-controlled trials. The application of these population-based ranges to an individual therefore assumes similarity to the target population. Differences in metabolism, physiology, and/or the underlying disease process can alter these relationships leading to unexpected and undesirable results. Phenytoin used for seizure control illustrates this limitation. A fairly good relationship between serum phenytoin concentration and seizure control exists. However, several factors have been shown to alter the **dose-response relationships** for phenytoin that might impact the utility of defined therapeutic ranges. The presence of concomitant liver disease significantly alters the pharmacokinetics of phenytoin due to its high degree of binding to albumin, a protein that is significantly altered with moderate–severe liver disease. As a result, individuals with liver disease are at significantly higher risk of toxicity at total serum phenytoin concentrations that are otherwise appropriate for individuals without liver disease. The recognition of this problem led to the development of methods for measuring phenytoin that is unbound to albumin in serum (also known as free phenytoin), which also correlates with seizure control. Nevertheless, this example illustrates the need for caution in applying population-based relationships to individual patients who may not be represented by the population used for defining the therapeutic range.

POINTS TO REMEMBER

Analytical factors affecting TDM results:

- Method selectivity for the target drug versus its metabolites
- Method precision relative to the desired therapeutic target range
- Availability of a standardized reference material
- Interfering substances in the sample (e.g., antibodies that react with assay reagents)

THERAPEUTIC DRUG MONITORING OF SPECIFIC DRUGS IN COMMON CLINICAL USE

Antiepileptic Drugs

Phenytoin

Phenytoin (diphenylhydantoin), most commonly available as *Dilantin* but also available in **generic** form, is used in the treatment of primary or secondary generalized tonic-clonic seizures, partial or complex-partial seizures, and status epilepticus. The drug is not effective for absence seizures. Phenytoin likely has many targets, but the most well-described mechanism of action is the modulation of voltage-gated sodium channels through prolonged channel inactivation, which reduces the ability of the neuron to respond at high frequency. The physiological effect of this action is reduction in central synaptic transmission, which aids in control of abnormal neuronal excitability.

Phenytoin is not readily soluble in aqueous solutions. When administered by intramuscular injection, most of the dose precipitates at the site of injection and is then slowly absorbed. A prodrug called *fosphenytoin* (Cerebyx) was introduced as a therapeutic form of phenytoin to improve phenytoin's pharmacology. Fosphenytoin has increased aqueous solubility for intramuscular injection. After injection, it is rapidly converted to phenytoin. Absorption of oral phenytoin is slow and sometimes incomplete. Variations in the drug preparation have been blamed for low bioavailability. Once absorbed, the drug is tightly bound to protein (90% to 95%). As with all drugs, the pharmacological effect of phenytoin is directly related to the amount present in the free (unbound) state. Only free phenytoin is available to cross biological membranes and interact at biologically important binding sites. The degree of protein binding can be reduced by the presence of other drugs, anemia, and hypoalbuminemia, which can occur in the elderly.

The optimal total phenytoin therapeutic concentration for seizure control without side effects is 10 to 20 μg/mL (40 to 79 μmol/L). Free phenytoin concentration in the range of 1 to 2 μg/mL (1 to 8 μmol/L) is frequently considered optimal. A side effect of phenytoin not related to plasma concentration is development of gingival hyperplasia.

Phenytoin is metabolized by the liver, and may become saturated within the therapeutic range due to its zero-order pharmacokinetics. Once metabolism is saturated, small dose increments result in large changes in blood concentration (Fig. 30.4); this phenomenon partially explains the wide variation in dosage among patients that is required to accomplish a therapeutic effect. Because of this saturation phenomenon, first-order kinetics do not generally apply to total phenytoin at blood concentrations in excess of 5 μg/mL (20 μmol/L).

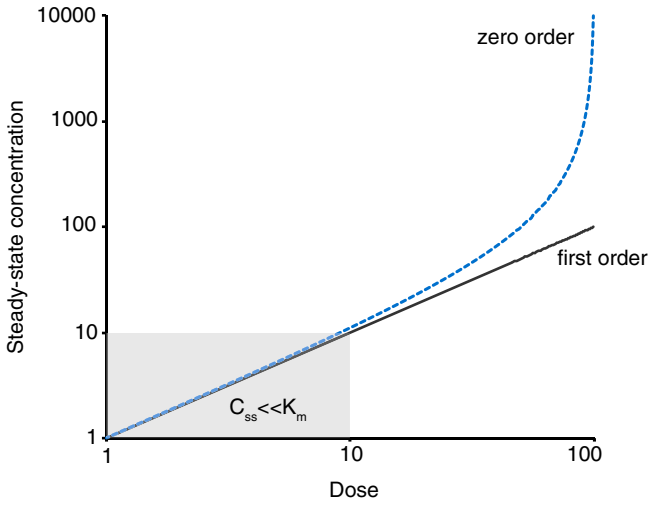


Fig. 30.4 Nonlinear response to dose changes. Drugs with first-order kinetics *(solid black line)* display serum steady-state concentrations (C) that vary proportionately with dose (D). In contrast, for drugs with zero-order kinetics *(dotted blue line)*, an increase in dose may result in a disproportionate increase in serum steady-state concentrations.

A number of **drug interactions** result in alteration of the disposition of phenytoin by enhancing metabolism (e.g., alcohol, barbiturates, and carbamazepine), competing with metabolism (e.g., chloramphenicol, cimetidine, disulfiram, isoniazid, and dicumarol), or displacing phenytoin from normal binding to serum protein (e.g., salicylate and valproic acid).

Carbamazepine

Carbamazepine, proprietary name *Tegretol,* is used in the treatment of generalized tonic-clonic, partial, and partial-complex seizures. It is also used for the treatment of pain associated with trigeminal neuralgia and as a mood-stabilizing drug in bipolar disorder. Like phenytoin, carbamazepine modulates the synaptic sodium channel, which prolongs inactivation, reducing the ability of the neuron to respond at high frequency. The physiological effect of this action is reduction in central synaptic transmission, which aids in control of abnormal neuronal excitability. The mechanisms of action in mood stabilization are less clear.

After oral administration, carbamazepine is slowly and erratically absorbed with wide individual variability. The drug is highly protein bound (80%). The elimination half-life early in therapy is approximately 24 hours. With chronic therapy, the enzymes responsible for metabolism are induced, and the elimination half-life is reduced to 15 to 20 hours. Because hepatic metabolism is the principal means by which the drug is eliminated from plasma, any reduction in liver function results in drug accumulation.

The therapeutic concentration range for optimal pharmacological effect of carbamazepine is 4 to 12 μg/mL (17 to 51 μmol/L); however, this range is dependent upon concomitant use of other AEDs. Toxicity associated with excessive carbamazepine ingestion occurs at plasma concentrations in excess of 15 μg/mL (63 μmol/L) and is characterized by symptoms of blurred vision, paresthesia, nystagmus, ataxia, drowsiness, and diplopia. Side effects unrelated to plasma concentration include development of an urticarial rash, which usually disappears on discontinuation of the drug, and hematologic depression (leukopenia, thrombocytopenia, and aplastic anemia).

The active metabolite of carbamazepine is carbamazepine–10,11–epoxide. This metabolite has been found to accumulate in children to concentrations equivalent to carbamazepine. It may contribute to symptoms of intoxication in children who have a therapeutic plasma concentration of the parent drug. Because carbamazepine is metabolized through the hepatic oxidative enzyme system, drugs that induce this system (phenytoin, phenobarbital) increase the rate of carbamazepine clearance.

Coadministration of phenobarbital, phenytoin, or valproic acid increases the rate of carbamazepine metabolism and reduces the blood concentration. Erythromycin and propoxyphene interfere with metabolism and increase carbamazepine concentrations.

Because of carbamazepine's relatively long half-life, the specimen yielding the most useful information is the one representing the trough concentration, although in the case of suspected mild intoxication, the peak value of the plasma concentration correlates more closely with toxicity. The peak specimen should be collected 4 to 8 hours after the oral dose in the setting of immediate-release formulations of carbamazepine.

Valproic Acid

Valproic acid, brand name *Depakene* or *Depakote*, is used for treatment of absence seizures. It has also been shown to be useful against tonic-clonic and partial seizures when used in conjunction with other AEDs such as phenobarbital or phenytoin. Beyond its use in the treatment of epilepsy, valproate also has mood-stabilizing effects that make it a useful agent, and alternative, in the treatment of bipolar disorder. The drug inhibits the enzyme gamma-aminobutyric acid (GABA) transaminase, resulting in an increase in the concentration of GABA in the brain. GABA is a potent inhibitor of presynaptic and postsynaptic discharges in the central nervous system.

Valproic acid is rapidly and almost completely absorbed after oral administration. Peak concentrations occur 1 to 4 hours after an oral dose. The principal metabolite, 2-*n*-propyl-3-ketopentanoic acid, has anticonvulsant activity comparable to that of valproic acid, although this metabolite does not accumulate in plasma. The single-dose half-life is 16 hours in healthy adults, but this reduces to 12 hours on chronic therapy and may be as short as 8 hours in children. In neonates and in hepatic disease, when metabolism is reduced, the half-life becomes prolonged. Valproic acid is highly protein bound (93%). In circumstances when competition for protein binding increases, such as in uremia, cirrhosis, or concurrent drug therapy, the percent of free valproic acid increases. The minimum effective therapeutic concentration of valproic acid is 50 μg/mL (346 μmol/L). Concentrations in excess of 100 μg/mL (693 μmol/L) have been associated with hepatic toxicity and acute toxic encephalopathy.

Ethosuximide

Ethosuximide, proprietary name *Zarontin,* is used for the treatment of absence seizures characterized by brief loss of consciousness. Ethosuximide reduces the flow of calcium through T-type calcium channels in the synapse of thalamic neurons; because thalamic neurons are the main source of 3-Hz spike-wave rhythms in absence seizures, reduction of calcium flow slows the rate of these seizure-inducing pulses.

Ethosuximide is readily absorbed from the gastrointestinal tract. The drug is cleared mainly by metabolism as either the hydroxyethyl compound or the glucuronide ester of the hydroxyethyl metabolite with a half-life of approximately 33 hours, although this may be prolonged in adults. The trough specimen yields the most useful information regarding therapeutic efficacy. The optimal therapeutic concentration of ethosuximide is 40 to 100 μg/mL (283 to 708 μmol/L). Toxicity related to an excessive blood concentration of ethosuximide is rare. Symptoms of gastrointestinal distress, lethargy, dizziness, and euphoria may be encountered early in therapy, but patients usually become tolerant to these symptoms.

Topiramate

Topiramate, brand name *Topamax*, is a sulfamate-substituted monosaccharide anticonvulsant that is also approved for use in migraine headache therapy. The mechanisms by which topiramate exerts these effects is not clearly established. It is proposed that several effects may contribute to topiramate's pharmacologic activity including blockage of voltage-dependent sodium channels, augmentation of the neurotransmitter gamma-aminobutyrate action at some of the subtypes of the GABA-A receptor, antagonism of the AMPA/kainite subtype of glutamate receptors, and inhibition of isozymes II and IV of carbonic anhydrase.

Brand name Topamax is generally well and rapidly absorbed following oral administration with a usual T_{max} between 2 and 4 hours and bioavailability of up to 95%. Coingestion of food can delay absorption by approximately 2 hours without effect on T_{max}. Average steady-state serum concentrations can fall by approximately 50% when either phenytoin or carbamazepine is coadministered. The metabolism of topiramate is not well understood or described, but renal clearance of unchanged drug has been reported to account for the majority of clearance in the absence of coadministered inducing drugs like phenytoin or carbamazepine. In the presence of the latter, the contribution of hepatic clearance increases. Thus, upon introduction of one of these inducing drugs into the patient's regimen, or withdrawal, closer monitoring of serum concentrations of topiramate is warranted. In the presence of significant renal disease, lowering the dosage and judicious use of TDM is recommended. Topiramate is only weakly bound to plasma proteins, and therefore circumstances that alter drug binding in serum do not affect topiramate clearance or steady-state concentrations. Based on retrospective studies and a concentration-controlled study, the usual range of concentrations is 5 to 20 μg/mL (15 to 59 μmol/L).

Lamotrigine

Lamotrigine, proprietary name *Lamictal,* is not a GABA analog, but binds to the GABA receptor; it is therefore considered a GABA receptor agonist. Lamotrigine acts like phenytoin and carbamazepine, and blocks repetitive nerve firings induced by depolarization of spinal cord neurons. Lamotrigine was approved by the FDA in 1994 for adjunctive therapy of partial seizures in adults. It has yet to be approved for use in children. Studies also suggest lamotrigine is effective against absence seizures.

Lamotrigine is well tolerated and completely absorbed from the gastrointestinal tract after oral administration. It is 60% bound to plasma proteins. Optimal response appears to occur with blood concentrations in the range of 2.5 to 15 μg/mL (~10 to 60 μmol/L). However, dizziness, ataxia, diplopia, blurred vision, nausea, and vomiting are signs of toxicity that may occur when the blood concentration exceeds 10 μg/mL (40 μmol/L). Half-life ranges from 15 to 35 hours with monotherapy. Elimination occurs through hepatic metabolism; the primary metabolite is the glucuronide ester. Coadministration with cytochrome P_{450}-inducing drugs such as phenobarbital, phenytoin, or carbamazepine results in reduced lamotrigine concentrations; dosage increases of approximately 30% are required to maintain optimal blood concentrations.

Levetiracetam

Levetiracetam, marketed under the trade name *Keppra*, is an anticonvulsant drug approved in the United States for adjunctive therapy in patients with partial seizures. Levetiracetam belongs to the lactam class of molecules that share a 5-member pyrrolidone ring, which include piracetam and ethosuximide. The mechanism of action for levetiracetam, although not fully understood, appears to be via binding to the synaptic vesicle protein, SV2A, leading to a reduction in the rate of synaptic vesicle release.

Levetiracetam, available in both oral and intravenous formulations, is mostly absorbed (>95%) within the gastrointestinal tract. Food intake causes a modest delay in absorption with reduced peak serum concentrations; however, overall bioavailability is unaffected. Levetiracetam is metabolized to the inactive acetamide form by hydrolysis (~27% to 34%) with the majority of the drug excreted by the kidneys unchanged into urine. Clearance of the drug is therefore affected by kidney disease. An increase in levetiracetam clearance has been observed during the third trimester of pregnancy; however, this change in clearance is variable. Because there is no significant hepatic metabolism of levetiracetam, liver function does not affect the pharmacokinetics of this drug.

Serum concentrations of levetiracetam appear to correlate linearly with dose within the typical dosing range. Although the optimal target range for serum concentration has not been fully defined, retrospective analysis of data from clinical trials of levetiracetam suggests that concentrations within the range of 12 to 46 μg/mL (70 to 270 μmol/L) are associated with efficacy.

Felbamate

Felbamate, proprietary name *Felbatol,* was approved by the FDA in 1993 for primary or adjunctive therapy of partial seizures. Its use is limited to those patients who fail other drug treatments because felbamate carries with it a substantial risk of aplastic anemia and liver failure that is not related to blood concentration. Biweekly monitoring of complete blood count, serum aminotransferases, and bilirubin is recommended to detect early onset of these side effects. Felbamate is particularly effective in control of Lennox-Gastaut syndrome.

Felbamate is completely absorbed from the gastrointestinal tract. The drug is 30% bound to plasma proteins, and optimal blood concentrations for felbamate, although poorly defined, has been suggested to range from 30 to 60 μg/mL (126 to 252 μmol/L). It is eliminated by hepatic metabolism, with its half-life ranging from 14 to 21 hours. Felbamate saturates metabolism when the concentration exceeds 120 μg/mL (504 μmol/L); at that concentration, metabolism converts from first order to zero order. There are currently no commercially available immunoassays for felbamate. High-performance liquid chromatography (HPLC) and capillary electrophoresis have been reported for felbamate analysis.

TABLE 30.1 Pharmacokinetic Parameters of Antiepileptic Drugs

Drug	RECOMMENDED THERAPEUTIC RANGE µg/mL	µmol/L	Mean Time to Steady State (days)	Observed Range of Half-Life in Adults (h)	Mean Oral Bioavailability (%)	Protein Binding (%)	Important Metabolizing Enzymes
Carbamazepine	4–12	17–51	2–4	8–12	70	75	CYP3A4
Clonazepam	0.015–0.060	0.048–0.190	3–10	17–56	>90	85	CYP3A4
Ethosuximide	40–100	283–708	7–10	30–60	>90	0	CYP3A4
Felbamate	30–60	126–252	3–4	14–21	>90	25	CYP3A4
Gabapentin	2–12	12–70	1–2	5–9	Variable	0	NA
Lamotrigine	2.5–15	10–59	3–6	20–30	>90	55	NA
Levetiracetam	12–46	70–270	1–2	6–8	>90	0	NA
Phenobarbital	10–40	43–172	12–24	70–140	>90	50	CYP2C19
Phenytoin	10–20 (free: 1.0–2.0)	40–79	5–17	30–100	80	90	CYP2C9, 2C19
Primidone	5–10	23–46	2–4	3–22	>90	20	CYP2C9, 2C19
Topiramate	5–20	15–59	4–5	20–30	80	15	NA
Valproic acid	50–100	346–693	2–4	11–20	>90	90	CYP2C9,2C19, 2B6, 2E1, 2A6
Zonisamide	10–40	47–188	9–12	50–70	65	50	CYP2C9, 3A4

Phenobarbital

Phenobarbital is used in the treatment of all seizures except absence seizures, and is known by a wide variety of proprietary names and found in combination with many other drugs. It is particularly useful for treatment of generalized tonic-clonic, partial, focal motor, temporal lobe, and febrile seizures.

Absorption of oral phenobarbital is slow but complete. The time at which peak plasma concentrations are reached is widely variable and ranges from 4 to 10 hours after the dose. Phenobarbital is 40% to 60% bound to plasma proteins. The elimination half-life is from 70 to 100 hours and is age dependent (children average 70 hours, geriatric patients 100 hours). Because hepatic metabolism is one of the prime routes of elimination, reduced liver function results in prolonged half-life.

The optimally effective therapeutic concentration of phenobarbital is between 15 and 40 µg/mL (66 to 177 µmol/L). The predominant side effect observed in adults at blood concentrations greater than 40 µg/mL (177 µmol/L) is sedation, although tolerance to this effect develops with chronic therapy.

Phenobarbital is metabolized in the liver to *p*-hydroxyphenobarbital, which is largely excreted as the glucuronide or sulfate ester. When renal and hepatic functions are decreased, patients experience decreased clearance of the drug. Elimination of phenobarbital may be decreased in the presence of valproic acid and salicylate if reduction in urinary pH occurs. During chronic administration of valproate or salicylate, the concentration of phenobarbital may increase 10% to 20%, and a dose adjustment may be necessary to avoid intoxication. Phenobarbital induces mixed-function oxidative enzymes, resulting in increased metabolism of other xenobiotics after approximately 1 to 2 weeks of therapy.

Because of the long elimination half-life of phenobarbital, the blood concentration does not change rapidly. Therefore, a serum specimen collected late in the dose interval (trough) is representative of the overall effect. Results from specimens collected 2 to 4 hours after the dose can be misleading, because they may be construed to be the peak concentration when in actuality they precede the peak. Table 30.1 summarizes pharmacokinetic data of anticonvulsant drugs.

Primidone

Primidone, proprietary name *Mysoline,* is effective in the treatment of tonic-clonic and partial seizures. The mechanism of action of this drug is similar to that described for phenobarbital, and the therapeutic effect is due partially to the accumulation of its major metabolite, phenobarbital. A second metabolite of primidone, phenylethylmalonamide, also has some **antiepileptic** activity.

Primidone is rapidly and completely absorbed after oral administration. Once absorbed, it is not highly protein bound and it has a half-life of approximately 10 hours. Disposition of the drug is not known to be significantly altered by other disease states or other drugs.

The optimal therapeutic concentration of primidone in adults has been established as 5 to 12 µg/mL (23 to 55 µmol/L). Because phenobarbital is an active metabolite of primidone, concurrent analysis of phenobarbital is required for complete result interpretation. The previously defined therapeutic range for phenobarbital applies to adequate primidone therapy. The phenobarbital concentrations rise gradually over a period of 1 to 2 weeks after therapy is initiated. Toxicity due to accumulation of primidone occurs at serum concentrations in excess of 15 µg/mL (69 µmol/L) and is usually associated with symptoms of sedation, nausea, vomiting, diplopia, dizziness,

TABLE 30.2 Pharmacokinetic Parameters of Commonly Monitored Antibiotics

Drug	Therapeutic Targets[a] (μg/mL [μmol/L])	Half-Life (h)	Oral Bioavailability (%)	Protein Binding (%)
Aminoglycosides				
Amikacin	C_{max}: 25–35 (43–60) C_{min}: 1–8 (1.7–13.7)	2[b]	NA	<11
Gentamicin	C_{max}: 5–12 (10.5–25) C_{min}: <1 (<2)	2[b]	NA	<30
Tobramycin	C_{max}: 5–12 (11–25.7) C_{min}: <1 (<2)	2[b]	NA	<15
Glycopeptides				
Vancomycin	C_{min}: >10–15 (>6.9–10.4)	6–12[c]	<1[d]	10–50
Other				
Chloramphenicol	C_{max}: 10–25 (31–77) C_{min}: 1–8 (3–25)	Adults: 1.5–4.1 Newborn: ≥24[f]	70–80[e]	60

[a]Target blood concentrations depend upon the infection (e.g., tissue compartment), organism and its sensitivity to the antibiotic (i.e., minimal inhibitory concentration).
[b]Clearance of aminoglycoside antibiotics is dependent upon kidney function. The half-life shown is the average elimination phase half-life in healthy individuals; however, the half-life may be significantly prolonged in individuals with renal disease.
[c]Vancomycin best conforms to a multicompartment (2 or 3 compartment) pharmacokinetic model. The half-life shown represents the terminal, elimination half-life.
[d]Absorption of vancomycin from the gastrointestinal tract leading to toxic concentrations has been observed in individuals with pseudomembranous colitis.
[e]Bioavailability of chloramphenicol depends upon the chemical form (succinate or palmitate).
[f]Chloramphenicol clearance is reduced in neonates due to limited glucuronide conjugating activity of the liver. Half-life is generally >24 h immediately after birth and decreases to ~10 h by 2 weeks of age.

ataxia, and a phenobarbital concentration greater than 40 μg/mL (177 μmol/L). Specimen collection is dictated by the same rules that apply for phenobarbital; the trough concentration is most useful.

Coadministration of acetazolamide with primidone results in decreased gastrointestinal absorption of primidone and subsequent diminished plasma concentrations. Primidone administered in association with phenytoin produces a modest increase in the phenobarbital/primidone ratio because phenytoin competes with the hepatic hydroxylating enzymes associated with phenobarbital's metabolism. Coadministration of valproic acid, for the same reasons outlined for phenobarbital, causes a modest increase in both primidone and phenobarbital serum concentrations.

Zonisamide

Zonisamide is the generic name used in the United States for a widely used seizure medication whose common brand name is Zonegran. The FDA approved zonisamide for use in 2000 with the suggestion that it be used together with other anticonvulsants in the treatment of partial seizures in adults.

Orally administered zonisamide is generally well absorbed with little to no effect of concomitant food consumption, and this drug is only weakly bound to plasma proteins. Zonisamide is extensively metabolized by oxidative, acetylation, and other pathways and has essentially linear pharmacokinetics resulting in linearity for doses ranging from 10 to 15 mg/kg per day. Concomitantly administered inducing drugs such as phenytoin or carbamazepine cause increased metabolism-based clearance and therefore a need for dose adjustment that can be aided by TDM of zonisamide. When concomitant therapy with an inducing drug is being withdrawn, adjustment of zonisamide dosing using TDM is warranted.

A target TDM range of 10 to 40 μg/mL (47 to 188 μmol/L) has been recommended, based largely on retrospective studies and, as with all other anticonvulsants, there is significant overlap of serum zonisamide concentrations between seizure-free patients and patients who do not respond to therapy with this drug, and between patients who encounter side effects and those who do not. Thus finding an optimal therapeutic concentration within the individual patient is an essential need, not simply titrating the patient to be within the target therapeutic range.

Antibiotics

Antibiotics that require TDM include aminoglycosides, chloramphenicol, sulfonamides, vancomycin, trimethoprim, β-lactams, and tetracyclines. Pharmacokinetic details of these antibiotics are summarized in Table 30.2.

Aminoglycosides

Aminoglycosides are polycationic agents that kill aerobic gram-negative bacteria. They act by binding to the 30S ribosomal subunit of bacterial mRNA, thereby inhibiting protein synthesis. They are inactive under anaerobic conditions because an oxygen-dependent active transport mechanism is involved in the transfer of aminoglycosides across the bacterial cell wall. The aminoglycoside class of drugs includes amikacin, gentamicin, kanamycin, neomycin, netilmicin, sisomicin, streptomycin, and tobramycin.

The aminoglycosides are a very polar group of compounds and are thus poorly absorbed from the intestinal tract. They are routinely administered intravenously or intramuscularly to achieve a high degree of bioavailability. When administered directly into the blood, they rapidly distribute to the extracellular fluid but do not cross cell membranes or bind to plasma proteins; this behavior is consistent with their unusually low volume of distribution. Most tissues and nonrenal or hepatic secretions contain very small concentrations of aminoglycosides, with the exception of the renal cortex, where the drug is concentrated, and bile, because of active hepatic secretion. The drugs are mainly excreted by glomerular filtration. Elimination half-lives are short, ranging from 2 to 3 hours. Because clearance is highly dependent on renal function, any impairment of glomerular filtration causes accumulation of these drugs.

Therapy with antibiotic agents like the aminoglycosides differs from the approach used for most other drugs discussed in this chapter. The goal is to achieve a concentration in plasma such that the bacteria are killed but the host remains undamaged. Because the organisms treated are variable and can become resistant to certain drugs, treatment with specific aminoglycoside agents should always be directed by susceptibility testing. Table 30.2 identifies target maximum peak and trough serum concentrations.

Dose corrections must be made in patients with compromised renal function because these patients have prolonged half-life and slower elimination. This should then be followed up by quantification of the blood concentration and dose adjustment following the method outlined by Gilbert.

Toxicity associated with aminoglycosides manifests as delayed-onset vestibular and cochlear sensory cell destruction and acute renal tubular necrosis. The degree and severity of cell damage are variable among the different drugs, but they all cause cell damage if the concentrations are high. Unfortunately, the therapeutic concentration guidelines identified in Table 30.2 do not guarantee the avoidance of toxicity; a small number of patients experience toxic effects regardless of the concentration. Fortunately, most patients reverse the toxic effects without direct intervention if the toxicity is associated with reasonable blood concentrations. Irreparable loss of vestibular, cochlear, or renal function usually correlates with administration of one of the aminoglycosides at increased blood concentrations for periods longer than 2 weeks.

Chloramphenicol

Chloramphenicol, proprietary name *Chloromycetin* and others, is used as a bactericidal agent. It acts by binding to the 50S ribosomal subunit of bacteria mRNA, and inhibits protein synthesis in prokaryotic organisms. Use of this drug depends on its relative toxicity against the microorganism versus the host. The drug is used against gram-negative bacteria such as *Hemophilus influenzae, Neisseria meningitidis, Neisseria gonorrhoeae, Salmonella typhi,* all *Brucella* species, *Bordetella pertussis, Vibrio cholerae,* and *Shigella.* These organisms are all susceptible to a concentration of 6 μg/mL (19 μmol/L). Organisms that require higher concentrations of 12 μg/mL (37 μmol/L) are *Escherichia coli, Klebsiella pneumoniae, Pseudomonas pseudomallei, Chlamydia,* and *Mycoplasma.*

Chloramphenicol is rapidly absorbed in the gastrointestinal tract. Peak serum concentrations occur 1 to 2 hours after the oral dose. In plasma, chloramphenicol is approximately 50% protein bound and is cleared with a half-life of 2 to 3 hours. Peak serum concentrations of chloramphenicol palmitate or succinate occur 4 to 6 hours after administration. Chloramphenicol distributes to all tissues, and it concentrates in the cerebrospinal fluid. The drug is actively metabolized by the liver by *N*-acetylation and glucuronidation. Thus, chloramphenicol accumulates in cases of hepatic disease. Renal disease does not dramatically reduce clearance.

Host toxicity displayed after chloramphenicol therapy includes hematologic toxicity and cardiovascular collapse; both show a modest relationship to blood concentration. The blood concentration–related hematologic toxicities include anemia, characterized by maturation arrest in the marrow; cytoplasmic vacuolation of early erythroid and myeloid cells; reticulocytopenia; and increases in both serum iron and serum iron–binding capacity. These symptoms all are associated with serum concentrations in excess of 25 μg/mL (77 μmol/L). Development of idiosyncratic aplastic anemia has also been observed, but this complication appears unrelated to dose or blood concentration. Cardiovascular collapse, which occurs primarily in newborns, has been related to a total serum chloramphenicol concentration in excess of 40 μg/mL (155 μmol/L). An oral dose of 50 mg/kg per day results in an optimal peak serum concentration of 10 to 25 μg/mL (31 to 77 μmol/L) in a healthy adult.

Procedures for the determination of chloramphenicol concentrations in blood serum include HPLC and immunoassay. Methods for chloramphenicol determination must be able to differentiate between the prodrug forms, chloramphenicol palmitate or succinate, and their active metabolite, chloramphenicol.

Vancomycin

Vancomycin is a glycopeptide that is bactericidal against gram-positive bacteria and some gram-negative cocci. Vancomycin is used because of its activity against methicillin-resistant staphylococci and corynebacteria. It has thus become popular for treatment of endocarditis and sepsis caused by these organisms.

Although the drug is generally poorly absorbed when given orally, absorption leading to toxicity has been observed in patients with pseudomembranous colitis. A 1-g dose given intravenously every 12 hours usually results in a peak blood concentration of 20 to 40 μg/mL (14 to 28 μmol/L) and a trough concentration of 5 to 10 μg/mL (4 to 7 μmol/L). It has an average elimination half-life of 5 to 6 hours. Blood concentration–related toxicity involves the auditory nerve. Concentrations less than 30 μg/mL (21 μmol/L) are rarely associated with this development. Toxicities not related to dose or blood concentrations include fever, phlebitis, and pain at the infusion site. Erythema or flushing of the face, neck,

and upper torso occurring within approximately 5 to 10 minutes following vancomycin infusion (sometimes referred to as "red man syndrome" or "red neck syndrome") has also been observed. This syndrome is due to acute, non-IgE-mediated mast cell degranulation, and is generally controlled by slow infusion and administration of antihistamines. In patients with impaired renal function, the serum concentration may increase to toxic concentrations because of reduced clearance.

Antifungal Antibiotics

Over the past decade, increasing evidence has accumulated to support the use of TDM to enhance the therapeutic safety and efficacy of antifungal medicines. The azole class of antifungals has the most data supporting concentration monitoring. There is also limited data supporting the monitoring of flucytosine. Studies of TDM for other classes of antifungal medicines including the echinocandins (i.e., caspofungin) and the polyenes (e.g., Amphotericin B) have generated data showing no or limited value to monitoring these drugs.

HPLC and liquid chromatography–tandem mass spectrometry (LC-MS) methods for detection of multiple azoles have been described, and represent the predominant methods used for clinical testing by reference laboratories. Although these methods are generally highly specific, it should not be assumed that results are transferrable across laboratories. A recent review of proficiency testing data that evaluated 5 years of data from 57 different laboratories around the world demonstrated that a wide variation in results, especially at low concentrations, is common leading to results that can deviate by more than 20%. New immunoassays for voriconazole and posaconazole have been described, but experience with these is limited. The availability of reliable immunoassays for these antifungals could foster more widespread experience in the effectiveness of their monitoring. Metabolism of the drugs can also affect interpretation of concentration data as discussed in more detail later.

Antineoplastic Agents

Methotrexate

Methotrexate has proved useful in (1) the management of acute lymphoblastic leukemia in children; (2) choriocarcinoma and related trophoblastic tumors in women; (3) carcinomas of the breast, tongue, pharynx, and testes; (4) maintenance of remission in leukemia; and (5) treatment of severe, debilitating psoriasis. High-dose methotrexate administration followed by leucovorin rescue is effective in treatment of carcinoma of the lung and osteogenic sarcoma. Intrathecal administration is effective in treating meningeal leukemia or lymphoma.

Methotrexate inhibits DNA synthesis by decreasing availability of pyrimidine nucleotides. Methotrexate competitively inhibits the enzyme dihydrofolate reductase, thus decreasing the concentrations of the tetrahydrofolate essential to the methylation of the pyrimidine nucleotides and consequently the rate of pyrimidine nucleotide synthesis. Leucovorin, a folate analog, is used to rescue host cells from methotrexate inhibition; as a synthetic substrate for dihydrofolate reductase, leucovorin administration allows resumption of tetrahydrofolate-dependent synthesis of pyrimidines and reinitiation of DNA synthesis. Methotrexate is a nonspecific cytotoxin, and prolongation of blood concentrations appropriate to killing tumor cells may lead to severe, unwanted cytotoxic effects such as myelosuppression, gastrointestinal mucositis, and hepatic cirrhosis.

Serum concentrations of methotrexate are commonly monitored during high-dose therapy (>50 mg/m^2) to identify the time at which active intervention by leucovorin rescue should be initiated. Criteria for blood concentrations indicative of a potential for toxicity after single-bolus high-dose therapy are as follows:

1. Methotrexate concentration greater than 10 μmol/L 24 hours after dose
2. Methotrexate concentration greater than 1 μmol/L 48 hours after dose
3. Methotrexate concentration greater than 0.1 μmol/L 72 hours after dose

Characteristically, blood concentrations are monitored at 24, 48, and 72 hours after the single dose, and leucovorin is administered when methotrexate concentrations are inappropriately high for a postdose phase. The route of elimination for methotrexate is primarily renal excretion. During the period of high blood concentrations, particular attention must be paid to maintaining output of a large volume of alkaline urine. The pK_a of methotrexate is 5.5; thus small decreases in urine pH result in significant reduction in its solubility. Keeping urinary pH alkaline diminishes the risks of intratubular precipitation of the drug and obstructive nephropathy during the treatment period. Monitoring blood concentrations therefore provides the basis for decisions for timing of initiation and continuance of leucovorin treatment and for managing urinary pH.

Methotrexate has been measured in biological specimens using a wide variety of techniques. Radioimmunoassay (RIA) and the folate reductase inhibition techniques have been used, but nonisotopic immunoassays are now the method of choice. Liquid chromatographic procedures have also been developed to allow for coanalysis of the drug and its metabolites.

Busulfan

Busulfan is a DNA alkylating agent available in both oral and intravenous formulations. High-dose busulfan is often used as part of the myeloablative preparative regimen for hematopoietic stem cell transplantation (HSCT). Clinical use of busulfan is complicated by significant interpatient variability in pharmacokinetic behavior with reported coefficients of variation of 23% and 25% for the oral and intravenous (IV) formulations, respectively. Age, obesity, underlying disease, and organ dysfunction also exert a significant influence on observed clearance for busulfan. Children younger than 6 typically display more than twice the average clearance of 2.5 mL/min per kilogram reported for adults. The observed variability in the pharmacokinetic behavior of busulfan is relevant because several studies have identified a pharmacodynamic relationship

between exposure to busulfan and both its toxicity and efficacy. Exposure to busulfan is typically estimated by measurement of several plasma concentrations over a 6-hour dose interval. This exposure is generally expressed as either the area under the concentration-time curve (referred to as the AUC) or the mean steady-state concentration ($\bar{C}_{SS}$). Dosing of busulfan is usually every 6 hours or once daily over 4 days for a total of 16 or 4 doses, respectively. As a result, a study is typically performed following the first dose, and the results are promptly reported to effect a dose adjustment as early as possible in the course of the treatment regimen; however, the completion and reporting of a busulfan pharmacokinetic study is not a simple task. A variety of chromatographic methods are used for measurement of busulfan in plasma with gas chromatography–mass spectrometry (GC-MS) and LC-MS/MS generally the preferred methods. Given the complexity of the analytical methodology, onsite testing is not always available. Sample extraction and analysis also usually require several hours to complete for a single patient. Nevertheless, using the above TDM approach with dose adjustment, exposures that are within less than 10% of the target exposure are possible, and toxicity can be avoided without a compromise in efficacy.

Immunosuppressive Drugs

Tacrolimus

Tacrolimus, proprietary name *Prograf* and formerly called FK506, is a macrolide antibiotic isolated from a strain of *Streptomyces tsukubaensis* that has significant **immunosuppressant** properties. Tacrolimus mediates its immunosuppressive action by entering the lymphocyte, and binding to a receptor known as FK506-binding protein (FKBP). Once bound to FKBP, this drug-receptor complex interacts with, and blocks, the calcium-dependent phosphatase, calcineurin, which is critical for the translocation of the nuclear factor of activated T cells (NFAT) to the nucleus, where the latter regulates the transcription of cytokines and other genes important for T cell activation, proliferation, and function. Calcineurin inhibition therefore leads to marked suppression of T cell–mediated immune responses such as those involved in solid organ transplant rejection and graft-versus-host disease following bone marrow transplantation. NFAT plays a role in other cell types, likely contributing to the toxic effects of tacrolimus.

Tacrolimus is administered predominantly orally in capsule form (Prograf) containing 0.5 mg, 1 mg or 5 mg of the drug. A sterile solution containing the equivalent of 1 mg/mL of tacrolimus for intravenous administration and an ointment containing 0.1% or 0.03% for topical use on skin are also available. Rapid but incomplete absorption are characteristic of standard oral tacrolimus formulations, with an average time to peak concentration of 1.6 to 2.3 hours and average oral bioavailability of 17% to 22%. Recently, several prolonged release formulations of tacrolimus have been developed with the first, Advagraf, now available commercially. Extended release formulations afford delayed maximal concentrations with improved bioavailability leading to a slight reduction in dose to achieve equivalent AUC.

There is a fairly good correlation between tacrolimus trough concentration, graft rejection, and toxicity. The target range for tacrolimus blood concentrations depends upon the concomitant use of other ISDs, the transplant type, and the time following transplantation. An example of the immunosuppressive regimens used at the authors' institution with the associated tacrolimus target ranges are shown in Table 30.3.

The steady-state trough concentration of tacrolimus per unit of dose varies widely within and between patients. The between-subject variability of dose-normalized tacrolimus has been estimated to be fivefold. Factors known to contribute to this variability are many and include hematocrit; plasma albumin concentration; patient age; genotypes of metabolizing enzyme CYP3A5; drug-drug, herb-drug, and food-drug interactions; and disease. The influence of one or more of these factors in transplant patients explains the wide within- and between-subject variability of tacrolimus concentrations. This, taken together with the narrow TI and the requirement of contemporary practice to lower the tacrolimus dosing and target trough concentration during the first transplant year and beyond to limit nephrotoxicity, is the basis for the need for close concentration monitoring of this drug.

Due to the high binding of tacrolimus to FK506 binding proteins within cells including erythrocytes, the plasma concentration of tacrolimus is typically 1.5% to 8% of whole blood. Methods for quantitative analysis of this drug therefore generally begin with cell lysis and protein precipitation of whole blood to liberate cell-bound tacrolimus to allow measurement of the total drug present within whole blood. The lysis step is most commonly performed as a manual step using a water-miscible solvent (e.g., methanol) solution containing $ZnSO_4$ followed by centrifugation. Performance of this initial step is one of the critical points in tacrolimus analysis, and it represents an important source of potential error if not performed correctly or consistently.

Interferences in the measurement of tacrolimus include metabolites of the drug and other substances present in human blood such as antibodies that react with components of the assay (for additional discussion on interference in immunoassay, refer to Chapter 15). The former represents the more frequent and challenging interfering substance as discussed previously in the section on Analytic Factors affecting testing results.

Cyclosporine

Cyclosporine, proprietary names *Sandimmune* (Cyclosporine A, or CsA) and *Neoral,* is a cyclic peptide composed of 11 amino acids, some of novel structure, isolated from the fungus *Trichoderma polysporum*. The compound has been shown to be effective in suppressing solid organ allograft rejection and graft-versus-host disease following bone marrow transplantation. CsA is approved for use in renal, cardiac, hepatic, pancreatic, and bone marrow transplants.

CsA acts by a mechanism that is very similar to tacrolimus. Following entry into the cell, CsA forms a complex with cytoplasmic receptors termed cyclophilins that are molecularly distinct from FKBP. These cyclosporine-cyclophilin

TABLE 30.3 Immunosuppressive Regimens and Associated Tacrolimus Target Ranges Used at the Author's Institution

Organ	Immunosuppressive Regimen	Time Posttransplant	TACROLIMUS THERAPEUTIC RANGE µg/L	TACROLIMUS THERAPEUTIC RANGE nmol/L
Kidney	Tacrolimus + MMF	0–1 months	8–12	10.4–15.6
		2–3 months	7–10	9.1–13
		4–6 months	6–8	7.8–10.4
		7–12 months	5–7 (6–8 for higher risk)	6.5–9.1 (7.8–10.4 for higher risk)
		>12 months	4–6 (5–7 for higher risk)	5.2–7.8 (6.5–9.1 for higher risk)
Heart	Tacrolimus + MMF	0–3 months	10–12	13–15.6
		4–6 months	10	13
		6–12 months	8–10	10.4–13
		>12 months	5–8	6.5–10.4
Liver	Tacrolimus + steroids +/– Azathioprine or MMF	0–3 weeks	8–12 (6–8 with renal insufficiency)	10.4–15.6 (7.8–10.4 with renal insufficiency)
		4–6 weeks	6–10 (6–8 with renal insufficiency)	7.8–13 (7.8–10.4 with renal insufficiency)
		7 weeks–9 months	5–8	6.5–10.4
		10 months–2 years	4–6 (6 for AIH; 3–4 with renal insufficiency if also on an anti-proliferative agent)	5.2–7.8 (7.8 for AIH; 3.9–5.2 with renal insufficiency if also on an anti-proliferative agent)
		>3 years	3–4 (6 for AIH)	3.9–5.2 (7.8 for AIH)
Lung	Tacrolimus + MMF + Steroids	0–12 months	8–12	10.4–15.6
		>12 months	6–8	7.8–10.4
Bone marrow	Tacrolimus		5–15	6.5–19.5

MMF, Mycophenolate mofetil.

molecular complexes interact with and inhibit the calcineurin phosphatase preventing NFAT activation that is critical for lymphocyte proliferation and function.

Absorption of CsA in the form of *Sandimmune* is highly variable, ranging from 5% to 40%. There is a poor relationship between dose and blood concentration; however, the whole-blood concentration of CsA correlates with the degree of immunosuppression and toxicity. A microemulsion form of CsA, *Neoral,* has more reproducible absorption, averaging 40%, and exhibits better correlation among dose, blood concentration, and clinical response.

Immunosuppression requires trough whole-blood concentrations of at least 100 ng/mL (83 nmol/L). Kahan and colleagues found that trough whole-blood concentrations exceeding 600 ng/mL (499 nmol/L) were associated with hepatic, renal, neurological, and infective complications. Shaw and colleagues have discussed strategies for reducing the toxicity of CsA and other ISDs.

Therapeutic trough blood concentrations of CsA for renal transplants are 100 to 300 ng/mL (83 to 250 nmol/L), whereas 200 to 350 ng/mL (166 to 291 nmol/L) is used as the target concentration for cardiac, hepatic, and pancreatic transplants; however, concomitant immunosuppression used in combination therapy, similar to combined drug therapy in other situations such as AED therapy in epilepsy, significantly affects the range of concentrations that are effective. Simultaneous immunosuppression with low-dose prednisone and either azathioprine (AZA) or mycophenolate mofetil (MMF) allows the patient to enjoy a good response to CsA at lower concentrations—some renal transplant patients obtain a satisfactory response with trough CsA concentrations of 70 ng/mL(58 nmol/L).

CsA is slowly absorbed, and peak concentrations are reached in 4 to 6 hours. Like tacrolimus, CsA is also highly (~90%) protein bound and concentrated in erythrocytes. The degree of concentration in erythrocytes is temperature dependent in vitro; thus measurement of plasma concentration requires strict attention to specimen temperature if reproducible results are to be obtained. Because of this effect, the best specimen for analysis is whole blood. CsA is also heavily metabolized, and immunoassays are generally subject to metabolite interference similar to the effects described for tacrolimus.

Several drugs alter the disposition of CsA. Ketoconazole, erythromycin, melphalan, amphotericin B, and aminoglycoside antibiotics all prolong metabolism of CsA sufficiently to increase the risk of nephrotoxicity. Coadministration of phenytoin, phenobarbital, carbamazepine, and rifampin results in induction of cytochrome P_{450} enzymes, which increase the rate at which CsA is metabolized. Intravenous administration of sulfadimidine and trimethoprim decreases CsA concentrations.

Currently, nonisotopic immunoassays performed on whole blood are the most commonly employed methods for

measurement of CsA; however, many laboratories use HPLC and LC-MS/MS methods, which are more specific and less subject to metabolite interference. It is therefore important to ensure consistency with TDM by following individuals with the same method over time and to interpret results within the context of the method used.

Sirolimus

Sirolimus, also known as rapamycin, is a macrocyclic antibiotic that is a fermentation product of the actinomycete *Streptomyces hygroscopicus* that was isolated from soil samples collected on Rapa Nui (Easter Island) following a search for novel antifungal agents. Structurally, sirolimus is a lipophilic macrocyclic lactone comprised of a 31-member macrolide ring. It demonstrates antifungal, antitumor, and immunosuppressive activity in animal model studies, and is approved in the United States for the prophylaxis of acute rejection in renal transplant patients.

The complex of sirolimus and the intracellular **immunophilin**, FKBP12, modulates the immune response by combining with the specific cell-cycle regulatory protein called mTOR (the mammalian target of rapamycin) and inhibiting its activation. This inhibition results in suppression of cytokine-driven T-lymphocyte proliferation, inhibiting the progression from the G_1 to the S phase of the cell cycle.

The metabolism of sirolimus by the human body is driven by oxidative metabolism via CYP3A in the gastrointestinal tract and liver. There are at least seven metabolites characterized as 41-*O*- and 7-*O*-demethyl, several hydroxy, hydroxy-demethylated, and didemethylated sirolimus. Total sirolimus metabolites accounted for 48% to 70%, and no single metabolite accounted for more than 10% of sirolimus in trough whole blood from stable renal transplant patients; however, the immunosuppressive activity of individual metabolites is reported to be lower than the parent drug.

Sirolimus is available as both an oral solution and tablet. Sirolimus is rapidly absorbed from the gastrointestinal tract with the average time to reach maximal concentration in whole blood of about 2 hours. The average bioavailability of sirolimus is low, at 15%, and attributable to extensive intestinal and hepatic metabolism by CYP3A and to counter-transport by the multidrug efflux pump P-glycoprotein in the gastrointestinal tract. This absorption barrier varies considerably within a single patient and from patient to patient. It is also the site of clinically important drug-drug and drug-food interactions.

Sirolimus distributes extensively into blood cells as reflected by the average blood-to-plasma ratio of 36:1 in renal transplant patients. Approximately 95% distributes into red blood cells, 3% in plasma, and 1% each in lymphocytes and granulocytes. The extensive and avid binding of sirolimus to the ubiquitously distributed intracellular FKBP12 accounts for the high blood-to-plasma sirolimus concentration ratio. Approximately 2.5% of the sirolimus within the plasma fraction is unbound with the remainder bound to plasma proteins.

The relationship between sirolimus whole-blood trough concentrations and efficacy and toxicity has been investigated in renal transplant patients who received concomitant full-dose CsA and corticosteroid therapy. According to these analyses, the minimum effective sirolimus concentration below which there is a significant increase in risk for acute rejection is 4 to 5 μg/L (4.4 to 5.5 nmol/L). The threshold concentration of 13 to 15 μg/L (14 to 16 nmol/L) was identified, above which the risks increase for the concentration-related side effects, thrombocytopenia (<100,000 platelets/mm^3), leukopenia (<4000 leukocytes/mm^3), and hypertriglyceridemia (>300 mg/dL or 3.4 mmol/L serum triglycerides). More studies are needed to define these relationships for other transplant populations and for different concomitant immunosuppressants.

Several chromatographic (i.e., LC-MS, LC-MS/MS, HPLC) detection methods have been validated and are in use in laboratories worldwide. A commercial, automated immunoassay has also been developed. This latter assay, while showing high precision, exhibits modest bias due to metabolite interference (~25%) when results are compared to chromatographic methods.

Everolimus

In April 2010, everolimus, a more water-soluble analog of sirolimus, was approved for use in CsA-sparing regimens including the requirement for adjusting everolimus doses using target trough blood concentrations in renal transplant patients. The target ranges were established from earlier retrospective studies and used prospectively to demonstrate equivalent acute rejection rates compared to the combination of standard dose CsA and empiric dose MMF. More recently, the use of everolimus has expanded with approval in liver transplantation in combination with low-dose tacrolimus.

The mechanism of action and pharmacodynamic effects of everolimus are comparable to those for sirolimus. Similar to the other immunosuppressive drugs, wide pharmacokinetic variability of everolimus has been observed. This variability is driven by the variable metabolism via CYP3A4/5 and p-glycoprotein along with frequent drug-drug interactions that are comparable to those observed for sirolimus. Everolimus trough concentrations show a significant, linear correlation with overall drug exposure as assessed by AUC with a coefficient (r^2) of 0.79. Similar degrees of correlation between everolimus trough concentration and thrombocytopenia, leucopenia, hypertriglyceridemia, or hypercholesterolemia have been observed in an investigation of 54 stable renal transplant patients (18 to 68 years). Based upon the robust correlation between trough concentration and both efficacy and toxicity, a trough concentration of 3 to 8 ng/mL (3.1 to 8.3 nmol/L) has been suggested as the optimal target range.

The bioanalytical method used for measurement of everolimus concentration in the pharmacokinetic assessments and prospective TDM protocols of many clinical investigations is a validated LC-MS/MS method. Use of this bioanalytical methodology provides sensitive and selective measurement of everolimus, which is important for reliable measurement of blood concentrations at the low end of the recommended target concentration range (3 ng/mL [3.1 nmol/L]) in currently utilized immunosuppression protocols. In the future, we can

anticipate the availability of immunoassays for everolimus, and it will be essential to understand the comparison between these methods and the current chromatographic methods.

Mycophenolic Acid

Mycophenolic acid (MPA) is a product of the *Penicillium* species that exhibits antitumor, antiviral, antifungal, antibacterial, and immunosuppressive activity. MPA is administered as its morpholinoethyl ester, MMF. MMF, also known as RS–61433, is considered a prodrug because its immunosuppressive activity is expressed only after its hydrolysis to MPA in the body. MPA inhibits inosine monophosphate dehydrogenase, an important enzyme in the purine metabolic pathway. T lymphocytes rely on this pathway for purine synthesis, whereas other cells use the hypoxanthine–guanosine ribosyl transferase salvage pathway for purine biosynthesis. Thus MPA selectively inhibits purine synthesis, and thus transcription, in T lymphocytes. MPA is of interest clinically because it has immunosuppressive activity similar to the thiopurine, AZA, but without many of its side effects.

MMF is completely absorbed and rapidly and completely metabolized to MPA, the active metabolite. The latter is metabolized in the liver by phase II enzymes to form the major metabolite, mycophenolic acid glucuronide (MPAG). The elimination half-life of MPA averages 18 hours, the volume of distribution averages 4 L/kg, and typical serum concentrations range from a peak of 12 μg/mL (37.4 μmol/L) to a trough value of 2 μg/mL (6.24 μmol/L). Based on the available studies, the optimal immunosuppression for MPA appears to be achieved with target AUCs in the 30 to 60 mmol*min/mL range. Although poorly correlated, this is approximated by trough serum concentrations in the range of 2 to 4 μg/mL (6.24 to 12.5 μmol/L).

Other Drugs

Digoxin

Digoxin, proprietary name *Lanoxin,* is one of a group of cardiac glycosides obtained from digitalis plants (e.g., *Digitalis lanata*). Although used substantially less frequently than in the past due to newer drugs, digoxin is still used for treatment of supraventricular arrhythmias such as atrial fibrillation due to its activity on atrioventricular nodal conduction. Digoxin also acts as an inotropic agent that restores the force of cardiac contraction in congestive heart failure. However, its use has decreased substantially. The drug binds to the extracytoplasmic side of the α-subunit of membrane-bound Na, K-ATPase, inhibiting both cellular Na^+ efflux, and K^+ influx in myocardial cells. This reduces the sodium/potassium gradient in the Purkinje fibers of the atrial, junctional, and ventricular myocardium, resulting in a decreased transmembrane potential. Inhibition of Na, K-ATPase is postulated to enhance movement of calcium ions in the cell, increasing calcium ion availability and improving cardiac contractility. In addition to direct effects on the excitable tissues within the heart through the Na, K-ATPase, digoxin has also been shown to alter autonomic activity within the heart to promote parasympathetic activity, which may further contribute to its mechanism of action.

At low concentrations, digoxin causes the atrium to be less electrically excitable. Moderate concentrations of digoxin are required to reduce the rate of depolarization in the spontaneously depolarizing conductive fibers (Purkinje fibers), and toxic concentrations of digoxin are necessary to diminish depolarization of the ventricular myocardium. Disagreement over the clinical value of digoxin measurements and the failure of the digoxin concentration to correlate with clinical toxicity are usually related to aberrations in serum and tissue concentrations of sodium, potassium, magnesium, and calcium. Increased sensitivity to digoxin can be noted in states of hypokalemia, hypomagnesemia, and hypercalcemia, which make establishment of the true therapeutic concentration of digoxin difficult because all parameters are interactive.

Absorption of digoxin is variable and dependent on the drug formulation. Digoxin is concentrated in tissues, and at steady state, the concentration of digoxin in cardiac tissue is 15 to 30 times that of plasma. Accumulation of digoxin in tissue lags behind the plasma concentration—that is, although the peak plasma concentration is reached 2 to 3 hours after the oral dose, the peak tissue concentration occurs 6 to 10 hours after an oral dose. Although pharmacological effects and toxicity correlate with tissue concentration rather than plasma concentration, the safe therapeutic plasma concentration of digoxin has been reported to range from 0.8 to 2.0 ng/mL (1 to 2.6 nmol/L). This range is not determined at the peak plasma concentration but rather at the time of peak tissue concentration. Thus, to ensure a correlation between plasma concentration and tissue concentration, the appropriate time to collect the *specimen* is 8 hours or more after the dose, at which time serum digoxin has reached distributional equilibrium with the drug in tissues. Results from specimens collected earlier than 8 hours after the dose are misleading because high concentrations may be misinterpreted as toxic concentrations, whereas they are more likely due to incomplete distribution.

Digoxin *toxicity* is characterized by nonspecific symptoms of nausea, vomiting, anorexia, and predominance of green/yellow visual distortion. Cardiac symptoms of intoxication include multiform premature ventricular contractions, ventricular bigeminy, ventricular tachycardia, and ventricular fibrillation. Children can tolerate higher concentrations and do not usually exhibit toxicity until the digoxin concentration exceeds 4 ng/mL (5.2 nmol/L).

Elimination of digoxin follows first-order kinetics; 50% to 70% is excreted unchanged or in the form of digoxigenin monosaccharides or disaccharides in the urine. A small amount is metabolized to dihydrodigoxin and also excreted by the kidneys. The remainder is found in the stool as digoxigenin and its saccharides.

Currently, immunoassays remain the most widely used method for measurement of digoxin in serum and biological fluids. Use of Digibind or DigiFab digoxin-specific Fab fragments, which neutralize the drug and are used in the setting of acute toxicity, can complicate digoxin measurement

by many immunoassays. Results of digoxin monitoring following administration of Digibind should therefore be interpreted with caution.

Lithium

Lithium, whose proprietary names include *Eskalith, Lithane, Lithonate,* and others, is administered as lithium carbonate and used for the treatment of the manic phase of affective disorders, mania, and manic-depressive illness. The mechanism of action for lithium is not entirely clear. Early research suggested that it acts by enhancing re-uptake of catecholamines, thereby reducing their concentration in the neuronal junction and producing a sedating effect on the central nervous system. More recently, lithium and other mood-affecting drugs such as valproic acid, carbamazepine, and the tricyclic antidepressants, have been shown to inhibit GSK3-β, a protein central to the Wnt/β-catenin signaling pathway that affects gene expression involved in many aspects of cellular behavior, neuronal polarity, plasticity and survival, and brain development. GSK3-β appears critical in the action of dopamine and serotonin on the brain affecting behavior. At least two mechanisms account for lithium's effects on GSK3-β. Lithium directly competes with magnesium, which is an important cation for GSK3-β activity. It also indirectly affects GSK3-β by inhibiting a critical phosphatase normally required for its activation. Although other mechanisms of action cannot be excluded, the mechanisms described above are likely to account for many of lithium's effects on mood.

Absorption of lithium from the gastrointestinal tract is complete, with peak plasma concentration reached 2 to 4 hours after an oral dose. This cation does not bind to protein. Lithium elimination is biphasic; during the first phase, 30% to 40% of the dose of lithium is cleared, with an apparent half-life of 24 hours. During the second phase, the remainder of lithium incorporated into the cellular ion pool is cleared, exhibiting a half-life of 48 to 72 hours. Clearance is predominantly a function of the kidneys, where active reabsorption occurs. Reduced renal function causes prolonged clearance times.

The optimal therapeutic response to lithium has not been related to a specific serum concentration; however, toxicity is related to serum concentration. Serum lithium concentrations are monitored to ensure patient compliance and to avoid intoxication. It is recommended that a standardized 12-hour postdose serum lithium concentration be used to assess adequate therapy. The interval of 0.5 to 1.2 mmol/L was identified as the optimal trough therapeutic concentration. Concentrations of 1.2 to 1.5 mmol/L signify a warning range, and a concentration in excess of 1.5 mmol/L *in a specimen drawn 12 hours after the dose* indicates a significant risk of intoxication. Early symptoms of intoxication include apathy, sluggishness, drowsiness, lethargy, speech difficulties, irregular tremors, myoclonic twitching, muscle weakness, and ataxia. These symptoms, although not life threatening, are uncomfortable for patients and indicate that the onset of life-threatening seizures is imminent.

Lithium excretion parallels that of sodium. It readily passes the glomerular membrane and is reabsorbed in the proximal convoluted tubules. In situations in which patients are vulnerable to dehydration (fever, watery stools, vomiting, loss of appetite, hot weather), the potential for lithium intoxication is increased. In dehydration, the proximal tubular response to reabsorption of sodium (and lithium) is reduction of clearance. Increased reabsorption of lithium leads to increased blood concentration of lithium. Severe intoxication, characterized by muscle rigidity, hyperactive deep tendon reflexes, and epileptic seizures, is usually associated with lithium concentrations in excess of 2.5 mmol/L.

The concentration of lithium in serum, plasma, urine, or other body fluids can be determined by several methods including flame emission photometry, atomic absorption spectrometry, ion-selective electrode or inductively coupled plasma–mass spectrometry (ICP-MS) or inductively coupled plasma–optical emission spectrometry (ICP-OES). In contemporary clinical practice, automated methods are primarily used for routine measurement of lithium in serum using a chromophore (e.g., a substituted porphyrin) that forms a colored product readily detected spectrophotometrically upon binding to lithium. The availability from NIST of a standard reference material (NIST SRM 3129a) provides manufacturers with the opportunity to prepare traceable calibrators and improve standardization of assays.

Theophylline

Theophylline, available under many proprietary names, relaxes bronchial smooth muscles to relieve or prevent asthma. The therapeutic effect of theophylline is likely due to antagonism of adenosine receptors in smooth muscle, whereas the toxic effects are due to inhibition of cyclic nucleotide phosphodiesterase. With increased use of β–adrenergic agonists, and because of the considerable toxicity associated with it, theophylline is now considered a second-level approach used only in treatment of persistent asthma.

Theophylline is readily absorbed after oral, rectal, or parenteral administration. If the drug is taken orally without food, the blood concentration peaks within 2 hours. If it is administered with food or as a slow-release formula, peak concentrations occur 3 to 5 hours after the dose. Once absorbed, it is 50% protein bound. The drug is rapidly cleared in adolescents and in adults who smoke due to its higher rate of metabolism due to increased levels of CYP1A2. In these individuals, the half-life ranges from 3 to 4 hours. Nonsmoking adults in good health have an elimination half-life of about 9 hours. The half-life in neonates and in adults with congestive heart failure can be prolonged to 20 to 30 hours, depending on the degree of liver immaturity or loss of liver function. Coadministration of cimetidine, ciprofloxacin, and ticlopidine leads to reduced clearance of theophylline.

There is a proportional relationship between forced expiratory volume and theophylline concentration, with the optimum therapeutic effect occurring at concentrations ranging from 8 to 20 μg/mL (44 to 111 μmol/L). Suppression of exercise-induced bronchospasm in asthmatic patients occurs at concentrations exceeding 10 μg/mL (56 μmol/L) and is optimal at 15 μg/mL (83 μmol/L). Neonatal apnea treated with theophylline

responds to slightly lower concentrations, ranging from 5 to 10 μg/mL (28 to 56 μmol/L). Relaxation of bronchial smooth muscle is directly proportional to blood concentration and continues at concentrations greater than 20 μg/mL (111 μmol/L). When the blood level exceeds 20 μg/mL (111 μmol/L), the secondary side effects become significant.

Theophylline typically exhibits first-order kinetics of elimination when used in most patients with serum concentrations in the range of 5 to 15 μg/mL (28 to 83 μmol/L); however, at serum concentrations greater than 20 μg/mL (111 μmol/L), theophylline exhibits zero-order pharmacokinetics with small dose increases leading to disproportionately large increases in serum concentration and intoxication. Symptoms of theophylline toxicity include nausea, vomiting, headache, diarrhea, irritability, and insomnia. Transient central nervous system stimulation occurring at initial administration is not directly related to blood concentration. This effect diminishes with chronic use. Serious toxicity characterized by cardiac arrhythmias and seizures is usually associated with serum concentrations in excess of 30 μg/mL (167 μmol/L). Once seizure activity begins, the final prognosis is very poor. Morbidity is reported in nearly all patients, and mortality can be as high as 50%.

Caffeine

A minor metabolite of theophylline in adults, caffeine has been shown to accumulate to significant concentrations in neonates. Caffeine itself is an effective inhibitor of apnea, which may explain the lower therapeutic concentration required for control of neonatal apnea. Therapy with caffeine alone has also been demonstrated as effective in the treatment of neonatal apnea; it is gaining popularity because of caffeine's long half-life in neonates (>30 hours). The optimal therapeutic concentration of caffeine in this situation ranges from 8 to 14 μg/mL (41 to 72 μmol/L). Caffeine can be measured by HPLC or immunoassay.

REVIEW QUESTIONS

1. Pharmacokinetics refers to the branch of pharmacology that deals with:
   - **a.** the time-dependent processes of absorption, distribution, metabolism, and the excretion of drugs and their metabolites from the body.
   - **b.** the relationship between a drug's concentration and its effects on its target receptor.
   - **c.** the toxic effects of a drug.
   - **d.** the distribution of drugs to patients.
   - **e.** the study of time-dependent effects of a drug on the body.
2. Which of the following statements are true regarding phenytoin?
   - **a.** Phenytoin exhibits zero-order kinetics with saturation of metabolism at conventional doses.
   - **b.** Phenytoin binds tightly to red blood cells and requires whole blood testing.
   - **c.** Phenytoin steady-state concentration is linearly proportion to the dose in most patients.
   - **d.** Phenytoin is excreted mostly unchanged in the urine.
   - **e.** Phenytoin concentration monitoring is best performed using collection of blood in a serum separator tube.
3. Which of the following best describes drug therapy with aminoglycosides?
   - **a.** The risk for ototoxicity with gentamicin is best correlated with serum trough concentration.
   - **b.** The risk of nephrotoxicity induced by amikacin is best correlated with peak drug concentration.
   - **c.** Aminoglycoside dosing should be adjusted for patients with renal insufficiency.
   - **d.** Aminoglycoside dosing should be adjusted for patients with liver dysfunction.
   - **e.** Once daily (extended interval) dosing of aminoglycoside antibiotics is frequently used in patients with end-stage renal failure.
4. Which of the following statements regarding therapeutic range are TRUE?
   - **a.** A therapeutic range is the range of drug concentrations in which patients will respond to therapy.
   - **b.** A therapeutic range is the range of drug concentrations in which the probability of efficacy is maximized and the probability of toxicity is minimized.
   - **c.** The concentration of a drug should not exceed its therapeutic range because patients will develop toxicity.
   - **d.** Therapeutic ranges are always validated in randomized prospective trials of therapeutic drug monitoring.
   - **e.** The therapeutic range of a drug is unaffected by other drugs used in combination and can therefore be applied to all situations in which a drug is used.
5. Which of the following best describes digoxin use and monitoring?
   - **a.** The time of blood collection for digoxin does not affect interpretation of digoxin concentration.
   - **b.** Digoxin is a mainstay of heart failure treatment due to its calcium channel–blocking effects.
   - **c.** Digoxin absorption is very consistent across different formulations.
   - **d.** Digoxin is frequently used to treat atrioventricular (AV) nodal dysfunction due to its enhancing effects on sympathetic nervous system activity within the heart.
   - **e.** The measurement of digoxin concentrations can be affected by prior administration of Digibind (Digoxin Immune Fab).

SUGGESTED READINGS

Atkinson, A. J. (2012). *Principles of Clinical Pharmacology* (3rd ed.). San Diego: Academic Press, Elsevier.

Bauer, L. A. (2014). *Applied Clinical Pharmacokinetics* (3rd ed.). New York: McGraw-Hill.

Bettinger, T. L., & Crismon, M. L. (2006). Lithium. In M. E. Burton (Ed.), *Applied Pharmacokinetics & Pharmacodynamics: Principles of Therapeutic Drug Monitoring* (pp. 789–812). Philadelphia: Lippincott Williams & Wilkins.

Carruthers, S. G., & K.L (2000). *Melmon, Melmon and Morrelli's Clinical Pharmacology: Basic Principles in Therapeutics* (4th ed.). New York: McGraw-Hill.

Ferrari, S., Sassoli, V., Orlandi, M., Strazzari, S., et al. (1993). Serum methotrexate (MTX) concentrations and prognosis in patients with osteosarcoma of the extremities treated with a multidrug neoadjuvant regimen. *Journal of Chemotherapy*, *5*(2), 135–141.

Gibaldi, M., & Prescott, L. F. (1983). *Handbook of Clinical Pharmacokinetics*. New York: ADIS Health Science Press.

Goodman, L. S., Brunton, L. L., Blumenthal, D. K., Murri, N., et al. (2011). *Goodman & Gilman's the Pharmacological Basis of Therapeutics*. New York: McGraw-Hill Medical.

Hendeles, L. & Weinberger M. (1981). Theophylline: therapeutic use and serum concentration monitoring, in Individualizing drug therapy: practical applications of drug monitoring. In W.J. Taylor, A.L. Finn, (Ed.), Gross, Townsend, Frank: New York. p. 32–65.

Jusko, W. J., Thomson, A. W., Fung, J., McMaster, P., et al. (1995). Consensus document: therapeutic monitoring of tacrolimus (FK-506). *Therapeutic Drug Monitoring*, *17*(6), 606–614.

Keating, G. M., & Lyseng-Williamson, K. A. (2013). Everolimus: a guide to its use in liver transplantation. *BioDrugs*, *27*(4), 407–411.

MacDonald, A., Scarola, J., Burke, J. T., & Zimmerman, J. J. (2000). Clinical pharmacokinetics and therapeutic drug monitoring of sirolimus. *Clinical Theraphy*, *22*(Suppl. B), B101–B121.

Patsalos, P. N., Berry, D. J., Bourgeois, B. F., Cloyd, J. C., et al. (2008). Antiepileptic drugs–best practice guidelines for therapeutic drug Monitoring: a position paper by the Subcommission on Therapeutic Drug Monitoring, ILAE Commission on Therapeutic Strategies. *Epilepsia*, *49*(7), 1239–1276.

Rowland, M., Tozer, T. N., & Rowland, M. (2009). *Clinical Pharmacokinetics and Pharmacodynamics: Concepts and Applications* (4th ed.). Philadelphia: Lippincott William & Wilkins.

Rybak, M. J., Lomaestro, B. M., Rotschafer, J. C., Moellering, R. C., et al. (2009). Vancomycin therapeutic guidelines: a summary of consensus recommendations from the Infectious Diseases Society of America, the American Society of Health-System Pharmacists, and the Society of Infectious Diseases Pharmacists. *Clinical Infectious Diseases*, *49*(3), 325–327.

Schentag, J. J. (1980). Aminoglycosides. In W. E. Evans, J. J. Schentag, & W. J. Jusko (Eds.), *Applied Pharmacokinetics: Principles of Therapeutic Drug Monitoring*. San Francisco: Applied Therapeutics.

Slattery, J. T., Clift, R. A., Buckner, C. D., Radich, J., et al. (1997). Marrow transplantation for chronic myeloid leukemia: the influence of plasma busulfan levels on the outcome of transplantation. *Blood*, *89*(8), 3055–3060.

Staatz, C., & Tett, S. (2007). Clinical pharmacokinetics and pharmacodynamics of mycophenolate in solid organ transplant recipients. *Clinical Pharmacokinetics*, *46*(1), 13–58.

Terra, S. G., Washam, J. B., Dunham, G. D., & Gattis, W. A. (1999). Therapeutic range of digoxin's efficacy in heart failure: what is the evidence? *Pharmacotherapy*, *19*(10), 1123–1126.

Touw, D. J., Neef, C., Thomson, A. H., Vinks, A. A., et al. (2005). Cost-effectiveness of therapeutic drug monitoring: a systematic review. *Therapeutic Drug Monitoring*, *27*(1), 10–17.

Wright, J. D., Boudinot, F. D., & Ujhelyi, M. R. (1996). Measurement and analysis of unbound drug concentrations. *Clinical Pharmacokinetics*, *30*(6), 445–462.

31

Clinical Toxicology

*Loralie Langman, Laura K. Bechtel, Christopher P. Holstege**

OBJECTIVES

1. Define the following terms:
   - Anion gap
   - Confirmatory test
   - Osmol gap (OSMg)
   - Screening test
   - Spot test
   - Sympathomimetic
   - Toxicology
   - Toxidrome
2. State and explain the formulas for calculating anion gap (AG) and OSMg; given appropriate information, calculate and interpret the results of these formulas, and list possible interferences with these calculated measurements.
3. For each of the following cellular hypoxia causative agents, describe the agent, the toxic effects, the treatment, and the method(s) of analysis, including specimen requirements and laboratory results, for each:
   - Carbon monoxide (CO)
   - Cyanide
   - Methemoglobin-forming agents
4. For each of the following alcohols, describe the pharmacologic action, the metabolites, the antidote, and the method(s) of analysis, including specimen requirements and laboratory results, for each:
   - Ethanol
   - Ethylene glycol
   - Isopropanol
   - Methanol
5. For each of the following analgesics, describe the toxic effects of overdose, the dose and time of symptom onset, the metabolizing enzyme and metabolite, the antidote, and the method(s) of analysis, including specimen requirements and laboratory results, for each:
   - Acetaminophen
   - Salicylate
6. For the following cholinergic/anticholinergic toxidrome agents, describe the clinical uses (if any), the mechanism of action, the metabolites, the methods of analysis, and the laboratory results for each:
   - Antihistamines
   - Antimuscarinics
   - Antipsychotics
   - Organophosphates/carbamates
   - Tricyclic antidepressants (TCAs)
7. For the following drugs, describe the manifestations of intoxication, the metabolizing enzyme and metabolite, the pharmacologic response, clinical uses if any, the antidote, and the screening tests and confirmatory methods for each:
   - Amphetamines/methamphetamines
   - Barbiturates
   - Benzodiazepines (BZs)
   - Cannabinoid
   - Cocaine
   - Marijuana
   - Opioids
   - Phencyclidine (PCP)
   - Synthetic opioids
8. List and describe five drugs used in drug-facilitated sexual assault (DFSA), including drug classification, pharmacologic effects, and methods of analysis.
9. For the following specimen types, state the advantages and disadvantages of each when used for drug analysis, and list a method of analysis for each:
   - Hair
   - Meconium
   - Saliva (oral fluid)
   - Sweat
10. Discuss the strengths and weaknesses, as well as the uses, of screening assays compared with confirmatory assays.

KEY WORDS AND DEFINITIONS

Amphetamine A sympathomimetic amine that has a stimulating effect on the central and peripheral nervous systems

Analgesics Agents that relieve pain without causing loss of consciousness

Analyte Substance or chemical constituent that is of interest in an analytical procedure

Anticholinergic toxidrome The signs and symptoms characteristic of poisoning caused by an anticholinergic agent (a substance that blocks the neurotransmitter acetylcholine in the central and peripheral nervous systems)

*A more extensive review of the subject and additional references are found in: Langman L.J., Bechtel L., Meier B.M., & Holstege C.P. (2018). Clinical toxicology. In N. Rifai, A.R. Horvath, C.T. Wittwer (Eds.), *Tietz Textbook of Clinical Chemistry and Molecular Diagnostics* (pp. 832–887). St Louis, Elsevier.

Antihistamines Antagonists of the H_1 or H_2 histamine receptors that are used to treat allergic reactions or gastric hyperacidity

Barbiturate Any of a class of sedative-hypnotic agents derived from barbituric acid or thiobarbituric acid and classified into long-, intermediate-, short-, and ultrashort-acting classes

Benzodiazepines Any of a group of minor tranquilizers that have a common molecular structure and similar pharmacologic activity, including antianxiety, sedative, hypnotic, amnestic, anticonvulsant, and muscle-relaxing effects

Cholinergic toxidrome A toxidrome that represents the acute phase of cholinesterase inhibitor poisoning

Clinical toxicology A subdivision of toxicology that involves the analysis of drugs, heavy metals, and other chemical agents in body fluids and tissues for the purpose of patient care

Confirmatory testing As used in programs that test for drugs of abuse, confirmatory tests are used to confirm a positive or sometimes negative result that had been presumptively classified as positive for a specific drug. Examples of techniques used for confirmatory testing include gas chromatography-mass spectrometry (GC-MS) and high-performance liquid chromatography (HPLC)-tandem mass spectrometry (LC-MS/MS)

Drug of abuse A drug that is repeatedly and deliberately used in a way other than prescribed or socially sanctioned (i.e., a drug that is taken for nonmedicinal reasons)

Enantiomer A molecule that exhibits stereoisomerism through the presence of one or more chiral centers (i.e., stereoisomers that are nonsuperimposable mirror images)

Ethylene glycol An ethylene compound with two hydroxy groups located on adjacent carbons. It is a common ingredient in antifreeze and is highly toxic if ingested

Gamma-hydroxybutyrate (GHB) A potent sedative, hypnotic, euphorigenic agent that is illicitly ingested for its pleasurable effects

Half-life ($t_{½}$) The length of time it takes for one-half of an administered drug to decrease due to biologic processes

Intoxication A state of impaired mental or physical functioning resulting from ingestion of alcohol or drug

Levorotary or (–) rotation A counterclockwise rotation of plane polarized light by a stereoisomer (e.g., L- or [l(–)]-methamphetamine)

Lysergic acid diethylamide (LSD) A derivative of an alkaloid found in certain fungi that has hallucinogenic properties

Marijuana A crude preparation of the leaves and flowering tops of (male or female plants) *Cannabis sativa*, usually used in cigarettes and inhaled as smoke for its euphoric properties

Medication half-life The time required for the plasma concentration of a drug to reach half of its original concentration

Miosis Constriction of the pupil of the eye, resulting from a normal response to an increase in light or caused by certain drugs or pathologic conditions

Mnemonics A mnemonic or mnemonic device is any learning technique that aids information retention, for example, an invented combination of letters (acronym) with each letter acting as a cue to an idea you need to remember (a free, nonprofit, online searchable database of medical mnemonics to help you remember the important details is available: http://www.medicalmnemonics.com/)

Opiate/opioid Opiate refers to any of a group of naturally occurring (poppy plant) or semisynthetic narcotic alkaloids with pharmacologic actions and chemical structure similar to morphine. Opioid is a general term that is applied to all substances with morphine-like properties, regardless of origin or chemical structure

Poison Any substance that, when relatively small amounts are ingested, inhaled, or absorbed, or when it is applied to, injected into, or developed within the body, has chemical action that may cause damage to structure or disturbance of function, producing symptoms, illness, or death

Screening test An initial test, such as that used to "screen" specimens to eliminate "negative" ones from further consideration and to identify presumptively positive specimens that then require confirmatory testing

Sedative-hypnotic drug A sedative that depresses activity of the central nervous system (CNS) and reduces anxiety and induces sleep

Toxidrome A syndrome caused by a dangerous concentration of toxins in the body

Toxin A poisonous substance that is produced by living cells or organisms and is capable of causing disease when introduced into the body tissues and is often capable of inducing neutralizing antibodies or antitoxins

---

Toxicology is a broad, multidisciplinary science whose goal is to determine the effects of chemical agents on living systems. Innumerable potential **toxins** inflict harm, including pharmaceuticals, herbals, household products, environmental agents, occupational chemicals, drugs of abuse, drugs used in sexual attacks, and drugs used by athletes for their performance-enhancing effects. Each year, millions of human exposure cases are reported to the American Association of Poison Control Centers. The Centers for Disease Control and Prevention has reported that **poisoning** is one

of the top 10 causes of injury-related death in the United States in all adult age groups.

This chapter provides a general overview of **clinical toxicology** and the laboratory services necessary to support the care of poisoned patients. Because a comprehensive discussion of all aspects of toxicology is beyond the scope of this chapter, the clinical significance and toxicity of only a select number of common drugs, drugs of abuse, and other chemicals are discussed.

Basic Information

In practice, it is neither possible nor necessary to test for the thousands of clinically significant toxins that may be encountered. In reality, up to 24 drugs or agents account for 80% or more of cases of **intoxication** treated in most emergency departments. The scope of clinical toxicology testing provided by the laboratory depends on the pattern of local drug use and on the available resources of the institution.

The value of drug and/or substance testing (screening) is well established (1) in the workplace, (2) for some athletic competitions, (3) to identify drug use, (4) to evaluate drug exposure and/or withdrawal in newborns, (5) to monitor patients in pain management and drug abuse treatment programs, and (6) to aid in the prompt diagnosis of toxicity for which a specific antidote or treatment modality is required (Table 31.1). In many other instances of drug toxicity, the value of drug screening, especially on an emergency basis, is controversial.

Clinical Considerations

To operate effectively, the laboratory should be closely associated with the health care team that is directly managing the patient. Through close and collaborative work, clinical information provided helps to guide appropriate ordering of tests and to ensure that interpretation of results is complete and accurate. For example, the team caring for the patient should provide the following information with the laboratory request:

1. The time and date of the suspected exposure, along with the time and date of sample collection.
2. History obtained from the patient or from witnesses that might aid in identification of the toxin
3. Assessment of the physical state of the patient at the time of presentation.

Such information is useful for guiding test selection and interpretation of results.

Analytical Considerations

Because of the wide variety of drugs, no single analytical technique is adequate for broad-spectrum drug detection. Therefore, several analytical approaches in combination are generally required. Other critical issues include speed of analysis, turnaround time, and availability of services. A drug analysis that requires several hours to complete or that is not available during all hours may be of little value in a clinical emergency. Alternatively, a rapid test that may provide false information could result in erroneous diagnostic and therapeutic decisions.

TABLE 31.1 Antidote or Specific Treatment for Intoxication

Toxin	Potential Antidote/Treatment
Acetaminophen	*N*-Acetylcysteine
Aluminum	Deferoxamine
Anticholinergics	Physostigmine
Arsenic	DMSA, BAL
Barbiturates	Multiple-dose oral activated charcoal, sodium bicarbonate
Benzodiazepines	Flumazenil
β-Adrenergic blockers	Calcium, glucagon, high-dose insulin
Calcium channel blockers	Calcium, glucagon, high-dose insulin
Carbamates	Atropine
Carbamazepine	Multiple-dose oral activated charcoal, extracorporeal techniques
Cardiac glycosides	Antidigoxin Fab fragments
Carbon monoxide	Oxygen (normobaric or hyperbaric)
Copper	d-Penicillamine
Cyanide	Hydroxocobalamin, nitrites, thiosulfate
Ethylene glycol	Fomepizole (4-methylpyrazole), ethanol, hemodialysis
Heparin	Protamine sulfate
Iron	Desferoxamine
Isoniazid	Pyridoxine
Lead	EDTA, BAL, DMSA, d-penicillamine
Mercury	BAL, DMSA, d-penicillamine
Methanol	Fomepizole (4-methylpyrazole) or ethanol, hemodialysis
Methemoglobin	Methylene blue, vitamin C
Methotrexate	Leucovorin
Opioids	Naloxone
Organophosphates	Atropine, pralidoxime
Sulfonylureas	Glucose, octreotide
Salicylates	Sodium bicarbonate, hemodialysis
tricyclic antidepressants	Sodium bicarbonate
Theophylline	Multiple-dose oral activated charcoal, extracorporeal techniques
tPA, streptokinase, urokinase	Aminocaproic acid
Warfarin	Phylloquinone (vitamin K1), plasma

BAL, Dimercaprol; *DMSA,* 2,3-dimercaptosuccinic acid.

Clinical Evaluation

When a health care team initially evaluates a patient who presents with a potentially toxicologically induced problem, the final diagnosis is often determined by reviewing the history, performing a directed physical examination, using ancillary tests (e.g., electrocardiogram [ECG], radiology), and applying a rational approach to laboratory testing.

Toxic Syndromes

Toxic syndromes (**toxidromes**) are clinical syndromes that are essential for the successful recognition of poisoning patterns and are a constellation of clinical signs and symptoms that suggests a specific class of poisoning. The most

TABLE 31.2 Symptoms of the Important Toxidromes

Toxidrome	Symptom
Anticholinergic	Agitation
	Blurred vision
	Decreased bowel sounds
	Dry skin
	Flushing
	Hallucinations
	Hyperthermia
	Ileus
	Lethargy/coma
	Mumbling speech
	Mydriasis
	Psychosis
	Tachycardia
	Urinary retention
Cholinergic	Diarrhea
	Diaphoresis
	Urination
	Miosis
	Bradycardia
	Bronchorrhea
	Emesis
	Lacrimation
	Salivation
Opioid	Bradycardia
	Decreased bowel sounds
	Hypotension
	Hypothermia
	Lethargy/coma
	Miosis
	Shallow respirations
	Slow respiratory rate
Sedative-hypnotic	Ataxia
	Blurred vision
	Confusion
	Diplopia
	Dysesthesias
	Hypotension
	Lethargy/coma
	Nystagmus
	Respiratory depression
	Slurred speech
Sympathomimetic	Agitation
	Diaphoresis
	Excessive motor activity
	Excessive speech
	Hallucinations
	Hypertension
	Hyperthermia
	Insomnia
	Restlessness
	Tachycardia
	Tremor

commonly encountered toxidromes include anticholinergic, cholinergic, opioid, sedative-hypnotic, and sympathomimetic (Table 31.2). Many toxidromes have overlapping features. For example, anticholinergic findings are highly similar to sympathomimetic findings, with an exception being effects on sweat glands. In this exception, anticholinergic agents produce warm, flushed dry skin, but sympathomimetic agents produce diaphoresis. Toxidrome findings may also be affected by individual variability, comorbid conditions, and coingestants. For example, tachycardia associated with sympathomimetic or anticholinergic toxidromes may be absent in a patient who is concurrently taking β-antagonist medications. In addition, although toxidromes may be applied to classes of drugs, one or more toxidrome findings may be absent for some individual agents within these classes. For instance, meperidine is an opioid **analgesic**, but it does not induce **miosis**, which helps to define the "classic" opioid toxidrome. When accurately identified, the toxidrome may provide invaluable information for diagnosis and subsequent treatment.

SCREENING PROCEDURES FOR DETECTION OF DRUGS

Screening procedures are designed for the relatively rapid and generally qualitative detection of drugs or other toxic substances. Screening procedures may be designed to detect a particular drug or drug class. In general, **screening tests** have adequate clinical sensitivity but lack specificity. Thus a negative result may rule out with reasonable certainty the presence of clinically significant concentrations of a particular drug but not necessarily all drugs within a class. Because of possible interferences, a positive result should be considered "presumptive positive" and should be confirmed by an alternate procedure of greater specificity. Screening tests frequently include simple visual color tests (spot tests) and immunoassays.

Spot Tests

Spot tests are less frequently used currently because some have been largely replaced by rapid immunoassays that may be performed at the point of care (POC) or in the central laboratory. For a more comprehensive description, refer to previous editions of this textbook.

Electrocardiogram

Because numerous drugs cause changes in the ECG, interpretation of the ECG in the poisoned patient has been known to significantly facilitate appropriate laboratory testing, diagnosis, and management of care.

Anion Gap

A basic metabolic panel is recommended and is an important initial screening test for all poisoned patients. When low serum bicarbonate is discovered on a metabolic panel, the clinician should determine whether an elevated anion gap (AG) exists. The formula most commonly used for the AG calculation is as follows:

$$AG = [Na^+] - [Cl^- + HCO_3^-]$$

The reference interval for this AG is accepted to be 8 to 16 mmol/L, but it has been suggested that because of changes in

the technique used to measure chloride, these limits could be lowered to 6 to 14 mmol/L.

An increase in the AG, accompanied by a metabolic acidosis, represents an increase in unmeasured endogenous (e.g., lactate) or exogenous (e.g., salicylates) anions. Clinically useful **mnemonics** for causes of high AG metabolic acidosis are the classic MUDPILES (representing Methanol, Uremia, Diabetes, Paraldehyde, Iron, Isoniazid, Inhalants, Lactate, Ethylene glycol, and Salicylate) and the more recently proposed GOLD MARK (Glycols [ethylene and propylene], Oxoproline, L-lactate, D-lactate, Methanol, Aspirin, Renal failure, and Ketoacidosis).

It is imperative that clinicians who admit poisoned patients initially presenting with an increased AG metabolic acidosis investigate the cause of that acidosis. Many symptomatic poisoned patients may have an initial mild metabolic acidosis upon presentation caused by processes resulting in elevated serum lactate (e.g., transient hypoxia or hypovolemia). However, with adequate supportive care (e.g., oxygenation and hydration), the AG acidosis should steadily improve. If, despite adequate supportive care, an AG metabolic acidosis worsens in a poisoned patient, the clinician should consider continued absorption of exogenous acids (e.g., salicylate), formation of acidic metabolites (e.g., ethylene glycol, methanol, toluene metabolites), and cellular ischemia with worsening lactic acidosis (e.g., cyanide) as potential causes.

Osmolal Gap

The main osmotically active constituents of serum are Na^+, Cl^-, HCO_3^-, glucose, and urea. Several empirical formulas based on measurement of these substances have been used to estimate serum osmolality. In practice, one has not shown itself to be superior to the others, yet each equation demonstrates significant differences in the osmolal gap reference interval. Therefore reference intervals must be validated on appropriate patient populations. Two commonly used formulas (in conventional and SI units) are presented here:

$$\text{OSMc (mOsm/kg)} = 2\ \text{Na (mmol/L)} + \text{glucose (mg/dL)}/18 + \text{urea (mg/dL)}/2.8$$

$$\text{OSMc (mOsm/kg)} = 2\ \text{Na (mmol/L)} + \text{glucose (mmol/L)} + \text{urea (mmol/L)}$$

or

$$\text{OSMc (mOsm/kg)} = 1.86\ \text{Na (mmol/L)} + \text{glucose (mg/dL)}/18 + \text{urea (mg/dL)}/2.8 + 9$$

$$\text{OSMc (mOsm/kg)} = 1.86\ \text{Na (mmol/L)} + \text{glucose (mmol/L)} + \text{urea (mmol/L)} + 9$$

The difference between the actual osmolality (OSMm), measured by freezing-point depression, and the calculated osmolality (OSMc) is referred to as delta-osmolality, or the osmol gap (OSMg).

$$\text{OSMg} = \text{OSMm} - \text{OSMc}$$

TABLE 31.3 Laboratory Findings Characteristic of Ingestion of Alcohols

Alcohol	Serum Osmolal Gap	Metabolic Acidosis With Anion Gap	Serum Acetone	Urine Oxalate
Ethanol	+	–	–	–
Methanol	+	+	–	–
Isopropanol	+	–	+	–
Ethylene glycol	+	+	–	+

Elevated OSMg implies the presence of unmeasured osmotically active substances. Compounds that increase serum osmolality when present in significant concentrations include volatile alcohols such as ethanol, methanol, isopropanol, acetone, and ethylene glycol. The calculation of OSMg is commonly used as a screen. However, it is important to remember that volatile alcohols are not detected when osmolality is measured with a vapor pressure osmometer. Therefore, for the purpose of determining the OSMg, only osmolality measurements based on freezing-point depression are acceptable.

What constitutes a normal osmolal gap is widely debated. Large variability is seen in the normal population. It would be expected that each 100 mg/dL (21.7 mmol/L) of ethanol (molecular weight = 46.068 g/mol) in serum results in an approximate increase of 21.7 mOsm/kg. However, this is not found to be the case. Applying a correction factor of 0.83 to the ethanol value will more closely approximate the contribution of ethanol to the OSMg. However, it has been observed that ethanol does not follow a completely predictable relationship with OSMg. In severe ethanol intoxication, OSMg is increased with increasing concentration, implying that something is present besides the alcohol.

A significant residual osmolal gap (>10 mOsm/kg) after correction for ethanol suggests the possible presence of one of the other volatile alcohols. This information, in conjunction with the presence or absence of metabolic acidosis or serum acetone, is helpful to the clinician when specific measurements of alcohols other than ethanol are not available on an emergency basis (Table 31.3). It must be realized that ketones and substances administered to patients such as polyethylene glycol, mannitol (osmotic diuretic), and propylene glycol (solvent for diazepam and phenytoin) may increase serum osmolality.

Immunoassay

Different types of immunoassays are useful in screening specimens for drugs. In some cases, these assays are relatively specific for a single drug (**lysergic acid diethylamide [LSD]**) but in others can detect several drugs within a class are detected (e.g., opiates). The detection limit for various

members of a class of drugs or the degree of cross-reactivity for similar drugs varies, and each manufacturer of immunoassay reagents should be consulted for specific information. These assays are easy to perform; many are available for use on automated instrumentation and may be able to provide "semiquantitative" results. For the vast majority of drugs of abuse, immunoassays are the methods of choice for initial screening. However, for more comprehensive drug screening, chromatographic procedures complement immunoassays.

Point of Care

Numerous POC drug test devices for urine (and oral fluid) are designed for easy, rugged, and portable use by nontechnical personnel. Although these devices are relatively simple to use, proper training of nonlaboratory users is important for optimal performance. These non–instrument-based immunoassay test devices are designed for use at the site of collection; results are available within minutes and are variously configured to detect only one drug or many drugs simultaneously. A more detailed description of these methods is found in the package insert for each specific test kit.

Gas Chromatography

Gas chromatography (GC) is relatively rapid, capable of resolving a broad spectrum of drugs, and widely used for qualitative and quantitative drug analysis. Common detectors for drug detection by GC are flame ionization and alkali flame ionization (nitrogen phosphorus [NP]) detectors. GC coupled to a mass spectrometer (GC-MS) provides an analytical system with the greatest accuracy of identification. Numerous methods for general drug screening by GC-MS spectrometry have been published

High-Performance Liquid Chromatography

The resolving power of high-performance liquid chromatography (HPLC) in separating widely divergent chemical constituents has been applied to the complex challenge of comprehensive drug screening in biologic fluids. Advantages of HPLC over GC include its usefulness in analyzing polar compounds without derivatization (e.g., morphine, benzoylecgonine [BE]) and thermally labile drugs (e.g., chlordiazepoxide). The advent of diode array detectors that provide a spectral scan of compounds as they elute from the column has greatly increased the discriminatory power of this technique. Analytical systems based on coupling of liquid chromatography (LC) with mass spectrometry (MS) (LC-MS or LC-MS/MS) significantly improve resolution of drugs from complex matrices.

PHARMACOLOGY AND ANALYSIS OF SPECIFIC DRUGS AND TOXIC AGENTS

The toxic consequences of several drugs and toxins are individually discussed in this section. They have been grouped into the following categories: agents that cause cellular hypoxia, alcohols, analgesics (nonprescription), agents related to anticholinergic toxidrome, drugs of abuse, and drug-facilitated crimes (DFCs).

AGENTS THAT CAUSE CELLULAR HYPOXIA

Carbon monoxide (CO) and methemoglobin-forming agents interfere with oxygen transport, resulting in cellular hypoxia. Cyanide interferes with oxygen use and therefore causes an apparent cellular hypoxia.

Carbon Monoxide

CO is a colorless, odorless, tasteless gas that is a product of incomplete combustion of carbonaceous material. Small amounts of CO are produced endogenously in the metabolic conversion of heme to biliverdin.

When inhaled, CO combines tightly with the heme iron (Fe^{2+}) of hemoglobin to form carboxyhemoglobin. The binding affinity of hemoglobin for CO is approximately 250 times greater than that for oxygen. Moreover, the binding of CO to a hemoglobin subunit increases the oxygen affinity for the remaining subunits in the hemoglobin tetramer. Thus, at a given tissue PO_2 value, less oxygen dissociates from hemoglobin when CO is also bound, shifting the hemoglobin-oxygen dissociation curve to the left. Consequently, CO not only decreases the oxygen content of blood, it also decreases oxygen availability to tissue. This produces a greater degree of tissue hypoxia than would result from an equivalent reduction in oxyhemoglobin due to hypoxia alone. CO may also bind to other heme proteins, such as myoglobin and mitochondrial cytochrome oxidase a_3; this may limit oxygen use when tissue PO_2 is very low.

Treatment for CO poisoning involves removal of the individual from the contaminated area and administration of oxygen. The **half-life ($t½$)** of carboxyhemoglobin in the body is variable. Low carboxyhemoglobin concentrations relative to the severity of poisoning may be observed if the patient was removed from the CO-contaminated environment several hours before blood sampling. Hyperbaric oxygen therapy for CO is highly debated, and current position papers have reported no evidence to support its use.

Analytical Methods

Carbon monoxide may be released from hemoglobin and then measured by GC, or it may be determined indirectly as carboxyhemoglobin by spectrophotometry. GC methods are accurate and precise even for very low concentrations of CO. Spectrophotometric methods are rapid, convenient, accurate, and precise, except at very low concentrations of carboxyhemoglobin (<3%).

GC methods measure the CO content of blood and are considered to be reference procedures. When blood is treated with potassium ferricyanide, carboxyhemoglobin is converted to methemoglobin and CO released into the gas phase. Measurement of the released CO may be performed by GC, using various methods and detectors. In clinical practice, CO binding capacity is determined after an aliquot of the blood specimen is treated with CO to saturate the hemoglobin. The results are then expressed as percentage of carboxyhemoglobin:

$$\% \text{ HbCO} = \frac{\text{CO}_{\text{content}}}{\text{CO}_{\text{capacity}}} \times 100$$

TABLE 31.4 Examples of Acquired Causes of Methemoglobinemia

Local anesthetics	Analgesics	Miscellaneous
Benzocaine	Phenazopyridine	Aminophenol
Lidocaine	Phenacetin	Aniline, *p*-chloroaniline
Prilocaine	**Nitrites and nitrates**	Bromates
Antimicrobials	Ammonium nitrate	Chlorates
Chloroquine	Amyl nitrite	4-Dimethyl-amino-phenolate (4-DMAP)
Dapsone	Butyl nitrite	Metoclopramide
Primaquine	Isobutyl nitrite	Nitrobenzene
Sulfonamides	Potassium nitrate	Nitroethane
Trimethoprim	Sodium nitrate	Nitroglycerin
	Nitrogen oxides	Phenazopyridine
	Nitric oxide	Potassium permanganate
	Nitrogen dioxide	Propanil

Spectrophotometric methods rely on the characteristic spectral absorption properties of carboxyhemoglobin. The most common are based on automated, multiwavelength measurements of several hemoglobin species, and they are rapid and convenient. Spectrophotometric methods generally compare favorably with GC procedures at carboxyhemoglobin concentrations greater than 2% to 3%, but their precision is poor below these concentrations. Therefore they are sufficiently accurate and precise for measurement of CO after exogenous exposure, but they are too insensitive to detect the increased endogenous production of CO that occurs in hemolytic anemia.

Fetal hemoglobin and adult hemoglobin differ slightly in their spectral properties. Consequently, falsely high carboxyhemoglobin values of 4% to 7% may occur when blood from neonates is measured by some spectrophotometric methods by using fewer wavelengths. Moreover, erroneous results may occur with lipemic and icteric specimens and in the presence of methylene blue (see later section, "Methemoglobin-Forming Agents").

Methemoglobin-Forming Agents

The heme iron in hemoglobin is normally present in the ferrous state (Fe^{2+}). When oxidized to the ferric state (Fe^{3+}), methemoglobin is formed, and this form of hemoglobin cannot bind oxygen. Congenital methemoglobinemia may result from a deficiency of NADH-methemoglobin reductase or, more rarely, from hemoglobin variants (hemoglobin M) in which heme iron is both more susceptible to oxidation and more resistant to reduction by the methemoglobin reductase system.

An acquired (toxic) methemoglobinemia may be caused by various drugs and chemicals (Table 31.4). The normal percentage of methemoglobin is less than 1.5% of total hemoglobin. All symptoms associated with methemoglobinemia are consequences of hypoxia associated both with diminished O_2 content of blood and with decreased O_2 dissociation from hemoglobin species. The PO_2 is normal in these patients and therefore so is the calculated hemoglobin oxygen saturation. Thus a normal PO_2 in a cyanotic patient is a significant indication of the possible presence of methemoglobinemia; therefore measurement of methemoglobin is important in these cases. Specific therapy for toxic methemoglobinemia involves the administration of methylene blue. Methylene blue and sulfhemoglobin cause spectral interference in the measurement of methemoglobin with some CO-oximeters but not with the Malloy-Evelyn method.

Analytical Methods

Methemoglobin may be measured using the manual spectrophotometric method of Evelyn and Malloy or by performing automated multiwavelength measurements with a CO-oximeter. Methemoglobin interferes with the noninvasive pulse oximetry method, measuring the absorbance of light at 660 nm (oxyhemoglobin) and 940 nm (deoxyhemoglobin). Because methemoglobin is not stable at room temperature, specimens should be kept on ice or refrigerated but not frozen. The stability of methemoglobin at 4°C has not been well studied. Some sources report significant decreases in methemoglobin concentration after 4 to 8 hours, whereas others describe little or no change after 24 hours. Freezing results in an increase in methemoglobin concentration.

Cyanide

Cyanide consists of one atom of carbon triple-bonded to one atom of nitrogen (C≡N). Inorganic cyanides (also known as cyanide salts) contain cyanide in anion form (CN^-) and are used in numerous industries. Organic compounds that have a cyano group bonded to an alkyl residue are called nitriles. Iatrogenic cyanide poisoning may occur during use of nitroprusside.

Cyanide in serum readily crosses all biologic membranes and avidly binds to heme iron (Fe^{3+}) in the cytochrome oxidase-a3 complex within mitochondria. When bound to cytochrome oxidase-a3, cyanide is a competitive inhibitor that causes decoupling of oxidative phosphorylation. Patients exposed to toxic concentrations of cyanide exhibit a wide variety of clinical effects depending on the poisoning severity. Hydroxycobalamin or the cyanide antidote kit should be administered as soon as cyanide poisoning is suspected.

Analytical Methods

After microdiffusion, whole blood CN^- is measured by spectrophotometry or by headspace GC.

ALCOHOLS OF TOXICOLOGIC INTEREST

Several alcohols are toxic and medically important; they include ethanol, methanol, isopropanol, acetone (also a metabolite of isopropanol), and ethylene glycol.

Ethanol

Ethanol is widely available and often abused chemical substance. Consequently, measurement of ethanol is one of the

TABLE 31.5 Stages of Acute Alcoholic Influence/Intoxication

Blood Alcohol Concentration (g/dL)	Influence	Clinical Signs/Symptoms
0.01–0.05	Subclinical	Influence/effects not apparent or obvious
		Behavior appears normal
		Impairment detectable by special tests
0.03–0.12	Euphoria	Mild euphoria
		increased sociability, talkativeness, self-confidence
		decreased inhibitions
		mild sensorimotor impairment
		Slowed information processing
		Loss of efficiency in finer performance tests
		Impairment of perception and memory
0.09–0.25	Excitement	Emotional instability; loss of critical judgment comprehension
		Decreased sensory response; increased reaction time
		Reduced visual acuity, peripheral vision, and glare recovery
		Sensorimotor incoordination; impaired balance
		Drowsiness
0.18–0.30	Confusion	Disorientation, mental confusion; dizziness
		Exaggerated emotional states (fear, rage, grief, etc.)
		Disturbances of vision (diplopia, etc.) and of perception of color, form, motion, dimensions
		Increased pain threshold
		Ataxia; dysarthria, apathy, lethargy
0.25–0.40	Stupor	General inertia; approaching loss of motor functions
		Markedly decreased response to stimuli
		Marked muscular incoordination; inability to stand or walk
		Vomiting; incontinence of urine and feces
		Impaired consciousness; sleep or stupor
0.35–0.50	Coma	Unconsciousness; coma; anesthesia
		Depressed or absent reflexes
		Subnormal temperature
		Impairment of circulation and respiration
		Possible death
0.45+	Death	Possible death from respiratory arrest

Modified from Dubowski, K. M. (1980). Alcohol determination in the clinical laboratory. *American Journal of Clinical Pathology, 74*, 747–504.

more frequently performed tests in the toxicology laboratory. The principal pharmacologic action of ethanol involves central nervous system (CNS) depression, and effects vary depending on the blood ethanol concentration (Table 31.5) and an individual's tolerance. A blood alcohol concentration of 80 mg/dL (0.08%) has been established as the per se limit for operation of a motor vehicle in many countries.

When consumed with other CNS depressant drugs, ethanol exerts a potentiation or synergistic depressant effect. This occurs at relatively low alcohol concentrations, and numerous deaths have resulted from combined ethanol and drug ingestion. Ethanol is also a teratogen, and alcohol consumption during pregnancy can result in a newborn with fetal alcohol spectrum disorder (FASD). FASD is an umbrella term that describes the variety of effects that can occur in an individual whose mother drank alcohol during pregnancy (http://www.nofas.org). These effects may include physical, mental, behavioral, and/or learning disabilities with possible lifelong implications and are 100% preventable when a woman completely abstains from alcohol during her pregnancy.

Ethanol is metabolized principally by liver alcohol dehydrogenase (ADH) to acetaldehyde, which is subsequently oxidized to acetic acid by aldehyde dehydrogenase (Fig. 31.1). The rate of elimination of ethanol from blood approximates a zero-order process. This rate varies among individuals, averaging approximately 15 mg/dL per hour for males and 18 mg/dL per hour for females. At both low (<20 mg/dL) and high (>300 mg/dL) ethanol concentrations, elimination becomes more nearly first order; it is accelerated at high

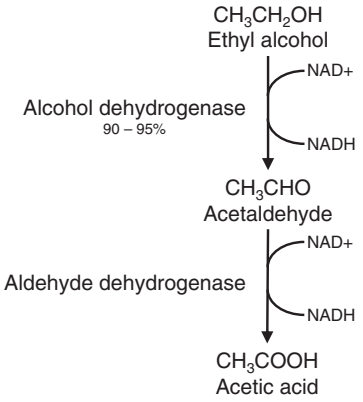


Fig. 31.1 Metabolism of ethanol.

concentrations. The elimination rate is also influenced by drinking practices (e.g., alcoholics have increased elimination rates caused by enzyme induction).

Methanol

Methanol is used as a solvent in several commercial products, as a constituent in windshield wiper fluid, copy machine fluids, fuel additives (octane boosters), paint remover or thinner, **antifreeze**, canned heating sources, deicing fluid, **shellacs**, and **varnishes**.

Methanol is oxidized by liver ADH (at approximately one-tenth the rate of ethanol) to formaldehyde. Formaldehyde in turn is rapidly oxidized by aldehyde dehydrogenase to formic acid, which may cause serious acidosis and optic neuropathy, resulting in blindness or death. The mainstay of therapy for methanol toxicity includes the administration of fomepizole or ethanol as a competitive ADH inhibitor, plus either folate or folinic acid, and hemodialysis when required.

Isopropanol and Acetone

Isopropanol is readily available to the general population as a 70% aqueous solution for use as rubbing alcohol but can also be found in cleaners, disinfectants, antifreezes, cosmetics, solvents, and inks. Isopropanol is rapidly metabolized by ADH to acetone, which is eliminated much more slowly. Therefore concentrations of acetone in serum often exceed those of isopropanol during the elimination phase after isopropanol ingestion. Acetone has CNS depressant activity similar to that of ethanol, and because of its longer half-life, it may prolong the apparent CNS effects of isopropanol. Supportive care is the mainstay of treatment, with dialysis only in severe intoxication.

Ethylene Glycol

Ethylene glycol is present in antifreeze products, deicing products, detergents, paints, and cosmetics. Ethylene glycol itself is relatively nontoxic; however, metabolism of ethylene glycol by ADH results in the formation of numerous acid metabolites, including oxalic acid and glycolic acid, which are responsible for much of the toxicity of ethylene glycol. It is difficult to define a serum ethylene glycol concentration associated with a high probability of death, and the determination of ethylene glycol and glycolic acid may provide useful information in cases of ethylene glycol ingestion. The mainstay of therapy for ethylene glycol toxicity includes administration of ethanol or fomepizole as a competitive ADH inhibitor and dialysis.

Analysis of Ethanol

Serum/Plasma and Blood Ethanol

Serum, plasma, and whole blood are suitable specimens for the determination of ethanol. Alcohol distributes into the aqueous compartments of blood; because the water content of serum is greater than that of whole blood, higher alcohol concentrations are obtained with serum as compared with whole blood. Experimentally, the serum-to-whole blood ethanol ratio is 1.18 (1.10 to 1.35) and varies slightly with hematocrit. Therefore laboratories that perform alcohol determinations should make clear the choice of specimen.

Because of the volatile nature of alcohols, specimens should be kept capped to avoid evaporative loss. Blood may be stored, when properly sealed, for 14 days at room temperature or at 4°C, with or without preservative. For longer storage or for nonsterile postmortem specimens, sodium fluoride should be used as a preservative to minimize changes in ethanol concentration.

To measure ethanol in serum/plasma, enzymatic analysis is the method of choice for many laboratories. In this method, ethanol is measured by oxidation to acetaldehyde with NAD, a reaction catalyzed by ADH. With this reaction, the formation of NADH, measured at 340 nm, is proportional to the amount of ethanol in the specimen:

$$\text{Ethanol} + \text{NAD} \xrightarrow{\text{ADH}} \text{Acetaldehyde} + \text{NADH}$$

Under most assay conditions, ADH is reasonably specific for ethanol, with interferences by isopropanol, acetone, methanol, and ethylene glycol of typically less than 1%. As a precaution, the venipuncture site is recommended to be cleansed with an alcohol-free disinfectant, such as aqueous benzalkonium chloride.

Spuriously increased results for ethanol have been described in the presence of both high concentrations of lactate dehydrogenase (LDH) and lactate. However, not all ethanol assays are known to have this interference. It is recognized that the lactate and LDH must hit some critical concentration for the interference to occur with those affected assays. Theoretically, other dehydrogenases and substrates may cause similar interference.

Breath Ethanol

Statutory laws for driving under the influence of alcohol were originally based on the concentration of ethanol in venous whole blood. Because collection of blood is invasive and requires the participation of medical personnel, determination of alcohol in expired air (breath) has long been the mainstay of evidential alcohol measurements. The fundamental principle for the use of breath analysis is that alcohol in capillary alveolar blood rapidly equilibrates with alveolar air in a ratio of approximately 2100:1 (blood/breath). To alleviate confusion and uncertainty surrounding conversion from breath to blood alcohol concentration, the traffic laws in many countries specify per se limits for blood and/or breath.

Urine Ethanol

Urine has been used as an alternative, less invasive specimen for determination of alcohol use. During the postabsorptive phase that follows alcohol ingestion, the concentration of alcohol in urine is approximately 1.3 times that in blood. However, the use of urine alcohol measurements to estimate blood concentrations is discouraged because the ratio is highly variable and, perhaps more important, because the urine alcohol concentration may better reflect an average of the blood alcohol concentration during the period in which urine is collected in the bladder.

Analysis of Volatile Alcohols

Flame ionization GC remains the most common method of detection and quantitation of volatile alcohols in biologic samples. Not only does it distinguish between ethanol, methanol, isopropanol, and acetone, it has the capability to measure concentrations as low as 10 mg/dL (0.01%).

Ethyl Glucuronide and Ethylsulfate

Ethyl glucuronide (EtG), ethyl sulfate (EtS), and phosphatidylethanols (PEth) are biomarkers of ethanol consumption. EtG and EtS are phase II metabolites of ethanol. EtG is a phase II metabolite of ethanol formed through the UDP-glucuronosyltransferase catalyzed conjugation of ethanol with glucuronic acid. EtS is formed also directly by the conjugation of ethanol with sulfate group. Because EtG has a long urinary elimination time (≤ 80 hours), its specificity for ethanol exposure, and the low detection limits, EtG is used as a marker of recent ethanol intake in a variety of clinical and legal settings, including medical monitoring for relapse, emergency department patient evaluation, postmortem assessment, and transportation accident investigation. Monitoring both EtG and EtS in urine improves sensitivity when using ethanol biomarkers for monitoring recent drinking. EtG, but not EtS, can be produced post specimen collection in diabetics due to glucose fermentation. Although upper respiratory infections, as well as β-glucuronidase hydrolysis, may lower levels of EtG, they but do not seem to affect EtS. However, challenges associated with factors such as establishing appropriate cutoff concentrations capable of distinguishing between drinking and incidental exposure such as nonbeverage sources of ethanol exposure (i.e., hand sanitizers, mouthwash), nonuniform laboratory reporting limits, sample stability, and microbial activity may complicate interpretation of results.

ANALGESICS (NONPRESCRIPTION)

Acetaminophen

Acetaminophen has analgesic and antipyretic actions. In normal doses, acetaminophen is safe and effective but may cause severe hepatic toxicity when consumed in overdose quantities. Initial clinical findings in acetaminophen toxicity are relatively mild and nonspecific and include nausea, vomiting, and abdominal discomfort. These conditions are not predictive of impending hepatic necrosis. The measurement of serum acetaminophen concentration becomes paramount for proper assessment of the severity of overdose and for appropriate decision making for antidotal therapy with *N*-acetylcysteine (NAC) using the Rumack-Matthew nomogram (Fig. 31.2).

Acetaminophen is normally metabolized in the liver to glucuronide (50% to 60%) and sulfate (≈30%) conjugates. A smaller amount (≈10%) is metabolized by a cytochrome P450 mixed-function oxidase pathway that is thought to involve formation of a highly reactive intermediate *N*-acetylbenzoquinoneimine (NAPQI). This intermediate normally undergoes conjugation with glutathione and then subsequent transformation to cysteine and mercapturic acid

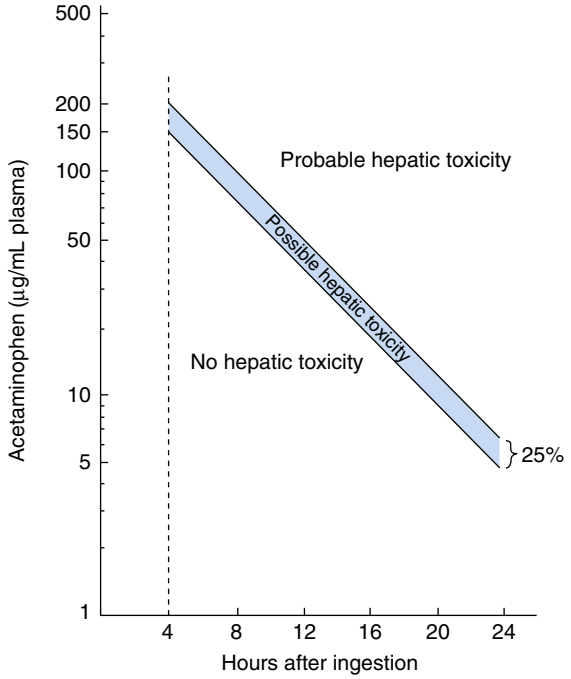


Fig. 31.2 Rumack-Matthew nomogram. (From Rumack, B. H., & Matthew, H. [1975]. Acetaminophen poisoning and toxicity. *Pediatrics, 55*, 871–876. Reproduced by permission of *Pediatrics*.)

conjugates of acetaminophen. With acetaminophen overdose, the sulfation pathway becomes saturated; consequently, a greater portion is metabolized by the P450 mixed-function oxidase pathway. When the tissue stores of glutathione become depleted, arylation of cellular molecules by the benzoquinoneimine intermediate leads to hepatic necrosis. Specific therapy for acetaminophen overdose is administration of NAC, which acts as a glutathione substitute. Antidotal therapy with NAC is most effective when administered before hepatic injury occurs, as signified by elevation of activities of aspartate transaminase (AST) and alanine transaminase (ALT).

Many spectrophotometric methods are available for determination of acetaminophen. In general, these methods are simple and relatively easy to perform but are subject to various interferences such as bilirubin or bilirubin byproducts absorbing at similar wavelengths. Some methods also measure the nontoxic metabolites and the potentially toxic parent acetaminophen, and thus may produce misleading results. Therefore only methods specific for parent acetaminophen should be used. Immunoassays are widely used for this purpose because they are rapid, easily performed, and accurate. A qualitative, one-step lateral flow immunoassay (cutoff of 25 µg/mL) may be suitable for point-of-care application, yet it has a low positive predictive value. Most chromatographic methods are very accurate and are considered reference procedures.

Salicylate

Acetylsalicylic acid (aspirin) has analgesic, antipyretic, and anti-inflammatory properties. Salicylates directly stimulate the central respiratory center and thereby cause hyperventilation and respiratory alkalosis. In addition, salicylates cause uncoupling

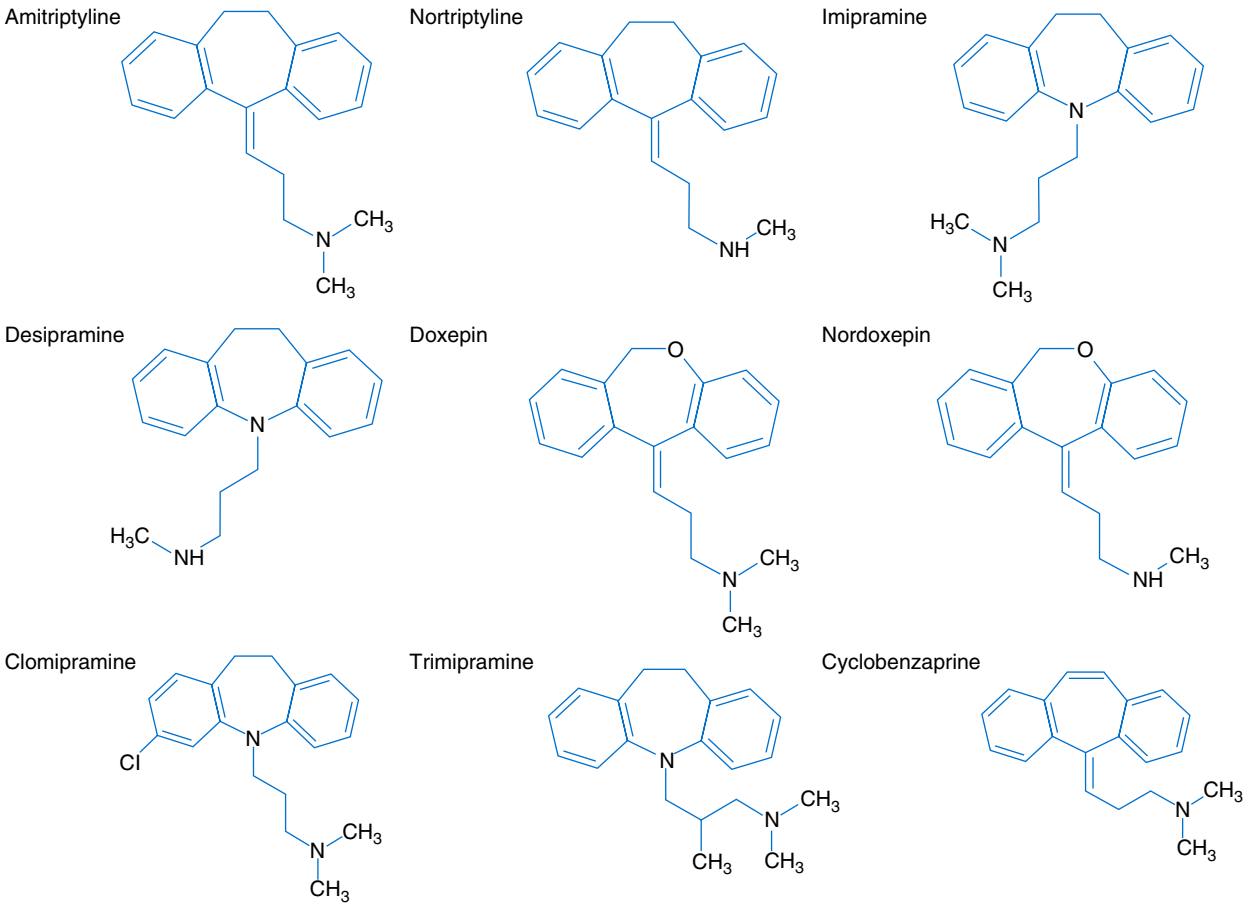


Fig. 31.3 Structure of tricyclic antidepressants *(TCA)* and related drugs.

of oxidative phosphorylation. As a result, heat production (hyperthermia), oxygen consumption, and metabolic rate may be increased. Salicylates also enhance anaerobic glycolysis but inhibit the Krebs cycle and transaminase enzymes, which leads to accumulation of organic acids and thus to metabolic acidosis.

Following acute salicylate overdose, patients initially may be asymptomatic, especially if that product is enteric coated. Patients may develop nausea, vomiting, abdominal pain, tinnitus, tachypnea, oliguria, and altered mental status ranging from agitation to lethargy to coma. Individuals with chronic intoxication present in a similar fashion as those with acute exposures, yet such exposures typically are more insidious and therefore are often misdiagnosed.

Use of salicylate concentrations to guide management must be done cautiously and only in conjunction with careful evaluation of a patient's clinical status. The absorptive phase of salicylates is often unpredictable (delayed or erratic) as a result of bezoar formation, enteric-coated product, gastric outlet obstruction, or pylorospasm. Therefore a specimen drawn soon after the original ingestion may not be reflective of the potential peak concentration; therefore monitoring of serial concentrations can be very useful. Because products containing salicylates are readily available, the clinical effects of salicylate toxicity are nonspecific, and lack of metabolic acidosis does not rule out the potential for salicylate toxicity, clinicians should have a low threshold for obtaining serum salicylate concentrations.

Classic methods for measurement of salicylate in serum are based on the method of Trinder. Despite the limitations of these methods, they continue to be used to assess salicylate overdose. The Trinder method results agree very closely with those of a reference HPLC procedure. A qualitative, one-step lateral flow immunoassay (cutoff of 100 μg/mL) is commercially available for point-of-care application but has a low positive predictive value (0.47). GC and LC methods are the most specific methods for salicylate, but their general availability, especially for emergency use, is limited and probably is not necessary.

AGENTS RELATED TO ANTICHOLINERGIC TOXIDROME

Numerous agents can induce an anticholinergic toxidrome include (e.g., tricyclic antidepressants [TCAs], antihistamines). These agents have divergent therapeutic applications but in overdose often share similar anticholinergic toxidrome.

Tricyclic Antidepressants and Related Drugs

TCAs are so named because of their three-ring structure (Fig. 31.3). They represent a class of drugs frequently prescribed for the treatment of depression. The *N*-demethylated

TABLE 31.6 Examples of Classical and Atypical Antipsychotics

Antipsychotics	Examples
Classical Antipsychotics	
Phenothiazines	Chlorpromazine
	Promethazine
	Perphenazine
	Fluphenazine
	Thioridazine
	Mesoridazine
	Trifluoperazine
Thioxanthenes	Flupentixol
	Zuclopenthixol
Dibenzoxazepine	Loxapine
Dihydroindoles	Molindone
Butyrophenones	Droperidol
	Haloperidol
Diphenylbutylpiperidines	Pimozide
Benzamides	Sulpiride
Atypical Antipsychotics	
Dibenzodiazepine derivatives	Clozapine
	Olanzapine
Dibenzothiazepine derivatives	Quetiapine
Benzisoxazole derivatives	Risperidone
Benzisothiazol piperazine derivative	Ziprasidone
Phenylindole derivatives	Sertindole

metabolites of several TCAs are pharmacologically active and contribute variably to overall pharmacodynamic activity.

The therapeutic mechanism of TCAs is the blockade of neuronal reuptake of serotonin and/or norepinephrine. However, TCAs have many other pharmacologic actions that contribute to adverse effects at both therapeutic and toxic doses, including: anticholinergic; antihistaminic; α_1-receptor blocker actions; inhibition of the GABA channel; and of cardiac potassium efflux and fast sodium channels.

Analytical Methods

TCAs are measured by chromatographic methods or by immunoassay. Immunoassays are rapid and relatively easy to perform but may be subject to interference by other drugs and they cannot necessarily be used to identify which TCA is being quantitated. In cases of overdose, qualitative identification (serum or urine) is sufficient because the severity of intoxication is more reliably indicated by an increase in the QRS interval on an ECG than by the serum concentration.

Antipsychotic Drugs

Antipsychotic drugs are generally used for numerous psychiatric disorders. In addition to their psychotherapeutic effects, certain members of this group are used for other indications. Antipsychotic compounds are traditionally divided and subdivided according to their chemical structure (Table 31.6).

The principal manifestations of phenothiazine toxicity involve the CNS and the cardiovascular system. The presentation for most of these drugs is related to their mechanisms of action, which include dopamine blockade (CNS depression), α_1-blockade (hypotension and miosis), anticholinergic, and K-efflux blockade (QT prolongation). Toxicity is strongly correlated with peak serum concentrations.

Neuroleptics are measured by chromatographic methods and by immunoassay. Chromatographically, GC methods with NP, electron capture (EC), and MS detectors are used, along with HPLC, LC-MS, and LC-MS/MS methods.

ANTIHISTAMINES

Antihistamines are medications that are popular among the general public for treatment of allergic reactions and as common sleep aids. Histamine is released from mast cells and plays an important physiological role in immediate hypersensitivity and allergic responses. Histamine functions as a neurotransmitter in the CNS and stimulates gastric acid secretion. Currently available antihistamine drugs clinically antagonize H_1 and H_2 histamine receptors. The principal H_2 receptor response is stimulation of gastric acid secretion, whereas other actions of histamine such as smooth muscle contraction, vasodilation, increased capillary permeability, pain, and itching are primarily mediated by H_1 receptors.

First-generation lipophilic antihistamines (e.g., diphenhydramine) bind H_1 receptors, exhibit peripheral and CNS effects, and bind to muscarinic receptors, resulting in anticholinergic activity. Second-generation antihistamines (e.g., fexofenadine) are highly specific for peripheral H_1 receptors and do not penetrate the CNS. Therefore second-generation H_1 receptor antagonists display minimal sedative and anticholinergic effects. Individuals overdosed with first-generation H_1 antihistamines present clinically with CNS depression or stimulation and peripheral anticholinergic effects.

Commonly available antihistamines are detected using GC-MS/MS and LC-MS/MS. However, the clinical necessity of quantitative serum antihistamine concentrations is questionable because poor correlation has been noted among patient's age, dose, blood concentration, clinical effects, and death. Antihistamines are detected by forensic laboratories as agents potentially used in facilitating sexual assault (see later section, "Drug-Facilitated Crimes").

Although detection of antihistamines is not typically clinically relevant for the acutely toxic patient, it is important to note that very high concentrations of first-generation antihistamines, such as promethazine and diphenhydramine, have been documented to cross-react with urine drugs of abuse immunomethods used for screening purposes. Therefore physicians and medical review officers (MROs) should be aware of potential false-positive results from on-site drug testing devices, as well as immunoscreening.

Antimuscarinic Agents

Besides numerous pharmaceutical agents, several plant and mushroom species contain antimuscarinic compounds (e.g.,

atropine, scopolamine) that cause anticholinergic symptoms when ingested. Testing for atropine and scopolamine is performed by LC-MS methods.

AGENTS RELATED TO CHOLINERGIC TOXIDROME

Agents inducing **cholinergic toxidrome** are diverse and act by either inactivation of cholinesterase enzymes or direct stimulation of acetylcholine receptors. Acetylcholine is an essential neurotransmitter that affects parasympathetic synapses (autonomic and CNSs), sympathetic preganglionic synapses, and the neuromuscular junction (see also prior section, "Toxic Syndromes"). The duration of acetylcholine action is controlled by acetylcholinesterase (AChE) and butyrylcholinesterase/pseudocholinesterase. AChE is found in red blood cells, nervous tissue, and skeletal muscle. Butyrylcholinesterase is found in plasma, liver, heart, pancreas, and brain.

Organophosphate and Carbamate Compounds

Organophosphate insecticides are toxic because they inactivate AChEs that are required for hydrolyzing acetylcholine at nerve junctions. This inhibition is a consequence of phosphorylation of the serine hydroxyl group at the active site of the cholinesterase catalytic triad (Ser-Glu-His). Subsequent hydrolysis results in irreversible dealkylation of AChE.

Carbamates are structurally different from organophosphates; carbamates exert their toxicity at the active site of AChE but inhibit enzyme activity through carbamylation rather than phosphorylation of the serine hydroxyl group. Most carbamates exhibit poor CNS penetration and have a shorter duration of action than organophosphates.

Excess synaptic acetylcholine stimulates muscarinic receptors (peripheral and CNSs) and stimulates but then depresses or paralyzes nicotinic receptors. Activation of peripheral muscarinic receptors causes signs and symptoms described by the mnemonics SLUDGE (Salivation, Lacrimation, Urination, Defecation, gastrointestinal [GI] distress, Emesis) or DUMBBELS (Diarrhea, Urination, Miosis, Bradycardia, and Bronchorrhea-bronchoconstriction, Emesis, Lacrimation, Sweating-salivation). Stimulation or paralysis of nicotinic receptors at the neuromuscular junction causes muscle fasciculations, cramping, weakness, and respiratory muscle paralysis. Stimulation of nicotinic receptors at sympathetic ganglia results in hypertension, tachycardia, pallor, and mydriasis. Thus the actual signs and symptoms observed with these toxins depend on the balance of muscarinic and nicotinic receptor activation.

Specific therapy for organophosphate and carbamate insecticide poisoning includes administration of atropine, to block the muscarinic (but not the nicotinic) actions of acetylcholine, and pralidoxime, to reactivate cholinesterase in organophosphate poisoning. Treatment requires immediate attention and should not rely on **confirmatory testing**. Cholinesterase activity is often monitored by clinicians in occupational settings, following acute intentional and accidental exposure, to determine response to therapy.

AChE and butyrylcholinesterase enzyme activities are typically monitored using spectrophotometric analyses. AChE activity present at nerve junctions is similar to that present in red blood cells and is an appropriate index of neurotoxicity and therefore correlates more closely with the degree of neurotoxicity. This assay is more sensitive than serum cholinesterase activity and often is used to confirm exposure and to predict enzyme reactivation during treatment.

Butyrylcholinesterase (pseudocholinesterase), also present in serum, is inhibited by these insecticides. The activity of butyrylcholinesterase declines then returns to normal more rapidly than is observed for the red cell enzyme. Serum butyrylcholinesterase is readily measured on hemolyzed samples and in clinical laboratories without isolation of red blood cells. However, interindividual variability is high; making interpretation of test results more difficult. Because butyrylcholinesterases are synthesized in the liver, this assay is particularly sensitive to interferences associated with pregnancy and liver disease.

In acute poisoning, symptoms generally begin when cholinesterase activity is inhibited by approximately 50% of the lower limits of the reference interval, and this degree of inhibition is of diagnostic value. However, the degree of cholinesterase inhibition generally does not correlate well with the clinical severity of poisoning. Interpretation of test results is made more difficult by considerable individual variability of reference activities.

DRUGS OF ABUSE

Many definitions exist for the term *drugs of abuse*. For this chapter, this term is defined as "a drug that is repeatedly and deliberately used in a way other than prescribed or socially sanctioned (i.e., a drug that is taken for nonmedicinal reasons)."

Drug use and abuse are widespread in society, and public awareness has been heightened as to their impact on public safety and on lost productivity in industry. To resolve these issues, governmental, industrial, educational, and sports agencies are increasingly requiring drug testing of prospective and existing employees, students, and participants in professional and amateur athletics. Moreover, drug abuse during pregnancy is a matter of concern, both medically and socially. Testing for drugs of abuse may be a medical requirement for organ transplantation candidates, pain management clinics, drug abuse treatment programs, and psychiatric programs. Drug testing for these purposes represents a significant activity for toxicology laboratories.

Testing for drugs of abuse usually involves testing a single urine specimen for various drugs. However, it should be noted that testing of a single random urine drug specimen detects only fairly recent drug use and it does not differentiate casual use from chronic drug abuse. The latter requires sequential drug testing and clinical evaluation. Moreover, urine drug testing (UDT) alone cannot determine the degree of impairment, the dose of drug taken, or the exact time of use. Because of these and other limitations of testing for drugs in urine, integrating the use of alternate biologic specimens for drug testing is a matter of growing interest.

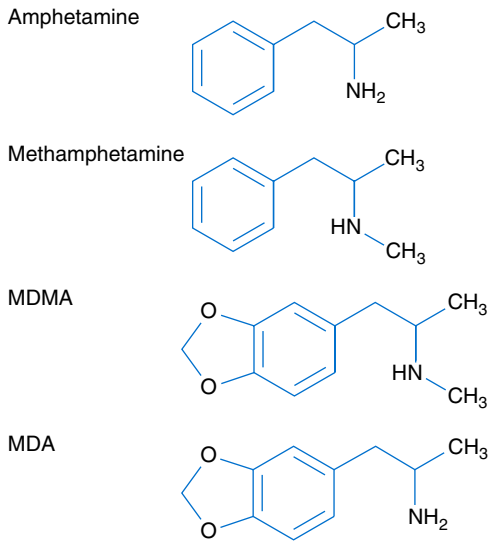


Fig. 31.4 Select amphetamine-type stimulants.

Initial screening tests for the previously listed drugs are typically immunoassays. These assays are calibrated at established cutoff concentrations. Specimens yielding responses greater than the cutoff (threshold) value are considered presumptively positive, whereas values less than the cutoff are considered negative. Immunoassays may demonstrate limited and variable specificity within certain drug classes. Similar drugs may result in a positive test; for example, pseudoephedrine, present in cold medications, may produce a positive response in immunoassays designed to detect amphetamine and methamphetamine. Therefore it is imperative that positive screening tests be confirmed by an alternate, more definitive test. The most widely accepted method for confirmatory testing is GC-MS; however, LC-MS/MS is also used and accepted.

In the following sections, the pharmacologic and analytical aspects of commonly measured drugs will be discussed.

Amphetamine Type Stimulants

Several stimulants and hallucinogens chemically related to phenylethylamine are referred to collectively as amphetamine-type stimulants (ATSs). They are considered to be sympathomimetic drugs because they mimic or enhance endogenous transmitters in the sympathetic nervous system.

Amphetamine and methamphetamine (Fig. 31.4) are CNS stimulant drugs that have limited legitimate pharmacologic use, including narcolepsy, obesity, and attention-deficit/hyperactivity disorders (ADHDs). They produce an initial euphoria and have a high abuse potential. These drugs have a stimulating effect on both the central and peripheral nervous systems. In the brain, a primary action is to increase the concentrations of extracellular monoamine neurotransmitters (dopamine, serotonin, norepinephrine).

Amphetamine and methamphetamine increase blood pressure, heart rate, body temperature, and motor activity, relax bronchial muscle, and depress the appetite. Abuse of these drugs may lead to strong psychological dependence, marked tolerance, and mild physical dependence associated with tachycardia, increased blood pressure, restlessness, irritability, insomnia, personality changes, and a severe form of chronic intoxication psychosis similar to schizophrenia. These unpleasant responses reinforce repetitive use of the drugs to maintain the "high." Tolerance and psychological dependence develop with repeated use of amphetamines. Long-term effects may include depression and impaired memory and motor skills, probably caused by a decrease in dopamine transporters and by damage to dopaminergic and serotonergic neurons. Methamphetamine has greater CNS efficacy, most likely because of its greater ability to penetrate the CNS.

The main metabolic pathways of amphetamine and methamphetamine include aromatic hydroxylation, aliphatic hydroxylation, *N*-demethylation, oxidative deamination, *N*-oxidation, and conjugation of nitrogen. Pharmacogenetics is thought to play a role in differences seen in the metabolism/elimination of these drugs. CYP2D6 is responsible for 4-hydroxylation of amphetamine and methamphetamine and *N*-demethylation of methamphetamine.

In addition to hepatic metabolism, amphetamine is eliminated as an unchanged drug in urine. Elimination is dependent on urine pH. Therefore elimination half-life (renal excretion and hepatic metabolism) varies with urine pH from 7 to 14 hours at acid pH to 18 to 34 hours at alkaline pH. These effects of urine pH on the elimination of unchanged amphetamines are a consequence of tubular reabsorption of nonionized amphetamine (pK_a, 9.9). Similarly methamphetamine is eliminated in urine in a pH-dependent manner, similar to that used for amphetamine.

Designer Amphetamines

The terms "designer drugs" and "club drugs" originated in the 1980s. These drugs include phenylethylamine, benzylpiperazine, phenylpiperazine, pyrrolidinophenone, and cathinone and derivatives (Table 31.7) and have gained popularity and notoriety among people who participate in all-night dance parties (raves) and who visit nightclubs. Most designer drugs produce feelings of euphoria and energy and a desire to socialize; they also promote social and physical interactions. They are used at these events to enhance energy for prolonged partying and/or dancing and to distort or enhance visual and auditory sensations. The moniker "club drug" does not imply that recreational use is restricted to this social environment. In this context, designer drugs mistakenly have the reputation of being safe; several experimental studies in rats and humans and epidemiologic studies have revealed risks to humans such as life-threatening serotonin syndrome, hepatotoxicity, neurotoxicity, psychopathology, and the abuse potential of such drugs.

Ephedrine and Pseudoephedrine

These amines are diastereoisomers that possess two asymmetrical carbon atoms and exist as isomers designated as 1*R*,2*S*- and 1*S*,2*R*-ephedrine and 1*R*,2*R*- and 1*S*,2*S*-pseudoephedrine.

TABLE 31.7 Designer Drugs Related to Phenylethylamine, Benzylpiperazine, Phenylpiperazine, Pyrrolidinophenone, and Cathinone

Phenylethylamines	3,4-Methylenedioxymethamphetamine (MDMA; Ecstasy)
	3,4-Methylenedioxyethylamphetamine (MDEA; "Eve")
	3,4-Methylenedioxyamphetamine (MDA), which is also a metabolite of MDMA
	Paramethoxyamphetamine (PMA)
	Paramethoxymethamphetamine (PMMA)
	2,5-Dimethoxy-4-methylamphetamine (DOM)
	2,5-Dimethoxy-4-methylthioamphetamine (DOT)
	4-Iodo-2,5-dimethoxyamphetamine (DOI)
	2,5-Dimethoxy-4-bromo-amphetamine (DOB)
	2,5-Dimethoxy-4-bromo-methamphetamine (MDOB)
	3,4-(Methylenedioxyphenyl)-2-butanamine (BDB)
	N-Methyl-1-(3,4-methylenedioxy-phenyl)-2-butanamine (MBDB)
	6-Chloro-3,4-methylenedioxymethamphetamine (Cl-MDMA)
	3,4-Methylenedioxymethcathinone
	4-Bromo-2,5-diemethoxy-phenylethylamine (2C-B)
	2,5-Dimethoxy-4ethylthio-phenylethylamine (2C-T-2)
	2,5-Dimethoxy-4 propylthio-phenylethylamine (2C-T-7)
Benzylpiperazines	1-Benzylpiperazine (BZP)
	1-(3,4-Methylenedioxybenzyl)-piperazine (MDBP)
Phenylpiperazines	1-(3-Trifluoromethylphenyl)piperazine (TFMPP)
	1-(3-Chlorophenyl)piperazine (mCPP)
	1-(4-Methoxyphenyl)piperazine (MeOPP)
Pyrrolidinophenone	α-Pyrrolidinopropiophenone (PPP)
	4-Methoxy-α-pyrrolidinopropiophenone (MOPPP)
	3,4-Methylenedioxy-α-pyrrolidinopropiophenone (MDPPP)
	4-Methyl-α-pyrrolidinopropiophenone (MPPP)
	4-Methyl-α-pyrrolidinohexanophenone (MPHP)
Cathinones	beta-keto analogues of methamphetamine
	• methcathinone
	• mephedrone (4-MMC)
	• methylone (MDMC)
	• butylone (bk-MBDB)
	• ethylone
	• 4-methylethcathinone (4-MEC)
	• 4-fluoromethcathinone (3-FMC)
	• 3-fluoromethcathinone (4-FMC)
	• ethcathinone
	• methedrone
	naphyrone
	α -Pyrrolidinophenone-derived
	• α-pyrrolidinovalerophenone (α-PVP)
	• α-pyrrolidinopropiophenone (PPP)
	• 4-methoxy-α-pyrrolidinopropiophenone (MOPPP)
	• 3,4-methylenedioxy-α-pyrrolidinopropiophenone (MDPPP)
	• 4-methyl-αpyrrolidinopropiophenone (MPPP)
	• 4-methyl-α-pyrrolidinohexanophenone (MPHP)
	• 3,4-methylenedioxypyrovalerone (MDPV)

The 1*R*,2*S*-ephedrine (ephedrine) and 1*S*,2*S*-pseudoephedrine (pseudoephedrine) isomers occur naturally in various plants of the *Ephedra* genus.

Ephedrine is both an α-adrenergic and a β-adrenergic receptor agonist. In addition, it enhances the release of norepinephrine from sympathetic neurons and is considered a mixed-acting sympathomimetic drug. Adverse effects such as hypertension, tremors, myocardial infarction, seizures, and stroke have resulted in fatalities. Because of this, the US Food and Drug Administration (FDA) banned the sale of dietary supplements containing ephedra in 2004.

Pseudoephedrine is used primarily as a decongestant because of its vasoconstrictive properties (α-adrenergic action). Pseudoephedrine can be used as a precursor for the illicit synthesis of methamphetamine. Because of this, the quantity per purchase of products containing these drugs is now restricted in many places.

Methylphenidate (Ritalin)

Methylphenidate (MPH) is a psychostimulant with pharmacologic properties similar to those of *S*(+) amphetamine. It is commonly used to treat ADHD and narcolepsy. It also exists

TABLE 31.8 Prescription Drugs That Either Are, or Are Metabolized to, Amphetamine or Methamphetamine

Drug		Drugs Detected[13]
Adderall	amphetamine	amphetamine
Dexedrine	D-amphetamine	D-amphetamine
Deprenyl	selegiline	L-methamphetamine L-amphetamine
Vyvanse	lisdexamfetamine	D-amphetamine
Didrex	benzphetamine	D-methamphetamine D-amphetamine

as an isomer, as (*R*,*R*)-MPH (d-MPH), and as (*S*,*S*)-MPH (l-MPH). The pharmacologic actions of MPH are almost solely performed by the d-isomer. Diversion and abuse of MPH have been increasing among children and adults because of its stimulant and purported aphrodisiac properties. In overdose, the clinical effects of MPH are similar to those of amphetamines.

Analytical Methods

The initial screening test for amphetamines and related drugs is typically immunoassay. For confirmatory testing of a presumptive positive test, a quantitative drug measurement is performed using GC-MS or LC-MS/MS.

Most amphetamine immunoassays have been designed to detect amphetamine/methamphetamine; others have been designed to detect MDMA and MDA; and others to more broadly capture the ATS group, all with varying cross-reactivity. Few amphetamine immunoassays were suitable for detection of the designer amphetamine or MPH, although some other chemically related compounds, such as pseudoephedrine, have been shown to produce positive results. In addition, many psychotropic medications have been reported to interfere with immunoassays. Regarding MPH, it should be noted that its detection by urine drug immunoassay is problematic because it does not cross-react well with amphetamine antibodies and is present in low concentration.

All positive immunoassay results should be confirmed by a second independent method. However, if the other designer amphetamines are suspected, due to the poor cross-reactivity, a negative amphetamine immunoassay screen cannot rule out the presence of these drugs. Fortunately, numerous GC- and LC-based methods for identification and quantitation of these drugs in biologic samples have been described.

Most published methods for analysis of members of the amphetamine class use liquid-liquid extraction (LLE) or solid-phase extraction (SPE). ATSs are considered volatile and are lost during a dry-down or evaporation step during extraction. In addition, because of their extreme volatility at the high temperatures encountered in GC-MS, and of lack of specificity of the mass spectrum, derivatization before analysis is frequently used. LC-MS and LC-MS/MS have been gaining in popularity.[1] Other prescription drugs that are metabolized to amphetamine or methamphetamine are listed in Table 31.8. Care must be taken in interpreting the results of drug screens.

TABLE 31.9 Half-life and Significant Active Metabolites of Select Barbiturates

Drug	Half-Life	Active Metabolite
Ultrashort Acting		
thiopental	6–46 h	pentobarbital
methohexital	1.2–2.1 h	
thiamylal	0.6–0.8 h initial 12–34 h terminal	
Short Acting and Intermediate Acting		
pentobarbital	15–48 h	
secobarbital	22–29 h	
butalbital	35–88 h	
aprobarbital	14–34 h	
amobarbital	15–40 h (dose dependent)	
butabarbital	34–42 h	
Long Acting		
phenobarbital	2–6 day	
mephobarbital	48–52 h	phenobarbital

Barbiturates

The success of barbital in 1903 and phenobarbital in 1912 led to the synthesis and testing of more than 2500 barbiturate derivatives, of which approximately 50 were distributed commercially. However, because of their low therapeutic index and high potential for abuse, they have been largely replaced by the safer **benzodiazepines** (BZs). Nevertheless, **barbiturates** continue to be available. The classification of barbiturates as ultrashort acting, short acting, intermediate acting, and long acting refers to the duration of effect, not to the elimination half-life (Table 31.9).

The barbiturates produce varying degrees of CNS depression, ranging from mild sedation to general anesthesia, and exert their effects by activate inhibitory $GABA_A$ receptors and inhibit excitatory AMPA receptors explain their CNS-depressant effects.

Analytical Methods

Numerous commercial immunoassays for barbiturates are available, and although the degree of cross-reactivity of other barbiturates varies with each assay, most have sufficient cross-reactivity to detect the major therapeutically used barbiturates.

Numerous confirmatory methods for barbiturates have been described. These include GC with flame ionization detection, nitrogen phosphorous detection, and MS, capillary electrophoresis-ultraviolet (UV), LC using ultraviolet (LC-UV) detection, and LC-MS.

Benzodiazepines

The prototype BZs are diazepam and nordiazepam (*N*-desmethyl diazepam). More than 30 members of this group are presently used worldwide. As a class of drugs, BZs are among the most commonly prescribed drugs in the Western hemisphere because of their efficacy, safety, low addiction potential, and minimal side effects and because of

TABLE 31.10 Half-Life of Select Benzodiazepines

Drug	Half-Life (h)	Significant Phase I Metabolites
Short Acting		
midazolam	1–4	α-hydroxy-midazolam
estazolam	10–24	3-hydroxy-estazolam
flurazepam	1–3	hydroxy-ethyl-flurazepam
	47–100 (*N*-desalkyl-flurazepam)	*N*-desalkyl-flurazepam[a]
temazepam	3–13	oxazepam
triazolam	1.8–3.9	α-hydroxy-triazolam
Intermediate Acting		
flunitrazepam[b]	9–25	7-amino-flunitrazepam
Long-Acting Agents		
diazepam	21–37	Nordiazepam[a] Oxazepam[a] Temazepam[a]
quazepam	39–53	3-hydroxy-quazepam *N*-desalkyl-2-oxo-quazepam 2-oxo-3-hydroxy-quazepam
alprazolam	6–27	α-hydroxy-alprazolam
chlordiazepoxide	6–27	nordiazepam[a] oxazepam[a]
clonazepam	19–60	7-amino-clonazepam
clorazepate[c]	2	nordiazepam[a]
	31–97 (nordiazepam)	oxazepam[a]
lorazepam	9–16	
oxazepam	4–11	

[a]Active metabolite.
[b]Not available in the United States.
[c]Converted to nordiazepam by gastric HCl.

the high public demand for sedative and anxiolytic agents. The BZs given by themselves or in combination with other drugs, particularly narcotic analgesics (opioids), are among the most widely abused drugs.

Virtually all results of the pharmacologic effects of BZs are caused by their actions on the CNS. The most prominent of these effects are sedation, hypnosis, decreased anxiety, muscle relaxation, anterograde amnesia, and anticonvulsant activity. Benzodiazepines are believed to exert most of their effects by interacting with inhibitory neurotransmitter receptors directly activated by GABA. Benzodiazepines act at $GABA_A$ but not $GABA_B$ receptors by binding directly to a specific site that is distinct from that of GABA binding. Ethanol increases both the rate of absorption of BZs and the associated CNS depression, and there are significant additive effects with other sedative or hypnotic drug.

Benzodiazepines may be divided into three categories based on their onset time and elimination half-lives: short-acting agents, intermediate-acting agents, and long-acting agents (Table 31.10). These pharmacokinetic properties in part determine the primary clinical applications for some BZs. Benzodiazepines undergo hepatic oxidation (phase I) and conjugation (phase II), often forming metabolites with pharmacologic activity (see Table 31.10)

Treatment of BZ toxicity is primarily supportive. Flumazenil may be used in select cases as a competitive inhibitor of the BZ site on the GABA complex.

Analytical Methods

Benzodiazepines are measured through a variety of techniques. However, their structural diversity and wide variation in potency provide a challenge for laboratories seeking to detect all relevant members in one analytical scheme.

Several commercial immunoassay systems are available for detection and screening. Cross-reactivity in screening immunoassays of the various BZs and their metabolites varies considerably from manufacturer to manufacturer, and screening assays cannot be used to distinguish between the individual BZs. Most assays are calibrated to a common metabolite such as oxazepam, temazepam, or nordiazepam. However, the large number of different functional groups that may be present on the BZ nucleus makes it difficult to detect all drugs in this class, and some compounds may not be detected by many assays. Other factors, such as low doses and short half-lives, make the detection of some BZs especially challenging. In the absence of sufficiently sensitive or specific immunoassays, direct analysis by a confirmatory method is warranted in suspected cases.

New-generation **sedative-hypnotic drugs** such as zolpidem (Ambien), eszopiclone (Lunesta), and zaleplon (Sonata) modulate the $GABA_A$ receptor, much like the BZs, yet they are structurally different, thus not detectable.

Analysts need to be aware that the specimen type will dictate the target substance. Blood analyses invariably will target the parent BZ and perhaps the major active metabolite (e.g., nordiazepam for diazepam and other analogues metabolized to nordiazepam), whereas in urine, a metabolite is often the required target species.

Benzodiazepines and their metabolites have been extracted from biologic specimens by LLE or SPE. A hydrolysis step may be necessary to cleave the glucuronide conjugates.

Many BZs are analyzed without derivatization by GC; these include diazepam, nordiazepam, flurazepam, and alprazolam. Drugs that are more polar, including those with hydroxyl groups (such as oxazepam, temazepam, and lorazepam) or a nitro group (clonazepam, nitrazepam), display poor chromatographic characteristics and require derivatization. Chlordiazepoxide is thermally unstable and may degrade at high temperatures in the GC. Some consider GC-MS as the definitive confirmatory method; however, LC with UV detection (240 nm) and LC-MS/MS are used to detect BZs and metabolites without derivatization.

Cannabinoids

Cannabinoids are a group of compounds found in the plant *Cannabis sativa* and has been used for centuries as a medicinal and a psychotropic agent. The main psychotropic effects are euphoria, distorted perceptions, relaxation, and a feeling of well-being.

Delta-9-tetrahydrocannabinol (THC) is the primary psychoactive component of the *C. sativa* plant. It binds to the endogenous cannabinoid receptors, CB1 (neuronal) and CB2 (immune cells). These transmembrane receptors are G-protein–coupled receptors that mediate signal transduction through inhibition of adenylate cyclase and calcium ions and through activation of potassium ion channels. The distribution pattern of CB1 receptors in the CNS accounts for most of the clinical effects of THC, such as those affecting mood, memory, cognition, pain, and appetite. CB2 may regulate immune and inflammatory processes.

When **marijuana** is smoked, THC diffuses into the plasma in seconds and is distributed into multiple phases. First, it distributes to highly vascularized tissues in minutes because of its lipophilic nature. THC then is redistributed back into the bloodstream, undergoes hepatic metabolism, and slowly accumulates into less vascularized and fatty tissues. After cessation of marijuana smoking, THC and its metabolites are slowly released from fat stores.

The main psychotropic effects after inhalation of marijuana occur within minutes and can persist for several hours. The peak plasma concentration of THC is dependent on the dose and occurs during the early acute phase (6 to 10 minutes). Numerous factors contribute to the variability in dose, such as method of consumption, depth of inhalation, exposure frequency, and cannabis potency. Onset of clinical symptoms and peak plasma concentrations after oral ingestion of THC is slower (2 to 6 hours) than after inhalation. The intensity of clinical effects described for smoked cannabis occurs during multiple phases. These phases are categorized as acute (0 to 60 minutes), postacute (60 to 150 minutes), and residual (>150 minutes). The ratio of THC to 11-nor-delta-9-tetrahydrocannabinol-9-carboxylic acid (THC-COOH) metabolite has been used to estimate the time of exposure to marijuana. This approach may be useful in naive users but is unreliable in chronic abusers of marijuana.

THC is metabolized by CYP2D6 liver enzymes to greater than 100 metabolites. The main active metabolite, 11-hydroxy-THC, is further oxidized to the most abundant inactive THC-COOH.

Cannabinoids have some medicinal use. For example, dronabinol (Marinol) contains synthetic THC and is used to treat anorexia and nausea in patients with acquired immunodeficiency syndrome (AIDS), nausea and vomiting associated with chemotherapy, or asthma and glaucoma.

Legitimate concern has been raised concerning positive results from "passive inhalation" of nearby users, or for use or consumption of hemp. Since 1998 the US Federal Government has prohibited the importation of *C. sativa* seeds and oil containing greater than 0.3% THC to reduce human exposure to THC. These measures were successful in reducing potential positive cannabinoid drug screen results from dietary sources.

Analytical Methods

An immunoassay method is typically used to screen for potential cannabinoids. Immunoassay screens have been designed to detect cannabis use in urine samples using antibody reagents developed against the inactive THC-COOH metabolite; these reagents cross-react with numerous other THC metabolites. A positive screening result for THC obtained by immunoassay is frequently confirmed by detection of THC-COOH by GC-MS/MS or LC-MS/MS analysis of the urine specimens.

A positive result from a urine cannabinoid screen or a positive result found by confirmatory testing does not indicate intoxication or degree of exposure. The window of detection for the urine concentration of THC-COOH varies among casual (2 to 7 days) and chronic abusers (up to 73 days) of marijuana and is dose dependent. Variables affecting the duration of detection include dose, frequency of exposure, route of exposure, body composition, fluid excretion, and method of detection. In practice, monitoring of abstinence is particularly challenging because dilution of urine due to normal biologic fluctuations (hydration) or ingested adulterants has caused a negative result one day and a positive on the next. To correct for hydration fluctuations, urine concentrations of THC-COOH are normalized per milligram creatinine in monitoring individuals who are resuming cannabis use. With the use of these normalized THC-COOH/creatinine concentrations, a ratio is calculated by comparing any normalized urine specimen (U2) with a previously collected normalized urine specimen (U1). "New use" is defined as a U2/U1 ratio greater than or equal to 0.5 to 1.5 in urine specimens taken more than 24 hours apart and containing THC-COOH concentrations greater than 15 ng/mL. Using the 1.5 cutoff rate results in decreased false-positive but increased false-negative decisions.

Cocaine

Cocaine is an alkaloid found in *Erythroxylon coca*. In clinical medicine, it is used for local anesthesia and vasoconstriction in nasal surgery and to dilate pupils in ophthalmology. Cocaine abuse has a long history, and cocaine is still one of the most common illicit drugs of abuse.

Cocaine is sold on the street in two forms: a hydrochloride salt (powder) and a free-base product known as "crack." The hydrochloride salt form of cocaine is administered by nasal insufflation ("snorting") or intravenously. Crack comes as a rock crystal that is heated and its vapors inhaled. It should be noted that "crack" cocaine is not to be confused with "free-basing," which is a process in which the user purifies cocaine HCl by mixing an aqueous solution of cocaine with baking soda or ammonia and adding diethyl ether, thereby extracting the free form of the drug into the organic solvent. However, because of the extremely flammable nature of diethyl ether, and therefore the risk of igniting any remaining ether, free-basing is no longer a common practice.

Chemically, cocaine is methylbenzoylecognine (COC). Its metabolism is complex (Fig. 31.5) and occurs via both nonenzymatic hydrolysis and enzymatic transformation in the plasma and liver, producing both active and inactive metabolites. COC is rapidly metabolized to BE and ecgonine methyl ester, both of which are inactive. It should be noted that

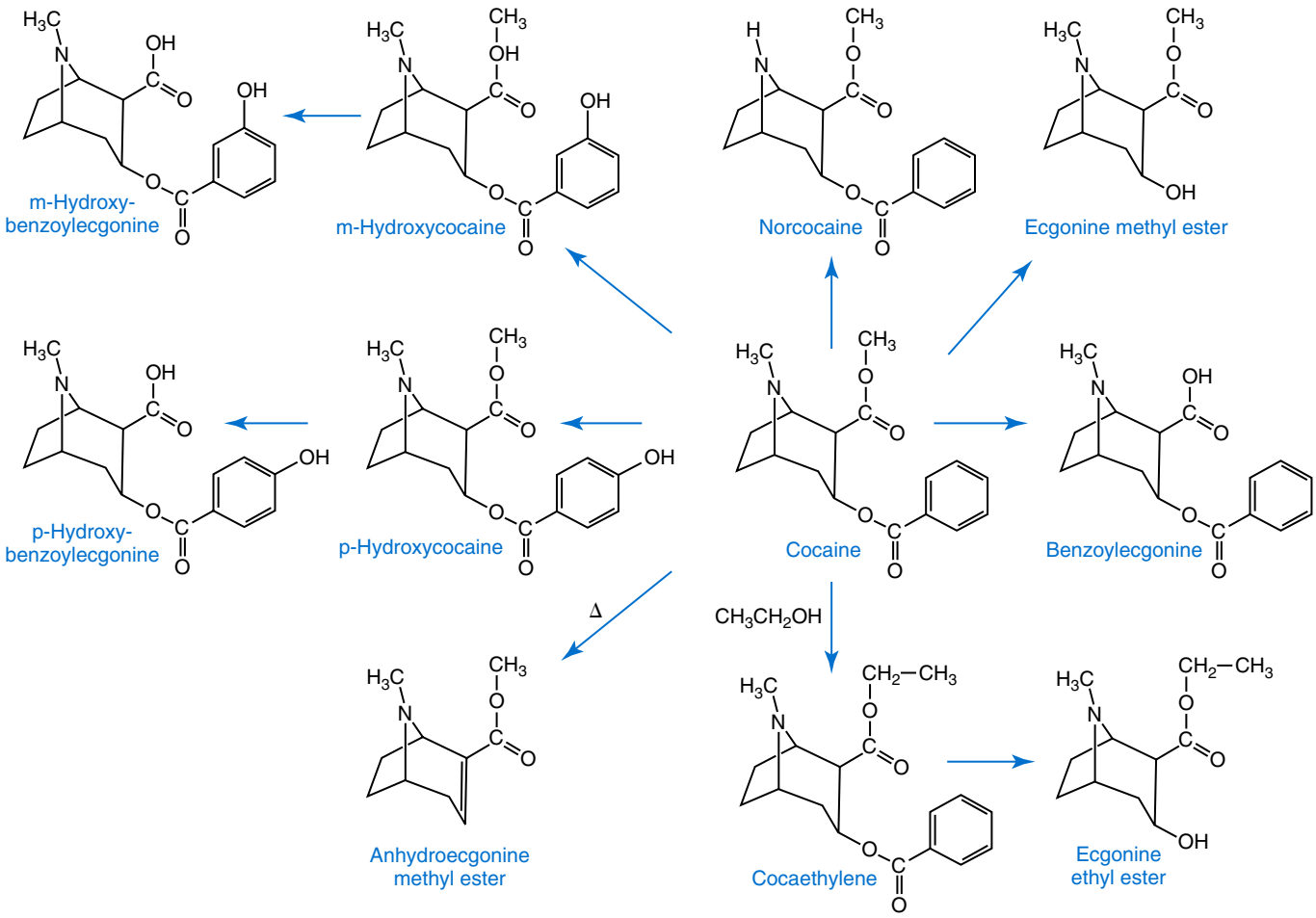


Fig. 31.5 Metabolism of cocaine.

cocaethylene, produced when cocaine and ethyl alcohol are used simultaneously, has the same CNS stimulatory activity as cocaine in experimental animals. Norcocaine (NC) is of clinical interest because of its conversion to potentially hepatotoxic metabolites: hydroxyl-NC and then NC-nitroxide.

BE has a **half-life** longer than that of COC. It is the most commonly monitored **analyte** in urine for determination of COC use. BE is further metabolized to minor metabolites such as m-hydroxybenzoylecgonine (m-HOBE), which has been shown to be an important metabolite in the meconium of cocaine-exposed babies. Anhydroecgonine methyl ester (AEME; methyl ecgonidine) has been identified as a unique COC metabolite after smoked COC ("crack") administration. Anhydroecgonine ethyl ester (AEEE; ethyl ecgonidine) has been identified in COC smokers who also use ethyl alcohol.

Cocaine is a potent CNS stimulant that elicits a state of increased alertness and euphoria with actions similar to those of amphetamine but of shorter duration. These CNS effects are thought to be largely associated with the ability of cocaine to block dopamine reuptake at nerve synapses, thereby prolonging the action of dopamine in the CNS. It is this response that leads to recreational abuse of cocaine. Cocaine also blocks the reuptake of norepinephrine at presynaptic nerve terminals; this produces a sympathomimetic response.

The CNS and cardiovascular effects of cocaine exhibit acute tolerance; its effects are more pronounced when the concentration of cocaine in blood is increasing than when it is at a similar but decreasing concentration *(clockwise hysteresis)*. Thus it is difficult to correlate isolated blood concentration values with psychomotor effects.

COC is frequently used with other drugs, most commonly ethanol. This combination produces greater euphoria and an enhanced perception of well-being relative to COC. CE appears to be equipotent to COC with regard to dopamine transporter affinity but is less potent than cocaine pharmacologically. It has been suggested that simultaneous COC and ethanol use carries an 18- to 25-fold increase in risk of toxicity than that associated with COC alone.

Analytical methods. The half-life of COC is 0.5 to 1.5 hours, whereas the half-life of BE is 4 to 7 hours. Thus BE is the analyte of choice in screening for cocaine use. The initial screening test for cocaine use is the immunoassay detection directed against the primary metabolite BE.

Most confirmatory assays offer quantification of both parent drug and metabolite. Numerous methods have been described for measurement of COC and various metabolites. GC techniques for analysis of COC and its metabolites require derivatization, especially of polar metabolites. GC-MS is the method of choice for many laboratories. In addition,

LC methods measure simultaneously many of the more polar potentially significant secondary metabolites such as CE, NC, AEME, and AEEE. For these reasons, LC-MS/MS methods are gaining in popularity.

Lysergic Acid Diethylamide

LSD is an extremely potent psychedelic ergot alkaloid derived from the fungus *Claviceps purpurea,* which grows on wheat and other grains. LSD shares structural features with serotonin (5-hydroxytryptamine), a major CNS neurotransmitter and neuromodulator. The drug LSD binds to serotonin receptors in the CNS and acts as a serotonin agonist. The principal psychological effects of LSD are perceptual distortions of color, sound, distance, and shape (synesthesia); depersonalization and loss of body image; and rapidly changing emotions from ecstasy to depression or paranoia. The most common adverse effects of LSD are panic attacks. In addition, unpredictable recurrence of hallucinations (flashbacks) may occur weeks or months after last drug use, and LSD may elicit psychotic reactions (thought disorders, hallucinations, depression, and depersonalization).

The drug is rapidly absorbed from the gastrointestinal tract, and its effects begin within 40 to 60 minutes, peak at approximately 2 to 4 hours, and subside by 6 to 8 hours. The elimination half-life is approximately 3 hours. LSD is extensively metabolized in humans and the metabolite 2-oxo-3-hydroxy-LSD is present in urine at concentrations 10- to 43-fold greater than those of LSD.

Analytical Methods

Because of the very high potency of LSD and its subsequent rapid and extensive metabolism, only approximately 1% to 2% of the drug is excreted unchanged in urine. Thus detection of LSD presents an especially difficult analytical challenge. Even with sensitive assays, the detection window for LSD is generally only 12 to 24 hours. The detection window may be extended, perhaps twofold to threefold, by including 2-oxo-3-hydroxy-LSD in the confirmatory test, while using sensitive techniques such as GC-MS-MS, LC-MS-MS, or LC-MS. To avoid degradation of LSD and 2-oxo-3-hydroxy-LSD or epimerization of LSD to iso-LSD, urine specimens should be protected from sunlight, bright fluorescent light, and elevated temperature at alkaline pH.

Opiates and Opioids

The term **opioid** describes compounds that encompass the natural and semisynthetic **opiates** that include variations on the structure of morphine and fully synthetic opioids with minimal structural homology to the natural alkaloids (Fig. 31.6). The defining characteristic of this class of drugs is their morphine-like activity stemming from interaction with opioid receptors. Other compounds that are referred to as "opioids" include receptor antagonists and mixed agonist/antagonists, as well as other opium-derived alkaloids such as papaverine.

Opioids interact with the family of opioid receptors that are variably distributed throughout the body. The classical opioid receptors are divided into the mu, delta, and kappa (μ, δ, and κ or MOR, DOR, and KOR, respectively) subfamilies. Opioid receptor agonists typically produce analgesia, and antagonists block this response. Most opioids have both substantial addictive capacity and potentially life-threatening side effects. Thus the benefits of their use in non–end-stage patients must be carefully weighed against the chance of serious consequences. The development of tolerance and the risk of prescription diversion complicate even further the process of monitoring long-term opioid therapy for compliance and efficacy.

One of the more important CYP enzymes, CYP2D6, is particularly notable for its role in variable clinical response to opioids; it will be discussed in greater detail in a later section. Many additional CYP enzymes, including CYP3A and CYP2C and others, are involved in opioid metabolism. It is important to note that several of these enzymes are subject to substrate inhibition and/or induction. Substrate-dependent changes in metabolic activity are affected by other drugs, herbal supplements, and endogenous compounds that are substrates of the same enzyme. For example, methadone concentrations may be lower than expected in a patient taking St. John's Wort—a noted CYP3A4 inducer—but higher in a patient ingesting a CYP3A4 inhibitor such as grapefruit juice.

Types of Opiates

Natural opium alkaloids. Opium is obtained from the unripe seed capsules of the poppy plant, *Papaver somniferum.* The milky juice is dried and powdered to make powdered opium, which contains several alkaloids. These alkaloids are divided into two distinct chemical classes: *phenanthrenes* and *benzylisoquinolines.* The principal phenanthrenes are morphine, codeine, and thebaine. The principal benzylisoquinolines are papaverine, which is a smooth muscle relaxant, and noscapine.

Poppy seeds from the poppy plant *P. somniferum* contain morphine and, to less extent, codeine. Ingestion of bakery products containing poppy seeds leads to excretion of morphine (and codeine) in urine. In practice, caution is required when the results of a positive urine test for morphine and codeine are interpreted.

Morphine. The archetypical opiate, morphine is used as the basis of comparison for relative characterizations of the opioid class Morphine interacts primarily with MOR to mediate its effects. Its major metabolites are glucuronide conjugates, including inactive morphine-3-glucuronide (M3G; ≈60%), active morphine-6-glucuronide (M6G; ≈10%), and a small amount of morphine-3,6-diglucuronide. With long-term administration, and when morphine concentrations are high, a minor fraction is converted to hydromorphone (up to 2.5% of the urine morphine concentration).

The detection time for morphine in urine is usually 48 to 72 hours, but this varies with individual differences in metabolism excretion and route and frequency of use.

Codeine. Because of its use as an antitussive and analgesic, codeine is frequently prescribed and is commonly combined with nonopiate analgesic agents such as aspirin and

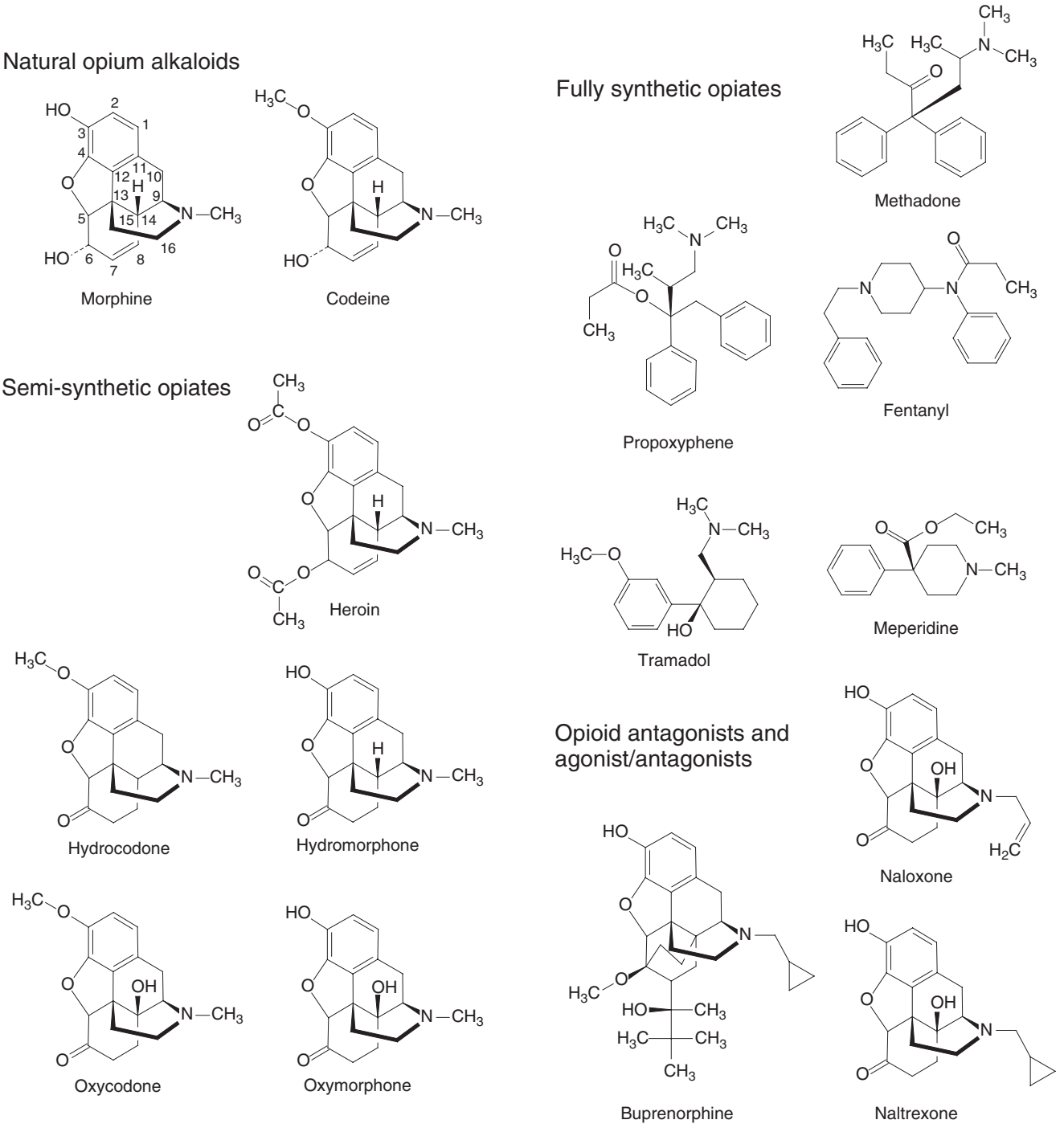


Fig. 31.6 Structure of common opioids.

acetaminophen. Codeine has only approximately one-tenth the analgesic potency of morphine and is generally considered a prodrug. The small fraction of codeine converted to morphine by CYP2D6 is considered to be responsible for the analgesic effect. During the early phase of excretion, codeine and conjugates predominate, but after this time, morphine conjugates are the major product. Approximately 3 days after codeine use, morphine and its conjugates are the only metabolites detected.

Genetic variation may play a significant role in the metabolism of codeine and several other opioids. More than 60 alleles have been described for CYP2D6, with resultant enzymatic activity ranging from essentially zero, in the case of null alleles, to many times higher than normal, in the case of amplified alleles (http://www.cypalleles.ki.se/cyp2d6.htm/). Thus, at the same codeine dose, patients with minimal CYP2D6 activity (poor metabolizers) would likely receive inadequate analgesia because of lack of conversion to morphine; however, patients with very high CYP2D6 activity (ultrarapid metabolizers) would be at risk for adverse responses to excessive morphine. Without knowledge of the CYP2D6 genotype, these clinical presentations

TABLE 31.11 Examples of Semisynthetic Opiates Include Heroin, Hydrocodone, Hydromorphone, Oxycodone, and Oxymorphone and Their Metabolites

Parent Drug	Primary Metabolite Active(s)	Minor Metabolite(s)
Natural and Semisynthetic Opiates		
morphine		hydromorphone (<3%)
codeine	morphine	hydrocodone (<11%)
heroin	6-acetylmorphine, morphine	
hydromorphone		
hydrocodone	hydromorphone, norhydrocodone	dihydrocodeine
oxycodone	oxymorphone, noroxycodone	
oxymorphone	noroxymorphone	
Fully Synthetic Opioids		
fentanyl	nor-fentanyl	
meperidine	normeperidine	
methadone	EDDP (2-ethylidene-1,5-dimethyl-3,3-diphenylpyrrolidine)	
tramadol	*N*-desmethyltramadol, *O*-desmethyltramadol	
Opioid Antagonists and Mixed Agonists/Antagonists		
Buprenorphine		
naloxone	noroxymorphone	
naltrexone		

are confusing. Therefore the possibility of pharmacogenetic effects is important to consider when appropriate dosing, patient compliance, and potential diversion or illicit uses are assessed. Like morphine with long-term administration, and when codeine concentrations are high, a minor fraction is converted to hydromorphone (up to 11% of the urine codeine concentration).

Semisynthetic opiates, fully synthetic and opioid antagonists, and mixed agonists/antagonists. Examples of semisynthetic opiates include heroin, hydrocodone, hydromorphone, oxycodone, and oxymorphone and their metabolites (outlined in Table 31.11).

Analytical Methods

Many different immunoassay methods are used to screen for opiates. Given their relatively rapid turnaround time and ability to detect several opiates, immunoassays are the methods of choice to screen urine samples for their opiate content. Antibodies in opiate abuse screens commonly target morphine, and there is significant variability in cross-reactivity such that many opiates and opioids with high abuse potential such as oxycodone are often poorly detected.

The cutoff concentrations used for drugs in federally regulated tests, in particular, opioids (2000 ng/mL), are too high to be of value in monitoring compliance of pain management patients or other clinical settings. For these circumstances, a cutoff of 300 ng/mL morphine (or morphine equivalents) is commonly used to distinguish negative from positive urine specimens. Monitoring compliance for oxycodone in pain management programs should use oxycodone-specific immunoassays because of the low cross-reactivity of oxycodone in most opiate immunoassays.

It is important for drug testing laboratories to communicate relevant aspects of the metabolic interconversion of opiates to physicians responsible for these programs.

GC-MS and LC-MS/MS are techniques of choice for confirmatory testing after a positive screening result or when looking for opiates that are not well detected by immunoassay. GC-MS has historically been considered the method of choice. However, despite the long-standing role of GC in opiate analysis, LC methods are common and are often analytically advantageous. One notable example is that LC provides the ability to analyze glucuronide-conjugated metabolites, as well as parent compounds. In addition, LC methods can be used to measure polar metabolites without prior derivatization, and on-column extraction is possible with some LC systems. One trend in method development is the ability to quantitate/identify combinations of morphine, other opiates, and opioids and their metabolites.

Phencyclidine and Ketamine

Phencyclidine (PCP) and ketamine are potent analgesics and general anesthetics. They are classified as dissociative anesthetics because they produce rapid-acting dissociation of perception, consciousness, movement, and memory. PCP and ketamine share similar structural features and pharmacologic actions. The effects are dose dependent and vary between individuals. PCP is listed as a Schedule II drug in the US Federal Controlled Substance Act and is not approved for human use. Ketamine is a Schedule III drug that is commonly used as an anesthetic. Both drugs have been used illicitly as a **drug of abuse**, as well as in cases of drug-facilitated sexual assault (DFSA). At anesthetic doses, it produces profound analgesia, but the individual is awake yet incapacitated, with

limited voluntary limb movement. They bind and antagonize the excitatory glutaminergic system by binding to NMDA receptors. They also decrease GABA transmission, disrupt cortical activity, and increase dopamine and norepinephrine synaptic reuptake.

Analytical Methods

Whether or not PCP and ketamine are included in a general urine drug screen depends on applicable regulations and on the prevalence of use in the local community. In some locations, the prevalence of PCP or ketamine use may be too low to warrant routine screening. Initial screening is typically done by immunoassay. Immunoassays for PCP are generally reliable; however, false-positives have been reported because of high concentrations of dextromethorphan (DXM), diphenhydramine, and thioridazine. Immunoassay-positive specimens should be confirmed using GC-MS or LC-MS techniques. Note, these techniques will also detect norketamine and dehydronorketamine, the active metabolites of ketamine.

Specimen Validity Testing

Several techniques are used by persons attempting to mask or adulterate drugs to avoid detection. These tactics may include the exchange of urine from a drug-free individual or dilution of the urine specimen by excessive consumption of water, use of a diuretic or simple addition of water to the specimen to reduce drug concentrations to less than cutoff limits. In addition, readily available adulterants, such as detergent, bleach, salt, alkali, ammonia, tetrahydrozoline, or acid, may be added to the specimen after collection in an attempt to interfere with immunoassay screening procedures. Other more sophisticated adulterants specifically marketed to avoid drug detection include glutaraldehyde (Urine Aid; Clear Choice), nitrite (Klear; Whizzies), chromate (Urine Luck; Sweet Pee's Spoiler), and a combination of peroxide and peroxidase (Stealth). These adulterants also interfere with immunoassays to variable degrees, and the oxidizing agents (nitrite, chromate, and peroxide/peroxidase) may result in destruction of morphine, codeine, and the principal metabolite resulting from marijuana use, thus interfering with their GC-MS confirmation and with immunoassays. Direct observation of urine collection is the most stringent means to guard against specimen exchange or adulteration. However, an individual's right to privacy and dignity must be weighed against the need for the highest degree of certainty of specimen integrity. Alternative measures to prevent specimen adulteration include limitations on clothing or other personal belongings allowed in the specimen collection area, addition of coloring agent to toilet water, and inactivation of the hot water tap.

DRUG-FACILITATED CRIMES

DFCs are defined as voluntary or surreptitious use of alcohol, drugs, and/or chemical agents to incapacitate an individual and facilitate a criminal act. This includes DDFSA. In addition to alcohol, the drugs that have been implicated in DFCs include choral hydrate, BZs, non-BZ sedative-hypnotics, **gamma-hydroxybutyric acid (GHB)**, dextramethorphan, ketamine, PCP, and nonprescription medications such as antihistamines and anticholinergics. The Society of Forensic Toxicologists Drug-Facilitated Crimes Committee has recommended analytical cutoff for testing of commonly encountered compounds (http://soft-tox.org/files/SOFT_DFC_Rec_Det_Limits_1-2014.pdf). They include impaired judgment, confusion, reduced inhibitions, sedation, hypnosis, loss of muscle coordination, and sometimes anterograde amnesia. Because of the amnesic properties of these drugs, victims often may not report their sexual assault for several days. Therefore sensitive analytical techniques are necessary to detect these drugs and their metabolites in urine or hair samples after a single dose.

Choral Hydrate

Chloral hydrate is classified as a sedative and as a hypnotic. The CNS depressant effects of chloral hydrate are believed to be due to its active metabolite trichloroethanol (TCE). The clinical diagnosis of chloral hydrate intoxication is difficult to differentiate from alcohol or sedative-hypnotics because it shares similar clinical effects. The exact mechanism of action of chloral hydrate has not been determined. The elimination half-life of choral hydrate is very short (4 minutes), but the half-life of its metabolite TCE is 6 to 10 hours. If coingested with alcohol, the metabolism of chloral hydrate is impaired. Because both ethanol and chloral hydrate are metabolized by CYP2E1 and ADH, coingestion not only exacerbates the clinical effects but also prolongs the duration of action.

Analytical Methods

Chloral hydrate is not detected on routine drug screens. Chloral hydrate and its metabolite TCE are detected with the use of LC-MS and capillary GC with electron-capture detection (GC-ECD), or GC flame ion detection (GC-FID).

Benzodiazepines

Benzodiazepines that have been reported in sexual assault victims are diazepam, triazolam, temazepam, and clonazepam (see the Benzodiazepines section earlier for more details). Flunitrazepam has achieved significant public awareness. It is a fast-acting sedative-hypnotic categorized as a Schedule I drug in the United States, but sexual predators are able to acquire this drug through illegal trafficking. Sexual assault predators use flunitrazepam because it is easily dissolved into a beverage, it is relatively tasteless and odorless, it quickly incapacitates their victims, and routine drug screens do not detect its presence.

Analytical Methods

Detection of flunitrazepam is especially challenging because of the low doses and the low degree of cross-reactivity of most immunoassays with the principal urinary metabolite, 7-aminoflunitrazepam (see BZ analytical methods earlier), although specific immunoassays have been developed. Direct analysis or confirmatory testing of 7-aminoflunitrazepam by GC-MS or LC-MS/MS is indicated in suspected cases.

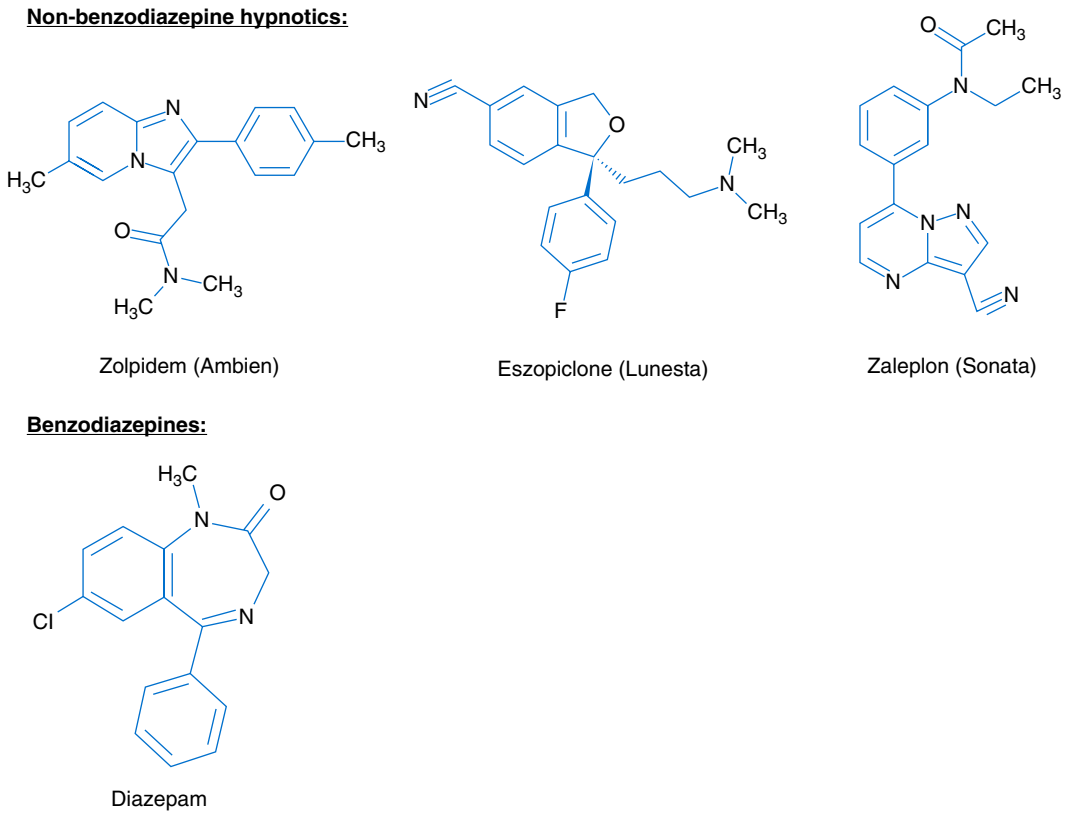


Fig. 31.7 Chemical structures of diazepam (typical benzodiazepine) and the nonbenzodiazepine sedative-hypnotics (zolpidem, eszopiclone, zaleplon).

Nonbenzodiazepine Sedative-Hypnotics

A new generation of sedative-hypnotics is available that are structurally different from BZs (Fig. 31.7). They include zopiclone, eszopiclone, zolpidem, and zaleplon and are prescribed as sleep aids through a prescription for a Schedule IV drug. However, they are readily prescribed, easily shared, and sold illegally. The rapid onset and amnesic properties of this class of drugs result in disinhibition, passivity, and retrograde amnesia, which are favorable characteristic for use in DFCs. The pharmacologic effects of these agents result from their interaction with a specific subtype of $GABA_A$ receptor complex. These drugs modulate the $GABA_A$ receptor chloride channel by binding to the BZ receptors in the brain without binding to peripheral BZ receptors.

Analytical Methods

Although these drugs do not cross-react with most BZ immunoassays, specific reagent systems (ELISA) directed against the non-BZ hypnotics are available. However, most screening and confirmatory testing are performed by GC-MS or LC-MS/MS.

Gamma-Hydroxybutyrate, 1,4-Butanediol, and Gamma-Butyrolactone

GHB and its synthetic precursor compounds, 1,4-butanediol (1,4-BD) and gamma-butyrolactone (GBL), are Schedule I agents in the United States, and availability is restricted in numerous other countries. GHB is an odorless and colorless liquid, or it can be obtained as an off-white powder that easily dissolves in liquids. When ingested, it has CNS depressant effects, resulting in sedation and hypnosis.

GHB is a naturally occurring substance that is produced in the brain. It is metabolized to GABA by multiple endogenous enzymes. Consumption of GHB, or the synthetic GHB precursor compound 1,4-BD or GBL, will promote GABA activity. In addition to increased metabolism to GABA, GHB has direct effects on the CNS by binding GHB-specific receptors and $GABA_B$ receptors.

GHB is rapidly metabolized ($t_{½} \approx 30$ minutes). Onset of its effects occurs in approximately 15 to 30 minutes and the duration of response is short, typically 1 to 3 hours for a normal dose (1 to 5 g) and 2 to 4 hours with excessive doses.

Analytical Methods

GHB is not detected on most screens but can be identified with the use of GC-FID or GC-MS. Because GHB is metabolized rapidly, timely sample collection is an important facet of GHB assay. Plasma samples should be collected within 6 to 8 hours after ingestion, and urine samples within 10 to 12 hours. Endogenous concentrations of GHB are typically 1.0 mg/L and after exogenous exposure may return to normal levels within 8 to 12 hours after ingestion. Therefore timely presentation of the sexually assaulted victims and physician recognition of GHB symptoms are essential for prosecution of sexual offenders.

Dextromethorphan

DXM is structurally related to the opioids, but it does not bind to opioid receptors at normal dose. The (L–) isomer of DXM, levorphan (not available in the United States), is a potent opioid analgesic and is an example of the stereoselective nature of opioid receptor binding. DXM lacks analgesic activity but does have antitussive activity comparable with that of codeine. DXM is present in numerous over-the-counter (OTC) cough medications, often in combination with other medications.

Analytical Methods

Clinically approved doses of DXM are not detected by most clinical opiate immunoassays, but larger doses may cross-react. ELISA assays are available to detect DXM and its major metabolite, dextrorphan. The presence of DXM or dextrophan in a sample is confirmed by GC-MS or LC-MS/MS. Dextrophan is the **enantiomer** of levorphanol, a potent opioid agonist available in the United States (Levo-Dromoran). Unless chiral analytical techniques are used, these enantiomers are not resolved.

PAIN MANAGEMENT

UDT is an essential tool in the field of pain medicine. In the late 1990s, awareness grew about the personal and economic costs of poorly controlled pain. Prominent physicians and professional organizations at local, national, and international levels supported judicious opioid use to address not only acute and cancer-related pain but society's pain at large. In short order, the negative consequences of widespread opioid use surfaced. Since 1999 the number of opioid-related deaths in the United States has more than doubled and now exceeds the number of deaths related to heroin and cocaine abuse combined. In 2012, drug overdose was the leading cause of injury death, surpassing that from motor vehicle crashes between the ages 25 and 64. For every one death from opioid abuse, there are 35 emergency department visits, 161 cases of abuse and 461 cases of misuse, contributing to more than $55 billion dollars in society and work-related costs.

Despite widespread agreement that treating chronic pain with opioids can be risky and requires commitment to monitoring, controversy in pain medicine exists as to the type and frequency of UDT. This controversy likely stems from the paucity of high-quality research demonstrating that UDT significantly reduces the risks of misuse, abuse, and diversion. In addition, evidence exists that inconsistent UDTs do not necessarily alter prescribing practices. Coincident are concerns about conflicts of interest and physician enticement because UDT is now a multibillion industry that has a history of incentivizing physicians to frequently order testing.

UDT is an important tool in clinical pain medicine. The current epidemic of prescription medication abuse underscores the need for a consistent approach to monitoring that uses both subjective and objective measures. Although some advocate that routine UDT should include all testable substances with abusive potential, such practices are costly and likely unnecessary in many cases. High-risk patients, classified by using validated clinical screening tools, require more careful and more frequent monitoring. POC and other enzyme immunoassay screening methods cannot differentiate between medications within the same drug class and may completely miss semisynthetic and synthetic opioids. For this reason, confirmatory testing with GC or LC should be performed at least annually and on all inconsistent samples. If misuse, abuse, or diversion is suspected, the physician must act in a way that promotes patient health and safety, while abiding by state and federal law. Such actions should be openly discussed and explicitly written in the opioid prescribing contract signed at the initiation of every opioid treatment plan.

DETECTION OF DRUGS OF ABUSE USING OTHER TYPES OF SPECIMENS

Blood, serum, and urine samples are typically collected for the purpose of determining exposure to various agents. However, sampling of blood is considered invasive, and collection of urine may require some invasion of privacy and loss of dignity. In addition, urine specimens are subject to adulteration or manipulation to evade detection. For these reasons, alternate biologic specimens have been investigated as alternative types of samples for drug analysis.

Meconium

The first intestinal discharge from newborns is meconium, which is a viscous, dark green substance composed of intestinal secretions, desquamated squamous cells, lanugo hair, bile pigments, and blood. Meconium also contains pancreatic enzymes, free fatty acids, porphyrins, interleukin-8, and phospholipase. Meconium begins to form during the second trimester and continues to accumulate until birth; drugs taken by the mother can be detected in the meconium of the newborn. Drug testing is necessary in infants to substantiate the clinical indication of prenatal exposure to illicit drugs.

The disposition of drug in meconium is not well understood. The proposed mechanism is that the fetus excretes drug into bile and amniotic fluid. Drug accumulates in meconium by direct deposition from bile or through swallowing of amniotic fluid. Therefore the presence of drugs in meconium has been proposed to be indicative of in utero drug exposure potentially in the second and third trimesters—a longer historical measure than is possible by urinalysis.

Meconium testing has some limitations. For example, meconium is usually passed by full-term newborns within 1 to 2 days; however, low-birth-weight infants may have delayed passage (median age of 3 days). Thus meconium collection is missed completely, or significant delay is seen in detection of intrauterine drug exposure.

In the clinical laboratory, meconium is a sticky material that is difficult to work with. Meconium drug screening has been adapted to various analytical immunoassays, but as with any immunoassay-based drug screen, confirmatory testing by LC-MS/MS or GC-MS is necessary. Confirmatory assays for meconium are more difficult than those for urine. Recovery of drugs from meconium is low (10% to 50%). Some debate

TABLE 31.12 Drugs and Metabolites of Significance in Meconium

Drug Class	Confirmation Compound
Cocaine	cocaine
	benzoylecgonine
	cocaethylene
	m-hydroxybenzoylecgonine
Opiates	morphine
	codeine
	6-monoacetylmorphine (6-MAM)
	hydromorphone
	hydrocodone
	oxycodone
Cannabinoids	9-carboxy-11-nor-delta-9-THC
	11-hydroxy-delta-9-tetrahydrocannabinol
	8,11-dihydroxy-delta-9-tetrahydrocannabinol
Amphetamines	amphetamine
	methamphetamine
	MDMA
	MDA
	MDEA
Ethanol	fatty acid ethyl esters
PCP	PCP

continues as to which are the most appropriate drug analytes that should be measured in meconium; Table 31.12 attempts to summarize current knowledge. Meconium drug testing is far less standardized than UDT. Assay cutoff limits and units (ng/g meconium or ng/mL extract) may vary; suitable reference or control materials are not yet available.

Overinterpretation of meconium data is a dangerous practice. Numerous reports have described the specificity and sensitivity of different analytes with the use of different testing methods. In our opinion, tremendous, and potentially inappropriate, value has been placed on a meconium result. On occasion, decisions about treatment or custody of the infant have been based solely on meconium drug screen results. It is critical to remember that a positive test could indicate drug exposure. However, a negative result does not rule it out. It is clear that additional work is necessary to address these important issues and to improve our understanding of the disposition of drugs in meconium.

Oral Fluid

Analysis of saliva for drugs was first done almost 30 years ago for the purpose of therapeutic drug monitoring. It has since been evaluated for use in forensic toxicology, with recognition of its advantages over other biologic matrices. Most studies on saliva in humans use whole saliva. It should be noted that the term "oral fluid" is now preferred for the specimen collected from the mouth.

Several advantages are associated with monitoring oral fluid as contrasted with monitoring plasma or serum concentrations. For example, collection of oral fluid is considered to be a noninvasive procedure, direct observation is possible (thus minimizing risk of adulteration), and some of the risks associated with drawing of blood are avoided. Furthermore, for the patient, the fear, anxiety, and discomfort that may accompany the drawing of blood are diminished. Significant disadvantages of oral fluid are (1) the window of detection is about equivalent to that of blood or serum and is short compared with that of urine, (2) glucuronide metabolites (size and charge) may poorly distribute into oral fluid, (3) pH may affect deposition into oral fluid, (4) protein binding diminish deposition, and (5) small volume of sample collected. The problem of small sample size is usually overcome by using methods that simultaneously extract multiple drug groups.

In principle, oral fluid drug concentration is related to plasma free-drug concentration; therefore oral fluid has the potential to show a relation between behavior/impairment and drug concentration, making it a possible matrix for monitoring drug intoxication or for conducting therapeutic drug management. Because disposition of drugs in oral fluid is dependent upon chemical and metabolic processes, appropriate interpretation is required. In recent years, great interest has been expressed in the use of oral fluid testing for roadside drug screening, monitoring the compliance of individuals on drug maintenance programs, and workplace drug testing. Low concentrations of drugs and metabolites necessitate sensitive screening methods. Many confirmatory methods have been developed for oral fluid testing of abused drugs, including GC-MS, LC-MS, and LC-MS/MS (refer to Chapter 6 for saliva sample collection).

Hair

For longer than 30 years, hair has been analyzed for trace metals, including lead, arsenic, and mercury. This was achieved with the use of atomic absorption spectroscopy. Subsequently, interest in analysis of hair for the purpose of detecting drug use has increased.

Hair is advantageous as a biologic specimen for drug analysis because it is easily obtained with less embarrassment, and it is not easily altered or manipulated to avoid drug detection. Hair also differs from other human materials used for toxicologic analysis in that it has a substantially longer detection window (months to years) depending on the length. The rate of hair growth depends on anatomic location, race, gender, and age. Once deposited in hair, drugs are very stable, and analysis can be performed even after centuries.

The exact mechanism by which chemicals are incorporated into hair is still being studied. However, it has been suggested that passive diffusion may be augmented by binding of the drug to intracellular components of hair cells, such as the hair pigment melanin. Factors that may affect how efficiently drugs are incorporated into hair are not well established but may include rate of hair growth, anatomic location of hair, hair color (melanin content), and hair texture (thick or fine, porous or not). These factors are determined by genetic factors and by the effects of various hair treatments. This may lead to biases in hair testing for drugs of abuse.

Drugs, when deposited in hair, are generally present in relatively low concentrations; thus sensitive analytical techniques are required for detection. Immunoassay procedures

have been modified for use with hair. For confirmatory testing, GC-MS is generally the method of choice; however, various GC-MS/MS or LC-MS/MS methods have been used. External exposure to drugs causes them to be detected in hair; one of the most crucial issues currently facing hair analysts is technical and false positives. The need to distinguish between passive exposure (environmental contamination) and active consumption is fundamental; consequently, decontamination procedures for hair are compulsory. These usually involve a washing step. Needless to say, hair analysis is a complex scientific undertaking.

Sweat

Drugs may be excreted in sweat, and sweat analysis may provide an alternative detection technique. Sweat patch collection devices that resemble an adhesive bandage may be worn for several days to several weeks. During this time, a drug, if present, accumulates in the absorbent pad in the patch, while water vapor escapes through the semipermeable covering. Thus sweat drug testing offers the possibility of monitoring drug use over extended periods without the need for frequent collection of urine. Sweat drug excretion may also be an important mechanism by which drugs enter hair.

REVIEW QUESTIONS

1. The best specimen type used to determine whether a neonate has been exposed to drugs during the second or third trimester of pregnancy is:
   a. fetal oral fluid.
   b. meconium.
   c. fetal hair.
   d. newborn infant urine.
2. The antidotal therapy for acetaminophen overdose is:
   a. forced alkaline diuresis.
   b. inhibition of the GABA channel.
   c. pralidoxime.
   d. *N*-acetylcysteine.
3. The best advantage of using high-performance liquid chromatography (HPLC) over gas chromatography (GC) for comprehensive drug screening for a broad range of chemicals is that:
   a. HPLC does not require a derivatization step.
   b. HPLC can be performed at the individual's bedside.
   c. GC gives diverse color hues and requires considerable technical skill.
   d. GC drug testing methods are subject to many types of interference.
4. Use of which of the following drugs is not consistently detected in the benzodiazepine immunoassay screen?
   a. Diazepam
   b. Clonazepam
   c. Oxazepam
   d. Nordiazepam
5. To confirm heroin use, testing of which of the following urinary metabolites is considered unique and specific for this drug?
   a. Morphine
   b. Codeine
   c. Meperidine
   d. 6-Acetylmorphine
6. For assessment of drugs of abuse, the most widely accepted laboratory method for drug *confirmation* is:
   a. immunoassay.
   b. GC-MS.
   c. HPLC-UV
   d. Time-of-Flight (TOF) Mass Spectroscopy

SUGGESTED READINGS

Baldessarini, R. J., & Tarazi, F. I. (2006). Chapter 18. Pharmacotherapy of psychosis and mania. In L. L. Brunton, J. S. Lazo, & K. L. Parker (Eds.), *Goodman & gilman's the pharmacological basis of therapeutics* (11th ed.). McGraw-Hill Companies, Inc. p. http://www.accessmedicine.com/content.aspx?aID=939118.

Baselt, R. C. (2011). *Dispositition of toxic drugs and chemical in man* (9th ed.). Seal Beach, CA: Biomedical Publications.

Buckley, N. A. (2004). Antipsychotic drugs (neuroleptics). In R. C. Dart (Ed.), *Medical toxicology* (pp. 861–870). Philadelphia, PA: Lippincott Williams & Wilkins.

Cone, E. J., & Huestis, M. A. (2007). Interpretation of oral fluid tests for drugs of abuse. *Annals of the New York Academy of Sciences*, *1098*, 51–103.

Drummer, O. H. (1998). Methods for the measurement of benzodiazepines in biological samples. *Journal of Chromatography*, *713*, 201–225.

Dubowski, K. M. (1980). Alcohol determination in the clinical laboratory. *American Journal of Clinical Pathology*, *74*, 747–750.

Ishihara, K., & Szerlip, H. M. (2004). Anion gap acidosis. *Seminars in Nephrology*, *18*, 83–97.

Jarvie, D. R., Heyworth, R., & Simpson, D. (1987). Plasma salicylate analysis: A comparison of colorimetric, hplc and enzymatic techniques. *Annals of Clinical Biochemistry*, *24*(Pt 4), 364–373.

Jones, G. (2004). Post-mortem toxicology. In A. Moffat (Ed.), *Clarke's analysis of drugs and poisons* (3rd ed.) (pp. 95–108). London, UK: Pharmaceutical Press.

Kadiev, E., Patel, V., Rad, P., Thankachan, L., Tram, A., Weinlein, M., et al. Role of pharmacogenetics in variable response to drugs: Focus on opioids. *Expert Opinion on Drug Metabolism & Toxicology, 4,* 77–91.

Lacy, T. L., & nichols, J. H. (2003). Therapeutic drugs iii: Neuroleptic (antipsychotic) drugs. In B. Levine (Ed.), *Principles of forensic toxicology* (2nd ed.) (pp. 315–325). Washington DC: AACC Press.

Langman, L. J., Bechtel, L. K., Meier, B. M., & Holstege, C. (2018). Chapter 41: Clinical toxicology. In N. Rifai, A. R. Horvath, & C. T. Wittwer (Eds.), *Tietz textbook of clinical chemistry and molecular diagnostics* (6th ed.) (pp. 832–887). St Louis, MO: Elsevier.

Middleberg, R. A., & Langman, L. J. (2015). Chapter 16. Clinical aspects of alcohol testing. In Y. H. Caplan, & B. A. Goldberger (Eds.), *Garriott's medicolegal aspects of alcohol* (6th ed.) (pp. 435–448). Tucson, AZ: Lawyers & Judges Publishing Company Inc.

Moore, K. A. (2003). Amphetamines/sympathomimetic amines. In B. Levine (Ed.), *Principles of forensic toxicology* (2nd ed.) (pp. 245–264). Washington DC: AACC Press.

Moore, C., Negrusz, A., & Lewis, D. (1998). Determination of drugs of abuse in meconium. *Journal of Chromatography*, *713*, 137–146.

Palmer, R. B. (2009). A review of the use of ethyl glucuronide as a marker for ethanol consumption in forensic and clinical medicine. *Seminars in Diagnostic Pathology*, *26*, 18–27.

Payne, J. P., Hill, D. W., & Wood, D. G. (1968). Distribution of ethanol between plasma and erythrocytes in whole blood. *Nature*, *217*, 963–964.

Ropper, A., & Samuels, M. (2009). Chapter 43. Disorders of the nervous system caused by drugs, toxins, and other chemical agents. In A. Ropper, & M. Samuels (Eds.), *Adams and victor's principles of neurology* (9th ed.). McGraw-Hill Companies, Inc. http://www.accessmedicine.com/content.aspx?aID=3640238.

SAMHSA SAaMHSA. (2012). *The role of biomarkers in the treatment of alcohol use disorders, 2012 revision SAMHSA Advisory*. HHS Publication No. (SMA) 12–4686.

The Centers for Disease Control Guideline for Prescribing Opioids for Chronic Pain—United States, 2016 details the use of urine drug testing (http://www.cdc.gov/mmwr/volumes/65/rr/pdfs/rr6501e1.pdf accessed 06/29/2016)

32

Toxic Elements

Frederick G. Strathmann, Lee M. Blum

OBJECTIVES

1. Identify several elements commonly associated with toxicity.
2. Compare available specimen types for applicability in the assessment of elemental toxicity.
3. Give examples of preanalytical concerns regarding elemental testing.
4. Summarize the key information regarding clinical presentation and treatment for elemental toxicity covered in this chapter.

KEY WORDS AND DEFINITIONS

Antidotal treatment The use of any chemical or physiological procedure to prevent, minimize, or terminate the adverse effects of toxicity.

Decontamination Neutralization or removal of toxic substances.

Neuropathy Damage to a nerve that may impair sensation, movement, or other aspects of health.

Signs of toxicity Objective evidence of toxicity to someone other than the exposed individual.

Symptoms of toxicity Subjective evidence of toxicity experienced by the exposed individual.

Toxicokinetics The study of the time course of xenobiotic absorption, distribution, metabolism, and excretion.

BACKGROUND

Elements have been recognized as toxins for centuries. Some elements are essential for life, but if an individual's exposure exceeds a certain threshold, toxicity may develop. When identified early, disease caused by elemental exposure is readily treatable with good outcomes. Conversely, if exposure is not identified and reduced, serious and sometimes irreparable damage to the nervous, renal, and cardiovascular systems can occur. The laboratory plays a key role in this process, and appropriate specimen collection coupled with accurate analysis can make a major difference in correct diagnosis.

CONTENT

This chapter explores toxic elements and the role of the clinical laboratory in diagnosing and monitoring toxicity associated with exposure. A general overview of diagnostic and treatment options for the poisoned patient is followed by detailed descriptions of 10 elements most commonly associated with toxicity. Each section highlights the following areas: (1) sources of exposure, (2) clinical presentation and treatment, (3) preanalytical and analytical aspects, and (4) regulatory and occupational exposure aspects.

Elements are woven into the very fabric of history. For example, arsenic poisoning was a favored way to dethrone royalty in the Renaissance era, and mercury poisoning was common in 18th-century Europe, where it was associated with the generation of felt from beaver pelts to make the popular top hat and was the origin of the phrase "mad as a hatter." This chapter explores these and other toxic elements with an emphasis on clinical presentation, occupational exposure, and analytical requirements.

PREVALENCE OF ELEMENTAL TOXICITIES

The incidence of elemental poisoning across large populations attributable to arsenic, cadmium, lead, or mercury appears to be on the same scale as the more common inborn errors of metabolism, such as neonatal hypothyroidism and phenylketonuria, and is the same order of magnitude as the incidence of adult-onset hemochromatosis. Screening for these latter diseases is indicated because they are treatable, and treatment significantly reduces long-term morbidity. The same may also be true for elemental toxicities. When identified early, disease caused by elemental exposure is readily treatable with good outcomes. Conversely, if exposure is not identified and reduced, serious and sometimes irreparable damage to the nervous, renal, and cardiovascular systems can occur. Review of the periodic table provides some insight into the determination of a metal's potential toxicity (Fig. 32.1).

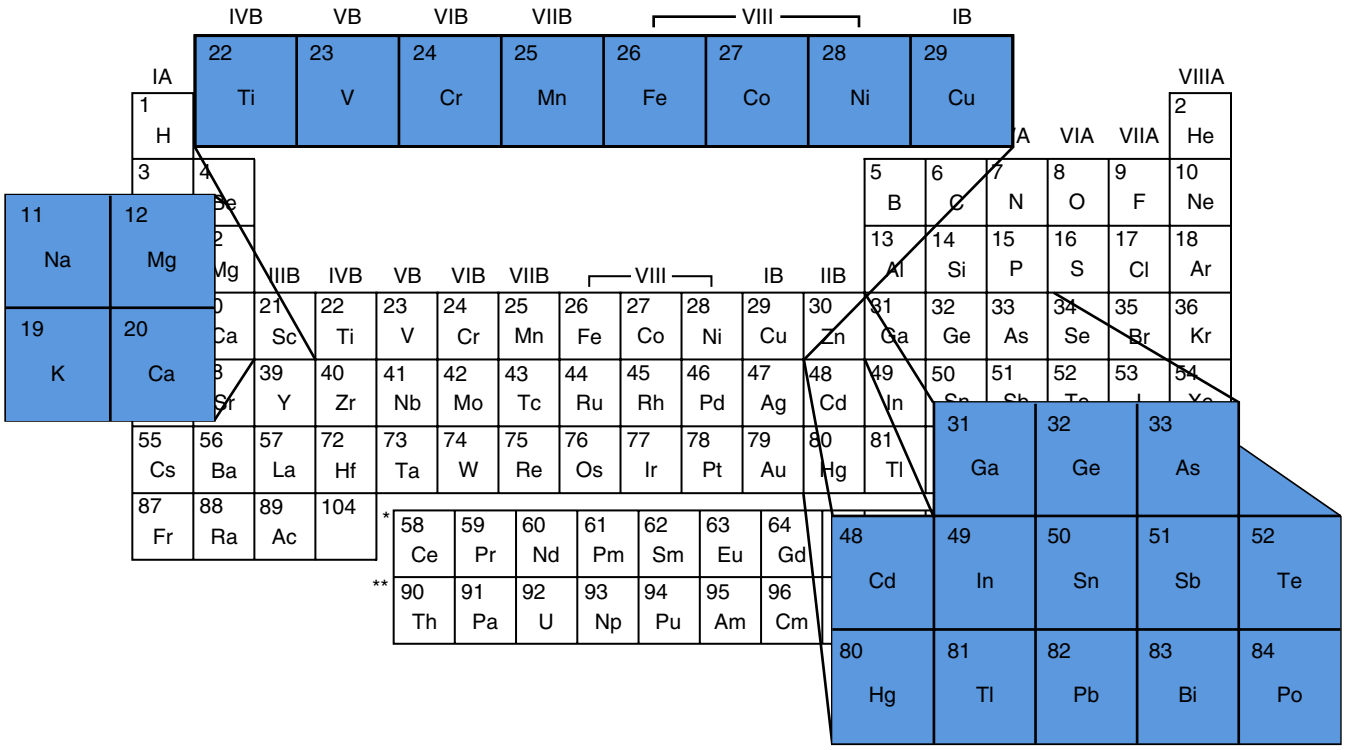


Fig. 32.1 Periodic table, with emphasis on toxic elements.

POINTS TO REMEMBER

- The incidence of elemental poisoning across large populations attributable to arsenic, cadmium, lead, or mercury poisoning appears to be on the same scale as the more common inborn errors of metabolism.
- When identified early, disease caused by elemental exposure is readily treatable with good outcomes.

DIAGNOSING TOXICITY

Confirming the diagnosis of elemental toxicity is difficult because **signs** and **symptoms** are similar to those of many non–element-dependent diseases. Diagnosis of elemental toxicity requires demonstration of all of the following factors: (1) a source of elemental exposure must be evident, (2) the patient must demonstrate signs and symptoms typical of the element, and (3) abnormal element concentration in the appropriate tissue must be evident. If one of these features is absent, a conclusive diagnosis of elemental toxicity cannot be made. The laboratory plays a key role in this process and appropriate specimen collection coupled with accurate analysis can make a major difference in correct diagnosis.

In clinical practice, analysis of toxic elements should always be considered in the clinical workup of the patient with (1) renal disease of unexplained origin, (2) bilateral peripheral **neuropathy**, (3) acute changes in mental function, (4) acute inflammation of the nasal or laryngeal epithelium, or (5) a history of exposure. Certain elements should be considered as the active, causative, or deficient agent in specific circumstances.

TREATMENT

The route of exposure is an important aspect in elemental toxicity that can influence treatment and clinical outcomes. The clinical approach to elemental exposure comprises one of the following treatments: (1) removal from exposure if the source is known, (2) **decontamination**, (3) enhanced elimination, or (4) **antidotal treatment**.

Removal from the source of exposure can be complicated if the element responsible is unknown or the source of exposure is in question. For less extreme cases of exposure or if evidence to support more invasive treatments is lacking, removal, often in conjunction with supportive care, may be the only form of treatment available.

Decontamination strategies may include gastric lavage, activated charcoal, and supportive therapies to maintain the ABCs (airway, breathing, blood circulation), as well as attention to water, electrolyte balance, and neurologic complications. Prevention of absorption in the gastrointestinal (GI) tract is one mechanism for treating the chronic copper overload seen in patients with Wilson disease.

Enhancement of elimination can be done through diuresis, biliary excretion, hemodialysis, or exchange transfusion. Distribution of the element, once absorbed, has an impact on the utility of each of these approaches, and for several elements there is a lack of supporting data demonstrating improved outcomes after their use.

Antidotal treatments range from highly specific (e.g., deferoxamine for iron and aluminum) to considerably less specific (e.g., ethylenediaminetetraacetic acid [EDTA] for

most divalent cations). The intent is to counteract the detrimental effects of the toxic element or to sequester it in a less toxic form. Although it is a highly useful method for treating elemental toxicity, the downside to chelation therapy is the potential for depletion of essential elements. The five most commonly used chelators are dimercaprol (British anti-Lewisite [BAL]), dimercaptosuccinic acid (DMSA or succimer), dimercaptopropane sulfonate (DMPS), EDTA, and D-penicillamine.

POINTS TO REMEMBER

Diagnosing Elemental Toxicity

- Confirming the diagnosis of elemental toxicity is difficult because signs and symptoms are similar to those of many non–element-dependent diseases.
- The route of exposure is an important aspect of elemental toxicity that can influence treatment and clinical outcomes.

SPECIFIC ELEMENTS

Aluminum (Al)

In 1972, Alfrey and colleagues first described an encephalopathy in patients undergoing prolonged hemodialysis for renal failure. The disease was characterized by abnormal speech, myoclonic jerks, and seizures. Patients with these signs also showed a predominance of osteomalacic fractures. Subsequently, it was found that exposure of patients to (1) Al-laden dialysis water, (2) Al-containing oral phosphate binders, and (3) Al-laden albumin administered during dialysis are the primary causes of these signs of Al toxicity. Al is also a developmental toxicant if administered parenterally.

Sources of Exposure

Under normal physiologic conditions, the usual daily dietary intake of Al is 5 to 10 mg, which is principally excreted by the kidney. Patients in renal failure lose this ability and are candidates for Al toxicity. The dialysis process is not highly effective at eliminating Al and can also be a significant source of exposure. Furthermore, it is common practice to administer Al-based gels orally to patients in renal failure to reduce the amount of phosphate absorbed from their diet to avoid excessive phosphate accumulation. A small fraction of this Al may be absorbed, and patients in renal failure can accumulate this. After dialysis, the albumin removed during dialysis may be replaced with albumin products that have high Al content resulting from the pharmaceutical purification process of passing the product through Al silicate filters.

Clinical Presentation and Treatment

A biochemical profile that characterizes Al overload disease has been defined. In human subjects with normal renal function, serum Al concentration is typically lower than 6 μg/L (0.2 μmol/L), but patients in renal failure invariably have significantly higher serum Al concentrations. Clinical guidelines published in 2006 suggest that patients with no signs or symptoms of osteomalacia or encephalopathy are likely to have serum Al concentrations less than 20 μg/L (0.74 μmol/L) and intact parathyroid hormone (PTH) concentrations of 150 to 300 ng/L (16.5 to 33 nmol/L), typical for secondary hyperparathyroidism associated with renal failure. Patients with signs and symptoms of osteomalacia or encephalopathy typically have serum Al concentrations greater than 60 μg/L (2.2 μmol/L) and PTH concentrations less than 65 ng/L (7.15 nmol/L), indicative of Al-related bone disease. Patients with serum Al concentrations greater than 20 and less than 60 μg/L (>0.7 and <2.2 μmol/L) were identified as candidates for likely onset of Al-related bone disease. These patients required aggressive efforts to reduce their daily Al exposure, including (1) switching from Al-containing phosphate binders to calcium-containing phosphate binders, (2) ensuring that dialysis water contains less than 10 μg/L (0.4 μmol/L) of Al, and (3) ensuring that albumin used during postdialysis therapy is Al-free.

Al-related bone disease has been diagnosed and treated with deferoxamine, an avid chelator of both iron and Al. The deferoxamine infusion test is useful for the ultimate diagnosis of Al overload disease, and the drug has demonstrated utility for treating acute Al overload.

Adverse health effects in workers occupationally exposed to Al compounds have included pneumoconiosis (aluminosis) and fibrosis of the lungs, and toxicity of the central nervous system, including balance disorders, difficulties with memory and concentration, impaired cognitive abilities, irritability, depression, and decreased psychomotor performances. Although rare, contact sensitization of the skin by Al is also possible.

Preanalytical and Analytical Aspects

Preanalytical considerations related to collection of specimens for Al analysis are significant. Most of the common evacuated blood collection devices used in phlebotomy today have rubber stoppers made of Al silicate. Puncture of the rubber stopper for blood collection is sufficient to produce an abnormal concentration of Al due to contamination. Typically, blood collected in standard evacuated blood tubes will be contaminated by 20 to 60 μg/L (0.7 to 2.2 μmol/L) of Al; this is readily demonstrated by collecting blood from a healthy volunteer into a standard evacuated phlebotomy tube. Special evacuated blood collection tubes from commercial suppliers are required for Al testing to avoid sample contamination, which leads to misinterpretation and misdiagnosis.

Analysis of Al is routinely performed by inductively coupled plasma-mass spectrometry (ICP-MS). Alternatively, atomic absorption spectrometry with electrothermal atomization may be employed, but considerable attention must be paid to matrix interferences.

Regulatory and Occupational Exposure Aspects

Occupational exposure to Al compounds primarily occurs through inhalation of airborne particles in dusts and fumes.

The solubility of the Al compounds affects its **toxicokinetics** and adds to subsequent health risks. In general, inhaled soluble particles (e.g., aluminum sulfate, hydrated aluminum chloride, and aluminum nitrate) are rapidly absorbed through the lungs, whereas the less or sparsely soluble particles (e.g., aluminum metal, aluminum oxide, aluminum hydroxide, aluminum phosphate, and aluminum silicate) are retained in the lungs and then slowly released into the systemic circulation. The amount of Al absorbed by inhalation also appears to be related to particle size. Al fumes are absorbed more readily than dust particles, and ultrafine nanoparticles of aluminum oxide can penetrate cell membranes.

Biological monitoring of occupational exposure to Al can be determined by analysis of blood, serum, or urine. The concentration of Al in these specimens can be affected by both the intensity of a recent exposure and the Al body burden. Studies of aluminum welders indicate a fair correlation between Al in serum and urine; however, in workers with normal kidney function, urinary Al concentrations are a more sensitive indicator of occupational Al exposure than serum Al levels. After exposures to low-level airborne Al, urinary Al levels are increased, whereas serum concentrations are generally within the range of control subjects. In chronically exposed welders, the Al concentration in urine collected after 1 or 2 exposure-free days is likely a good indicator for determining the magnitude of the Al stored in the body; whereas in newly exposed workers, the urinary Al level is probably more influenced by recent exposures than by the body burden.

Antimony (Sb)

Antimony compounds have been known since ancient Egyptian times and were used as cosmetics by the women of that era. In the 16th century, Sb preparations were thought to be wonder drugs and in the 19th century were prescribed for various conditions.

Sources of Exposure

Pure metallic Sb is very brittle; however, alloys of Sb are used in various fields of technology. For example, addition of Sb to lead, tin, and copper increases the hardness of these elements when used as electrodes, bullets, type metal for printing, and ball bearings. Other uses include fire-resistant chemicals, pigments, and dyes.

Before the introduction of the anthelmintic drug praziquantel, several Sb-containing compounds were used in the treatment of leishmaniasis, a parasitic disease in which protozoa of the genus *Leishmania* are spread by sand flies, and schistosomiasis, a disease carried by freshwater snails infected with one of the five species of the parasitic flatworm *Schistosoma*.

Clinical Presentation and Treatment

Symptoms of acute exposure include a metallic taste, headache, nausea, dizziness, and, after a short interval, vomiting, diarrhea, and intestinal spasms. In chronic intoxication, adverse health effects include cardiac arrhythmias, upper respiratory tract and ocular irritation, spontaneous abortion, premature birth, and dermatitis. Lymphocytosis, eosinophilia, and a reduction in leukocyte and platelet counts are also seen, and indicate liver and spleen damage. Evidence supports increased risk for the development of lung cancer in Sb smelter workers, but the increased risk may be multifactorial, and may be due to, among other factors, the presence of arsenic in the work environment. It is important to remember that when intoxication occurs with metallic Sb, the effect is caused not only by Sb but also by the lead, arsenic, and other elements that may accompany it.

Preanalytical and Analytical Aspects

Careful selection of specimen tube type, even for certified trace element–free products, is important for the accurate measurement of Sb in blood. ICP-MS is typically used for Sb detection because of its increased sensitivity over that of other instrumentation. Historic methods for Sb measurements in blood have used time-consuming heated acid digestion for achieving better sensitivity, while simplified sample preparation methods have been published more recently.

Regulatory and Occupational Exposure Aspects

For purposes of biological monitoring after exposure to Sb, urine samples are preferable over blood. In one study, urinary Sb concentrations in workers were correlated to the intensity of airborne exposure in nonferrous smelters producing antimony pentoxide and sodium antimonite ($r = 0.86$).

Arsenic

Arsenic (As) is perhaps the best known of the elemental toxins, having gained notoriety from its extensive use by Renaissance nobility as an antisyphilitic agent and an antidote against acute As poisoning. Long-term administration of low As doses protects against acute poisoning by massive doses—a historic example of hepatic enzyme induction. This agent was memorably used in the well-known tale *Arsenic and Old Lace* as a means of terminating undesirable acquaintances. Arsenic is listed as the number 1 toxicant on the Agency for Toxic Substances and Disease Registry (ATSDR) 2013 Substance Priority List of Hazardous Substances based on a combination of frequency, toxicity, and potential for human exposure. Despite their inherent toxicity, As-containing compounds are used for therapeutic reasons.

Sources of Exposure

Nontoxic forms of As are present in many foods with arsenobetaine and arsenocholine, the two most common forms. The foods that most commonly contain significant concentrations of organic As are shellfish and other predators in the seafood chain (e.g., cod, haddock). In a large US population study, for all participants over 6 years of age, the inorganic As metabolite dimethylarsinic acid (DMA) and arsenobetaine had the greatest contribution to the total urinary As output, with arsenobetaine being the primary contributor.

Clinical Presentation and Treatment

The symptoms of As toxicity may be nonspecific and often overlap with symptoms of other toxicants. Acute As toxicity

can be characterized by GI distress, including vomiting and diarrhea, and cardiac arrhythmias. Chronic toxicity may be characterized by renal failure, cardiac arrhythmias, liver dysfunction, and peripheral neuropathy, and As is a known carcinogen. Transverse white bands on the fingernails (Mees lines), hypopigmented macules, hyperpigmentation, and hyperkeratosis have been documented after As exposure.

Arsenic has been shown to interfere with the activity of several heme-biosynthetic enzymes, including aminolevulinate synthase, porphobilinogen deaminase, uroporphyrinogen III synthase, uroporphyrinogen decarboxylate, coproporphyrinogen oxidase, ferrochelatase, and heme oxygenase with resultant increases in copro/uro and copro I/III ratios. Interestingly, the madness of King George III, historically attributed to acute hereditary porphyria, has been hypothesized to instead be a case of arsenic-induced porphyria based on retrospective analyses of hair.

An effective antidote for treating As intoxication is BAL; the active agent in BAL is dimercaprol, a sulfhydryl-reducing agent. BAL was originally developed during World War II in response to the use of Lewisite, an As-based chemical warfare agent. Other dimercaprol derivatives are available for treatment, including DMSA (succimer) and dimercap-to-1-propanesulfonic acid (DMPS), although the latter is not currently approved for As chelation therapy in the United States.

Preanalytical and Analytical Aspects

To distinguish among toxic inorganic species, high-performance liquid chromatography (HPLC) techniques for evaluating biological fluids and tissues have been developed. In general, a 24-hour urine specimen with measurable arsenic will have 95% present as the organic, nontoxic species and less than 5% present as the inorganic toxic species. Despite the availability of HPLC-ICP-MS methods for As speciation, it is noted that the use of a screening method for total As before speciation is of outstanding utility in reducing unnecessary costs.

Hair analysis is frequently used to document time of As exposure. Circulating in the blood, it will bind to protein by formation of a covalent complex with sulfhydryl groups of the amino acid cysteine. Because As has a high affinity for keratin, which has high cysteine content, the As concentration in hair or nails is greater than in other tissues. Several weeks after exposure, transverse white striae, called Mees lines, may appear in the fingernails, caused by denaturation of keratin by elements such as As, Cd, Pb, and Hg. Because hair grows at a rate of approximately 1 cm/month, hair collected from the nape of the neck can be used to document recent exposure. Axillary or pubic hair is used to document long-term (6 months to 1 year) exposure. Hair As greater than 1 μg/g dry weight indicates excessive exposure, though reports vary for results determined in hair due to analytical, environmental, and dietary factors. In one study, the highest hair As observed was 210 μg/g dry weight in a case of chronic exposure that was the cause of death.

The body treats As like phosphate, incorporating it wherever phosphate would be incorporated. Absorbed As is rapidly circulated and distributed into tissue storage sites. Blood is the least useful specimen for identifying As exposure. Blood As concentrations are elevated for only a short time after administration and rapidly disappear into the large body phosphate pool. Abnormal blood As concentrations are detected for only a few hours (<4 hours) after ingestion. This test is useful only to document an acute exposure when the As is likely to be greater than 20 ng/mL (0.3 μmol/L) for a short period. Typically, serum As is less than 40 ng/mL (0.5 μmol/L).

Arsenic has been accurately measured by ICP-MS. Mass response from the argon plasma is monitored for As (mass-to-charge ratio $[m/z] = 75$); however, the method must reduce the potential for interference from argon chloride ($m/z = 75$) with the use of a dynamic reaction cell or collision cell with kinetic energy discrimination. For more details on the ICP-MS technology, refer to Chapter 13. Urine is the sample of choice for As analysis because As is excreted predominantly by the kidney.

Regulatory and Occupational Exposure Aspects

Cardiovascular diseases and cerebrovascular effects are associated with chronic exposures to inorganic As compounds in the workplace. Inhalation exposures to inorganic As compounds have been correlated to an increased incidence of lung cancer. The American Conference of Governmental Industrial Hygienists (ACGIH) classifies As and its inorganic compounds as A1 carcinogens (confirmed human carcinogens). The recommended ACGIH Biological Exposure Index (BEI) for inorganic As plus methylated metabolites is less than 35 μg/L (0.5 μmol/L) in an end-of-workweek urine sample after the exposure to As and soluble inorganic compounds (excluding gallium arsenide and arsine). After exposures to poorly soluble inorganic compounds, such as gallium arsenide, the urinary As concentration is more representative of the amount absorbed rather than the total dose inhaled or ingested.

Cadmium (Cd)

An appreciation for the toxicity of Cd extends back to 1858 and the observed GI and respiratory effects after exposure to Cd-containing polishing agents. Famously, Cd toxicity was at the center of itai-itai disease (Japanese for "Ouch-Ouch"), a bone disease associated with fractures and severe pain that was identified after World War II in Japan. No biological function has been identified for Cd in humans; however, Cd serves as a metal cofactor of Cd-carbonic anhydrase in diatoms.

Sources of Exposure

Cd is a by-product of zinc and lead smelting. It is used in industry in electroplating, in the production of rechargeable batteries, as a common pigment in organic-based paints, and in tobacco products. Spray painting of organic-based paints without the use of a protective breathing apparatus is a common source of chronic exposure. Auto repair mechanics are a work group that has significant opportunity for exposure to Cd.

Clinical Presentation and Treatment

Chronic exposure to Cd causes accumulated renal damage. Breathing the fumes of Cd vapors leads to nasal epithelial

deterioration and pulmonary congestion resembling chronic emphysema. Renal dysfunction with proteinuria of slow onset (over a period of years) is the typical presentation. Normal blood Cd concentration is less than 5 ng/mL (44 nmol/L), with most concentrations in the interval of 0.5 to 2 ng/mL (4 to 18 nmol/L). Moderately increased blood Cd (3 to 7 ng/mL or 27 to 62 nmol/L) may be associated with tobacco use. Acute toxicity is observed when the blood concentration exceeds 50 ng/mL (444 nmol/L). Usual daily excretion of Cd is less than 3 μg/day (0.03 μmol/day). Cd concentrations also increase with age and may be involved with senescence.

Preanalytical and Analytical Aspects

Collection of urine samples using a rubber catheter has been known to result in elevated results because rubber contains trace amounts of Cd that are extracted as urine passes through it. Brightly colored plastic urine collection containers should be avoided because the pigment in the plastic may be Cd-based. Cd has historically been quantified by atomic absorption spectrometry, and more recently by ICP-MS.

Regulatory and Occupational Exposure Aspects

Inhalation is the primary route of Cd exposure in the occupational setting, and the amount absorbed from the lungs depends on the solubility of the Cd compound and particle size. Absorption through the GI tract can occur from the clearance of particles deposited in the lungs and from contaminated hands and food. After chronic low-level exposures, approximately half of the Cd body burden is stored in the liver and kidneys. Urine is the main route of elimination, and thus urinary Cd levels are widely used to biomonitor chronic exposures in workers. However, because of the high binding capacity for Cd in the body, a urine level may provide no information about exposure in someone newly exposed to Cd. The lag time needed before the urinary Cd concentration correlates with exposure, depends on the intensity of the integrated exposure. Until sufficient time of chronic Cd exposure has passed before urine testing can be used appropriately (~1 year), blood testing is suggested for newly exposed workers or when changes in exposure occur, because blood levels primarily reflect recent exposures.

The exposure assessment of Cd in the US Occupational Safety and Health Administration (OSHA) Cadmium Standard is part of an overall periodic environmental, medical, and biological monitoring program. Biomonitoring requires testing of urine for both Cd and β_2-microglobulin with standardization to grams of creatinine for each component (μg/g creatinine) and of blood for Cd with standardization to liters of whole blood (μg/L).

Chromium

Chromium (Cr), from the Greek word *chroma* ("color"), makes rubies red and emeralds green. Among its many uses, it is best known for its application in making stainless steel, which is resistant to corrosion and discoloration. The requirement of Cr for sugar and lipid metabolism has resulted in considerable debate on its potential role in insulin resistance. Of late, Cr continues to make headlines as one of several metal ions released with wear because of highly publicized recalls of misaligned and accelerated failure rates of specific metal-on-metal (MoM) prosthetics.

Sources of Exposure

Occupational exposure to Cr represents a significant health hazard. Cr is used extensively (1) in the manufacture of stainless steel, (2) in chrome plating, (3) in the tanning of leather, (4) as a pigment in paints and dye for printing and textile manufacture, (5) as a cleaning solution, and (6) as an anticorrosive in cooling systems. The toxic form of Cr is hexavalent Cr^{6+} (Cr[VI]), and a strong oxidizing environment is required to convert the common form trivalent Cr^{3+} (Cr[III]) to Cr^{6+}. This can occur when Cr^{3+} is exposed to high temperatures in the presence of oxygen or during high-voltage electroplating.

Considerable attention has been given to Cr toxicity after release of Cr ions during normal and abnormal wear of MoM prosthetics. However, the vast majority of Cr released from MoM prosthetics has been shown to be in the trivalent form, which is considerably less toxic than hexavalent Cr. In May 2011, the US Food and Drug Administration (FDA) issued orders for postmarket surveillance studies to manufacturers of MoM hip replacement systems in response to increasing concern over failed implant devices reported in Europe. In the United States, no guidelines have been established for the assessment of metal ions in asymptomatic patients because of a lack of knowledge regarding the prevalence of adverse events in the US population and no clear threshold levels associated with an adverse event.

Clinical Presentation and Treatment

Symptoms associated with Cr toxicity vary based on the route of exposure, Cr species, and dose, and may include dermatitis, impairment of pulmonary function, gastroenteritis, hepatic necrosis, bleeding, and acute tubular necrosis. In the case of failing MoM prosthetics, elevated Cr in serum or blood is rarely the initial finding and instead is preceded by more telling physical symptoms such as reduced range of motion, swelling and inflammation around the joint, and general discomfort or pain. Although offered by many laboratories, measurement of Cr in joint fluid has relatively little clinical utility.

Preanalytical and Analytical Aspects

Use of a plastic cannula for blood sampling was shown to be unnecessary in the assessment of Cr; however, sporadic contamination from stainless steel needles was observed. Quantification of total Cr in urine can be used to assess exposure to all forms of Cr. The presence of Cr in erythrocytes is suggestive of exposure to Cr^{6+} within the past 120 days, because Cr^{6+} crosses biological membranes but Cr^{3+} does not. Increased serum Cr concentrations are observed in association with orthopedic implants made from Cr alloys. ICP-MS is the preferred technology for quantification of Cr in body fluids but suffers from considerable interference because of polyatomics. Use of dynamic reactive cell technology or a

collision cell with kinetic energy discrimination is required for reproducible and accurate measurements.

Regulatory and Occupational Exposure Aspects

Industrial Cr exposures can involve trivalent and hexavalent forms, which exhibit different toxicokinetics and toxicities in the body. Urine Cr levels are the most useful biomarker for assessing occupational exposure to water-soluble hexavalent Cr compounds; however, other nonoccupational sources of both trivalent and hexavalent Cr from the diet, supplements, and the environment can contribute to the total Cr concentration in urine. Furthermore, because Cr has been shown to accumulate in the body, urine Cr levels can be affected by both recent and past workplace exposures. The measurement of Cr in erythrocytes has been used to gauge the intensity of exposure to hexavalent Cr; however, insufficient data are available to determine a relationship between erythrocyte Cr levels and the risks associated with exposures.

Occupational exposures to hexavalent Cr compounds can cause respiratory irritation and tissue damage of the nose, throat, and lungs after inhalation, and can cause irritation, burns, and ulcers on the skin with contact. Hexavalent Cr has been associated with cancers of the lungs, nose, and nasal sinuses, and is classified as a human carcinogen by the Internal Agency for Research on Cancer. Hexavalent Cr compounds are categorized as human carcinogens by the ACGIH (A1) and the German Research Foundation (DFG) in Germany (category 1). The ACGIH suggests testing total Cr in urine for workplace exposure to hexavalent Cr (water-soluble fumes).

Cobalt

The word cobalt (Co) is derived from the German *kobalt* meaning "goblin" because of its superstitious reputation among miners. Cobalt is widely distributed in the environment, is the essential cofactor in vitamin B_{12}, and has been widely used throughout history for decorative accents to glass and ceramics.

Sources of Exposure

Although relatively rare, Co is widely distributed in nature and is found in green vegetables, animal foods, and seafood. Of the Co in the body, 85% is in the form of cyanocobalamin or vitamin B_{12}, a Co complex similar in structure to that of porphyrins. Cobalt is used in alloys in which the presence of Co results in high melting point, strength, and resistance to oxygen. The addition of Co in MoM prosthetics adds resistance to surfaces undergoing heavy wear, and the release of Co from failed prosthetics remains of outstanding concern. Cobalt is found in rechargeable batteries and provides color to plastics, inks, glass, and paint.

Clinical Presentation and Treatment

Cobalt is not highly toxic, but large exposures will produce pulmonary edema, allergy, nausea, vomiting, hemorrhage, and renal failure. Occupational exposure occurs during production and machining of metal alloys and has been known to result in interstitial lung disease. Cardiomyopathy and renal failure are symptomatic of acute Co exposure; this was exemplified by an incident of mass population exposure to Co when beer contaminated with cobalt salts was consumed. Chronic exposure may cause pulmonary syndrome, skin irritation, allergy, GI irritation, nausea, cardiomyopathy, hematologic disorders, and thyroid abnormalities. Cobalt exposure alone may not lead to toxicity and must be considered within the context of exposure to multiple elements.

Serum Co concentrations are increased above normal (>1 μg/L or >17 nmol/L) in patients with well-positioned and functioning orthopedic implants made from Co alloys. Considerable attention has been given to Co toxicity in the context of failed MoM prosthetics, with several case reports of severe neurologic and cardiac abnormalities described. In the case of failing MoM prosthetics, elevated Co in serum or blood is rarely the initial finding and instead is preceded by more telling physical symptoms such as reduced range of motion, swelling and inflammation around the joint, and general discomfort or pain. Although offered by many laboratories, measurement of Co in joint fluid has relatively little clinical utility.

Preanalytical and Analytical Aspects

Quantification of active vitamin B_{12} is the usual way to assess nutritional status; quantification of serum, blood, or urine Co concentration is not typical for assessing vitamin B_{12} status. Cobalt deficiency has not been reported in humans. Quantification of Co in urine or biological specimens is accomplished by atomic absorption spectrometry or by ICP-MS.

Regulatory and Occupational Exposure Aspects

Absorption of Co through the lungs from inhalation of dust and fumes is the primary route of exposure in occupational settings, but absorption from dermal contact can occur. Pulmonary absorption of the inhaled Co depends on the particle size and its solubility, with smaller particles deposited in the lower respiratory tract, where they dissolve or are phagocytized and translocated. After inhalation, Co is eliminated from the body in several phases with a relatively rapid initial phase having a half-life ranging from hours to a few days and subsequent slower phases with half-lives ranging from months to years. Cobalt concentrations in urine increase proportionally more than in the blood of workers during Co exposures; Co concentrations in urine are primarily affected by recent exposures, but increase through the workweek under stable exposure conditions.

The skin and respiratory tract are the two main target organs in workers exposed to cobalt metal, cobalt salts, and cobalt-containing dusts. Contact of Co with the skin can cause mild irritation and allergic reactions. Inhaled cobalt dust and fumes can cause shortness of breath. Respiratory symptoms after chronic inhalation of Co can range from cough to respiratory hypersensitivity, progressive dyspnea, decreased pulmonary function, permanent disability, and death. Pathogenic lesions in the parenchymal regions of the lungs, known as "hard metal disease," can lead to severe

alveolitis that progresses to end-stage pulmonary fibrosis. In the hard metal industry, the coexposure to tungsten and Co has proved more toxic than metallic Co alone.

Lead

Lead (Pb) has been used extensively throughout history, but use has been drastically reduced in the last few decades as a result of a better appreciation for its toxicity, especially during early childhood development.

Sources of Exposure

Lead is present at high concentration (up to 35% w/w) in many paints manufactured before 1972. The Pb content of paints intended for household use was limited to less than 0.5% in 1978, but Pb is still found in paint products intended for non-domestic use and in artists' pigments. Ceramic products for use in homes available from noncommercial suppliers (e.g., local artists) have been known to contain significant amounts of Pb; Pb is leached from the products by weak acids such as vinegar and fruit juices. Leaded crystal contains up to 10% Pb, which is leached during long-term storage of fluids such as fruit juice, wine, and spirits. Lead is also found in dirt from areas adjacent to homes painted with Pb-based paints, in soil near abandoned industrial sites, and on highways where it has accumulated from the use of leaded gasoline in automobiles. Use of leaded gasoline has diminished significantly since the introduction of unleaded gasoline, which has been required in personal automobiles in the United States since 1978. Water transported through Pb or Pb-soldered pipes contains some Pb, with higher concentrations found in water that is weakly acidic. Some foods (e.g., moonshine distilled in Pb pipes) and some traditional home medicines also contain Pb. Exposure to Pb from any of these sources by ingestion, inhalation, or dermal contact has been known to cause significant toxicity.

Clinical Presentation and Treatment

The development of Pb toxicity follows a progressive pattern. Fig. 32.2 describes this progression through a series of symptoms.

The finding that Pb contributes significantly to decreased intellectual capability in the very young is of particular concern. Young children are particularly prone to the effects of Pb because they have greater opportunity for exposure. Children tend to spend a lot of time on the floor. In older homes that have been previously treated with Pb-based paints, Pb-laden paint chips and dust accumulate on the floor, which children are likely to ingest. The Centers for Disease Control and Prevention (CDC) has continued to lower the suggested normal limit for children and is now recommending 5 μg/dL (0.24 μmol/L) based on the 97.5% percentile of nonexposed children. In addition, the term *blood lead level of concern* has been abandoned in favor of the viewpoint that all Pb is potentially harmful and therefore unsafe.

The World Health Organization (WHO) has defined blood Pb concentrations greater than 30 μg/dL (1.5 μmol/L) in adults as indicative of significant exposure. Pb concentrations above 60 μg/dL (2.9 μmol/L) require chelation therapy.

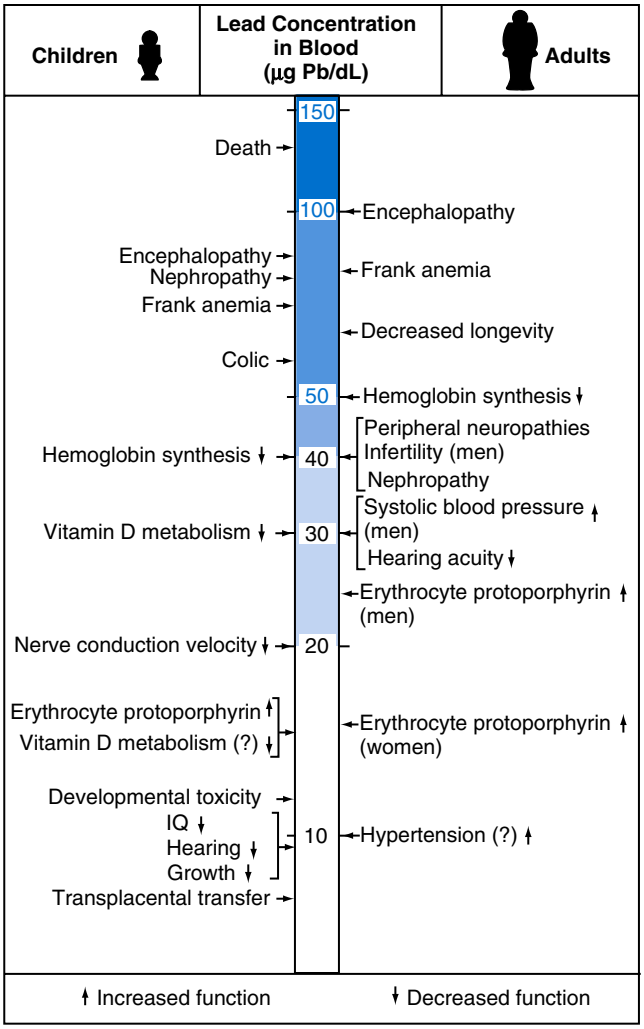


Fig. 32.2 Effects of inorganic lead on children and adults (lowest observable adverse effect concentrations). (Reprinted with permission from Royce, S. E., & Needleman, H. L., [Eds.], *Case studies in environmental medicine: lead toxicity.* Washington, DC: US Public Health Service, Agency for Toxic Substances and Disease Registry; 1990.)

Similar to the situation seen in children, adult blood Pb concentrations have dropped to a mean value of 1.4 μg/dL (0.07 μmol/L) for ages 20 to 49 and a mean value of 1.9 μg/dL (0.09 μmol/L) for ages 50 to 69.

Avoidance of continued exposure to Pb is paramount when blood Pb concentrations exceed acceptable reference intervals. Oral dimercaprol has become a standard therapy and is being used in the outpatient setting for all except those with the most severe Pb poisoning. Although chelation therapy is effective in reducing blood Pb concentrations, a 2003 study indicated that chelation therapy given to preschool children with Pb concentrations in the range of 20 to 44 μg/dL (1 to 2.1 μmol/L) showed no beneficial effect on tests of cognition or behavior. Thus, prevention is the best therapeutic option.

Preanalytical and Analytical Aspects

The definitive test for Pb toxicity is measurement of blood Pb. Analysis of Pb is routinely performed by ICP-MS,

electrothermal atomic absorption spectrometry, or anodic stripping voltammetry. Because Pb is concentrated in the erythrocytes, EDTA-anticoagulated blood is the specimen of choice for Pb analysis. Sodium heparin may also be used; however, samples that are not analyzed within 48 hours are frequently clotted and must be rejected. Care must be taken when obtaining capillary blood. Surface contamination, insufficient collection volume, or inadequate mixing with EDTA results in frequent sample rejection. Urinalysis can also be performed; urine quantification correlates with exposure.

If ICP-MS is used to measure Pb concentrations, it is essential to sum the masses of 206, 207, and 208 *m/z* to account for the natural isotopic variation of Pb in the environment. Failure to sum masses will skew results above or below the actual concentration because the isotopic abundance of a particular mass in the calibrator might not match the sample. However, this isotopic variation has been exploited to determine the source of Pb exposure. By determining the relative abundances of Pb in blood and of potential sources of exposure (e.g., paint chips, soil), it is possible to identify a matching pattern. The exposure source with the same ratio of major Pb isotopes as that of the blood should then be avoided or removed from the patient's environment.

Regulatory and Occupational Exposure Aspects

Biomarkers of exposure and biomarkers of effect have been used in assessing occupational Pb exposures. The direct measurement of Pb in blood reflects the concentration of Pb in the soft tissues at steady state, and therefore is an indicator of recent exposures. Most of the Pb body burden is stored in bones, so the bone Pb concentrations are a good indicator of cumulative or long-term exposures. X-ray fluorescence techniques are used to measure Pb levels in bone.

Exposures to Pb have been found to cause encephalopathy, peripheral neuropathy, neurologic and neurobehavioral effects, renal effects, hypertension, and reduced fertility. Because of these adverse effects, OSHA has enacted standards to assess inorganic Pb exposure in the workplace. Workers with a single blood Pb level meeting the numerical criteria for medical removal must have their blood Pb level retested within 2 weeks. If a worker is medically removed, a new blood Pb level must be measured monthly during the removal period. Workers are permitted to return to work when their blood Pb level is 40 μg/dL (1.9 μmol/L) or less according to current guidelines.

According to the OSHA Lead Standards, measurement of zinc protoporphyrin (ZPP) is required on each occasion that a blood Pb measurement is made. OSHA recommends a hematocrit determination to exclude anemia whenever a confirmed ZPP of 50 μg/100 mL (0.8 μmol/L) is obtained. Zinc protoporphyrin levels in excess of 100 μg/100 mL (80 μmol/L) are considered elevated. However, ZPP is an insensitive biomarker for blood Pb levels less than approximately 20 μg/dL or 1 μmol/L.

Mercury

Mercury's elemental symbol, Hg, is derived from the Greek word *hydrargyros,* meaning "water silver," and is commonly referred to as quicksilver. Mercury is one of only two elements that are liquid at standard temperature and pressure (the other is bromine). The various forms of Hg have remarkably different toxicity profiles from largely benign unless inhaled (elemental Hg), to highly toxic (Hg salts and organic forms). The phrase "mad as a hatter" was first used in reference to the Hg-induced dementia seen in workers with chronic exposure to Hg in the production of felt for hats.

Sources of Exposure

Mercury is widely found in the environment and occurs both naturally and as the result of industrial processes, with the single largest source of Hg being its natural out-gassing from granite rock. Mercury is also found in deposits throughout the world, mostly as cinnabar (mercuric sulfide), which is the source of the red pigment vermilion.

In the past, Hg was extensively used in the manufacture of devices such as thermometers, barometers, manometers, and sphygmomanometers. However, concerns about its toxicity have resulted in the phasing out of Hg-based instruments, which have been replaced with alcohol-filled, digital, or thermistor-based versions. It is still used as a dental amalgam and in lighting as mercury vapor lamps, although these are being replaced by sodium vapor bulbs. Mercury is used in the pulp and paper industry as a whitener, as a catalyst in the synthesis of plastics, and as a potent fungicide in antifouling and latex paints.

Chemically, it is possible to convert Hg from its elemental state to its ionized state; in industry, this is frequently accomplished by exposing Hg^0 to a strong oxidant, such as chlorine. Elemental Hg is also bioconverted to both Hg^{2+} and alkyl Hg by microorganisms that exist in the normal human gut and in the bottom sediment of lakes and rivers. When Hg^0 enters bottom sediment, it is absorbed by bacteria, fungi, and related microorganisms; these organisms metabolically convert it to Hg^{2+}, CH_3Hg^+, $CH_3CH_2Hg^+$, and similar species. Consequently, the methyl mercurials are accumulated in the aquatic food chain and reach their highest concentrations in predatory fish. The normal bacterial flora present in the mouth converts a fraction of Hg^{2+} to CH_3Hg^+; the latter has been shown to be incorporated into body tissue. As a consequence of accumulation of methylmercury in the aquatic food chain, most human exposure to Hg happens through the eating of contaminated fish, shellfish, and sea mammals. For example, commercially distributed fish considered safe for consumption contain less than 0.3 μg/g, but some game fish contain more than 2 μg/g and, if consumed on a regular basis, contribute to significant body Hg loads.

In the late 1980s, the public became concerned about exposure to Hg from dental amalgams. However, later studies have failed to confirm a causal relationship. Basic to the initial concerns was the fact that restorative dentistry used an Hg-silver amalgam for approximately 90 years as a filling material. In 1989, Hahn and colleagues showed that a small (2 to 20 μg/day) release of Hg^0 from amalgam occurs when it is mechanically manipulated, such as by chewing. In addition, the habit of gum chewing can cause release of Hg from dental

amalgams in even larger amounts. In 2010, the FDA issued rules that classify dental amalgam, reclassify dental mercury, and specify special controls for dental amalgam, mercury, and amalgam alloy.

Concerns have been raised about the possible relationship between Hg exposure from vaccines and autistic disorders. In the United States, the prevalence of autism has risen from 1 in approximately 2500 in the mid-1980s to 1 in approximately 300 children in the mid-1990s. Some investigators believe that this rise occurs because of the Hg that is present in vaccines as the preservative thimerosal (sodium ethyl mercury thiosulfate). However, this causality has been questioned by numerous other studies, which have not been able to confirm this relationship. In 2001, the Committee on Immunization Safety Review of the Board on Health Promotion and Disease Prevention of the Institute of Medicine initiated a study to review the connection between Hg-containing vaccines and neurodevelopmental disorders, including autism. The Committee reported that the hypothesis was biologically plausible, but that evidence was insufficient to accept or reject a causal connection. The findings of this report have been challenged, and thimerosal has been removed from most vaccines in the United States. The evidence linking Hg exposure from vaccines with autistic disorders has been highly criticized and resulted in the journal *Lancet* retracting the article linking autism to measles, mumps, and rubella (MMR) vaccines.

Historically, Hg-containing compounds have been used for therapeutic reasons. For example, Hg has been used in medications that were touted as cures for syphilis and dysentery, to treat constipation, and as diuretics. Intriguingly, the use of Hg-containing laxatives, termed "Thunder Clappers," on the Corps of Discovery Expedition with Lewis and Clark has left a discoverable trail of their movements, in more ways than one!

Clinical Presentation and Treatment

Each form of Hg manifests differently after a toxic exposure. For elemental Hg, poor absorption in the GI system reduces toxicity but can result in GI distress with prolonged contact or deposition. With repeated injections of elemental Hg, seen in suicide attempts or psychiatric patients, the presentation was predominantly pleuritic chest pain, with other classic Hg-associated symptoms being absent. Common target organs with inorganic Hg poisoning include the GI system, kidneys, and brain.

Experience with Hg poisoning has been gained from investigation of the 1951 to 1963 industrial dumping of Hg-laden waste sludge into Minamata Bay, Japan. Fish in Minamata Bay became heavily laden with Hg through the food chain. The local human population, whose diet depended on fish from the bay, exhibited symptoms of methylmercury poisoning including ataxia, impaired speech, visual field constriction, hearing loss, and somatosensory change, characterized histologically by cerebral cortex necrosis; this is collectively known as Minamata disease.

In adults, cases of methylmercury poisoning are characterized by the focal degeneration of neurons in regions of the brain such as the cerebral cortex and the cerebellum. Depending on the degree of in utero exposure, methylmercury may result in effects ranging from fetal death to subtle neurodevelopmental delays. Consequently, because pregnant women, women of child-bearing age, and young children are particularly at risk, the FDA recommends that they avoid eating shark, swordfish, mackerel, and tilefish.

Preanalytical and Analytical Aspects

Analysis of blood, urine, and hair for Hg concentrations is used to determine exposure. The quantity of Hg found in blood and urine correlates with the degree of toxicity, although the form of Hg can negatively affect the reliability of the relationship. Hair analysis has been used historically to document the time of peak exposure. However, it should be noted that hair analysis for elements in general is difficult because of contamination. Normal whole-blood Hg concentration is usually lower than 10 μg/L (49 nmol/L). Treatment with BAL or penicillamine will mobilize Hg, allowing for its excretion in the urine. Therapy is usually monitored by following urinary excretion of Hg; therapy may be terminated after the daily urine excretion rate falls below 50 μg/L (249 nmol/L).

Regulatory and Occupational Exposure Aspects

In industry, elemental Hg is present in the air as a vapor and its inorganic salts are present as aerosols. Inhalation is the primary route of uptake from occupational exposures to Hg, although dermal absorption can also occur. Once absorbed, some elemental Hg is oxidized to the divalent form, which can bind to sulfhydryl groups in proteins such as albumin and metallothionein. Elemental Hg is mainly distributed to the kidney and brain, whereas inorganic Hg salts are mostly delivered to the kidney. Mercury is eliminated primarily through the feces and urine. Initially after exposure, more Hg may be eliminated in the feces than in the urine. This may be due to the presence of binding sites in the kidneys. The half-life of Hg in urine is approximately 40 days. Therefore the Hg levels in urine do not necessarily reflect recent exposures in newly exposed workers.

Urinary Hg concentrations are primarily used to monitor long-term exposures to elemental Hg and its inorganic salts, whereas blood Hg concentrations are mainly useful for short-term, higher-level exposures of these compounds. Urine is the preferred specimen for biological monitoring of workplace exposure to elemental Hg and its inorganic compounds, because Hg levels in blood can be affected by the presence of organic Hg compounds from the consumption of dietary fish. Because biological monitoring in urine reflects long-term Hg exposures, the airborne exposure levels and urinary Hg concentration can be correlated, provided the exposures are relatively constant over a sufficiently long period, sample collection is standardized, and biomonitoring is initiated at least 6 months after exposure begins.

Nickel

Found in coins since the mid-19th century, nickel (Ni) remains valuable for its corrosion resistance in plating, but is

often found at the center of reports involving hypersensitivity and allergic reactions to a wider range of products from cell phones to belt buckles.

Sources of Exposure

Ni is frequently used in the production of metal alloys (which are popular for their anticorrosive and hardness properties), in Ni-based rechargeable batteries, and as a catalyst in the hydrogenation of oils. Ni alloyed with transition metals is considered nontoxic, except that it will induce inflammation at the point of contact. Nickel carbonyl ($Ni[CO]_4$), used in petroleum refining, is one of the most toxic chemicals known to humans.

In nonindustrial settings, Ni is found in a wide variety of consumer products, including utensils, jewelry, eyeglass frames, clothing buttons, braces, orthopedic implants, and some varieties of cell phones. Despite its incorporation into MoM prosthetics and increased concern with failed MoM devices (see earlier section on "Chromium"), Ni has not been implicated as a significant contributor to the associated clinical presentation.

Clinical Presentation and Treatment

Contact dermatitis is the symptom often associated with Ni exposure in the nonoccupational setting. Ni allergy is the most common contact allergen in women because of its ubiquitous use in jewelry and cosmetic products. Patch test studies have demonstrated a strong dose-response relationship between degree of exposure and severity of symptoms.

Patients exposed to Ni carbonyl exhibit rapid onset of pulmonary congestion and inability to oxygenate hemoglobin, followed by development of lesions of the lung, liver, kidney, adrenal glands, and spleen. Patients undergoing dialysis are exposed to Ni and accumulate Ni in blood and other organs.

Preanalytical and Analytical Aspects

Use of a plastic cannula for blood sampling was shown to be unnecessary in the assessment of Ni; however, sporadic contamination from stainless steel needles was observed.

Regulatory and Occupational Exposure Aspects

Workplace exposures to various forms of Ni occur in a variety of industrial processes. Exposure to sparingly soluble Ni compounds, such as nickel sulfide, nickel oxide, nickel carbonate, and nickel sulfidic ores, generally occurs during the mining of nickel ores, the smelting and refining processes, and the grinding and welding of Ni-containing alloys. The electroplating industry is usually the main source of workplace exposures to soluble forms of Ni such as nickel acetate, nickel chloride, nickel hydroxide, and nickel sulfate. Exposure to nickel carbonyl (or nickel tetracarbonyl), an easily absorbed and highly toxic form of Ni, likely occurs in Ni refining, nickel coatings, plating glass, and as a catalyst in chemical reactions.

Occupational Ni exposure mainly occurs by inhalation of dusts from sparingly soluble Ni compounds, aerosols from soluble Ni solutions, or gaseous Ni generally from nickel carbonyl. Skin contact with airborne Ni and Ni-containing solutions is also possible, as is oral ingestion in facilities with poor industrial hygiene practices and workers who practice poor personal hygiene. The solubility of the Ni compound will affect the toxicokinetics of Ni in the body. Soluble Ni compounds are more rapidly absorbed and eliminated compared to the poorly soluble forms. Inhaled particles of sparingly soluble Ni compounds can be retained in the lung tissue and regional lymph nodes, where they can accumulate and gradually be released over time. The major elimination route for absorbed Ni is through the kidneys and into the urine.

The toxicities of most significance after occupational exposure to Ni are the allergenic and carcinogenic effects. Ni exposure can cause sensitization of the skin (e.g., contact dermatitis) and respiratory tract. Occupational exposure to Ni compounds can also increase the risk for lung and nasal cancer in workers. The International Agency for Research on Cancer (IARC) classifies Ni compounds as human carcinogens (group 1). Biological monitoring of Ni concentrations in body fluids after occupational Ni exposure has been extensively studied in workers, and Ni concentrations in body fluids are affected by the chemical species of the Ni to which the individual was exposed, and the time of sample collection. In general, except for exposure to nickel carbonyl, urinary Ni levels can be useful in assessing the severity of the poisoning. However, elevated Ni concentrations in body fluids may not be directly correlated to exposure levels or risk for disease and only provide information about Ni uptake.

Thallium

Thallium (Tl) is often referred to as the "poisoner's poison" because it is nearly tasteless and highly toxic in low doses. Isotopes of Tl are used in diagnostic imaging and were once used medicinally to treat ringworm.

Sources of Exposure

Thallium is a by-product of zinc and lead smelting and cadmium production. Interest in Tl derives primarily from its former use as a rodenticide, with accidental exposure the most likely reason for toxicity. Acute and often fatal poisoning has been reported after criminal use and suicide attempts. Additionally, environmental concerns are growing because Tl is a waste product of coal combustion and the manufacturing of cement. Thallium is used in photoelectric cells, lamps, semiconductors, and scintillation counters.

Clinical Presentation and Treatment

The clinical presentation of Tl toxicity varies based on dose, age, and acute or chronic exposure. Patients exposed to high doses of Tl (>1 g) demonstrate alopecia (hair loss), peripheral neuropathy, seizures, and renal failure. Peripheral neuropathy of the feet and less commonly the hands occurs within 1 week after exposure, and hair loss begins at the same time and continues for several weeks. GI symptoms, including pain, diarrhea, and constipation have been reported in acute ingestion, in addition to myalgias, pleuritic chest pain, insomnia, optic neuritis, hypertension, cardiac abnormalities, Mees lines, and liver injury.

Preanalytical and Analytical Aspects

Typical serum concentrations in the absence of significant exposure are less than 10 ng/mL (49 nmol/L), and daily urine excretion is less than 10 μg/day (49 nmol/day). Exposed patients have been observed to have serum concentrations as high as 50 ng/mL (245 nmol/L), with urine output in excess of 500 μg/day (2446 nmol/day). Tl is routinely measured by ICP-MS in blood and urine.

Regulatory and Occupational Exposure Aspects

Urine is the suggested specimen for biological monitoring of Tl. Although there is insufficient information available correlating occupational exposure levels and urinary Tl concentrations to establish a threshold value, the biological monitoring of Tl in urine can assist in assessing the effectiveness of workplace control measures, especially because the ACGIH gives Tl and Tl compounds a "skin" notation designating the potential for a significant contribution to the overall exposure by the cutaneous route. For exposures to Tl and inorganic Tl compounds, the Finnish Institute of Occupational Health (FIOH) has no established biomonitoring action limit but recommends testing for Tl in a post-shift urine specimen.

REVIEW QUESTIONS

1. An individual has eaten a large meal of seafood including shellfish and haddock before submitting a urine specimen for metal analysis. What metal would likely be present in a high concentration in his urine specimen?
   a. Arsenic
   b. Copper
   c. Mercury
   d. Cadmium
2. Lead inhibits _____, an enzyme that catalyzes synthesis of heme from porphyrin.
   a. ATPase
   b. RNA polymerase
   c. glutathione peroxidase
   d. aminolevulinic acid dehydratase
3. Which of the following reasons best explains why elemental toxicity is difficult to diagnose?
   a. Physicians are poorly informed about prevalence of toxic elements.
   b. Laboratories are unable to properly test for toxic elements.
   c. Signs and symptoms in elemental toxicity mirror numerous disease states.
   d. Elemental toxicity is relatively rare in developed countries.
4. Analysis of toxic elements is recommended in which of the following presentations?
   a. Renal disease due to chronic hypertension
   b. Unexplained, bilateral neuropathy
   c. Chronic inflammation of the liver
   d. No history of exposure on record
5. Which of the following correctly identifies an accepted approach to elemental exposure?
   a. Wait and see if symptoms become more severe
   b. Prospective chelation challenges with postchelation laboratory assessments
   c. Enhanced absorption
   d. Antidotal treatment
6. The toxicity of Cd, As, Hg, and Pb are similar in regard to damage of which organ?
   a. Eye
   b. Kidney
   c. Pancreas
   d. Distal phalanges
7. Which element in its hexavalent state is toxic while in its trivalent state is considered largely benign?
   a. Cr
   b. Hg
   c. Pb
   d. As
8. Which element in its elemental state is largely benign unless inhaled while its inorganic form is the basis for the phrase "mad as a hatter"?
   a. As
   b. Cr
   c. Pb
   d. Hg
9. Regarding Ni exposure, _____ Ni compounds are more rapidly absorbed and eliminated compared to the _____ forms.
   a. soluble, gaseous
   b. insoluble, soluble
   c. soluble, less soluble
   d. alloyed, pure
10. The measurement of biological samples from a worker for elements found in a workplace following occupational exposure assesses which of the following?
   a. The air concentration of the measured element
   b. The nutritional status of the worker
   c. The worker's ability to schedule time off
   d. The effectiveness of environmental control strategies used in the workplace

SUGGESTED READINGS

Alfrey, A. C., et al. (1972). Syndrome of dyspraxia and multifocal seizures associated with chronic hemodialysis. *Transactions - American Society for Artificial Internal Organs*, *18*(0), 257–261, 266–257.

Caldwell, K. L., Jones, R. L., Verdon, C. P., et al. (2009). Levels of urinary total and speciated arsenic in the US population: national health and nutrition examination survey 2003-2004. *Journal of Exposure Science & Environmental Epidemiology*, *19*, 59–68.

Dart, R. C., & Sullivan, J. B. (2004). Mercury. In R. C. Dart (Ed.), *Medical toxicology* (3rd ed.). (pp. 1437–1448). Philadelphia: Lippincott Williams & Wilkins.

Hadrup, N., & Lam, H. R. (2014). Oral toxicity of silver ions, silver nanoparticles and colloidal silver: a review. *Regulatory Toxicology and Pharmacology*, *68*, 1–7.

Haglock-Adler, C. J., & Strathmann, F. G. (2015). Simplified sample preparation in the simultaneous measurement of whole blood antimony, bismuth, manganese, and zinc by inductively coupled plasma mass spectrometry. *Clinical Biochemistry*, *48*, 135–139.

Hahn, L. J., et al. (1989). Dental "silver" tooth fillings: a source of mercury exposure revealed by whole-body image scan and tissue analysis. *FASEB Journal*, *3*, 2641–2646.

Hansen, G., et al. (2004). Evaluation of the mercury exposure of dental amalgam patients by the Mercury Triple Test. *Occupational and Environmental Medicine*, *61*, 535–540.

Ibrahim, D., Froberg, B., Wolf, A., et al. (2006). Heavy metal poisoning: clinical presentations and pathophysiology. *Clinics Laboratory Medicine*, *26*, 67–97, viii.

Jin, T., & Berlin, M. (2007). Titanium. In G. Nordberg, B. Fowler, M. Nordberg, et al. (Eds.), *Handbook on the toxicology of metals* (3rd ed.). (pp. 861–870). Amsterdam: Academic Press.

Kazantzis, G. (2007). Diagnosis and treatment of metal poisoning: general aspects. In (3rd ed.). G. F. Nordberg, B. A. Fowler, M. Nordberg, et al. (Eds.), *Handbook on the toxicology of metals* (Vol). (pp. 303 -318). San Diego, CA: Elsevier.

Lauwerys, R. R., & Hoet, P. (2001). Biological monitoring of exposure to inorganic and organometallic substances. In *Industrial chemical exposure: Guidelines for biological monitoring* (3rd ed.). (pp. 21–181). Boca Raton: Lewis Publishers.

Lison, D. (2007). Cobalt. In G. Nordberg, B. Fowler, M. Nordberg, et al. (Eds.), *Handbook on the toxicology of metals* (3rd ed.). (pp. 511–528). Amsterdam: Academic Press.

McDiarmid, M. A., Gaitens, J. M., & Squibb, K. S. (2012). Uranium and thorium. In E. Bingham, B. Cohrssen, & F. A. Patty (Eds.), *Patty's toxicology* (6th ed.). (pp. 769–816). Hoboken, N.J: John Wiley & Sons.

Michalke, B. (2003). Element speciation definitions, analytical methodology, and some examples. *Ecotoxicology and Environmental Safety*, *56*, 122–139.

Michalke, B., & Fernsebner, K. (2014). New insights into manganese toxicity and speciation. *Journal of Trace Elements in Medicine and Biology*, *28*, 106–116.

Nelson, L. S. (2011). Copper. In L. Nelson, & L. R. Goldfrank (Eds.), *Goldfrank's toxicologic emergencies* (9th ed.). New York: McGraw-Hill Medical. xxviii.

Newton, A. W., Ranganath, L., Armstrong, C., et al. (2012). Differential distribution of cobalt, chromium, and nickel between whole blood, plasma and urine in patients after metal-on-metal (MoM) hip arthroplasty. *Journal of Orthopaedic Research*, *30*, 1640–1646.

Nordberg, G. F., Nogawa, K., & Nordberg, M., et al. Cadmium. In: G. Nordberg, B. Fowler, M. Nordberg, et al., (Eds.) *Handbook on the toxicology of metals.* (3rd ed.). Amsterdam: Academic Press 445–486.

Riihimäki, V., & Aitio, A. (2012). Occupational exposure to aluminum and its biomonitoring in perspective. *Critical Reviews in Toxicology*, *42*, 827–853.

Rogan, W. J., & Ware, J. H. (2003). Exposure to lead in children: how low is low enough? *The New England Journal of Medicine*, *348*, 1515–1516.

Siddique, A., & Kowdley, K. V. (2012). The iron overload syndromes [Review]. *Alimentary Pharmacology & Therapeutics*, *35*, 876–893.

Skerfving, S., & Bergdahl, I. A. (2007). Lead. In G. Nordberg, B. Fowler, M. Nordberg, et al. (Eds.), *Handbook on the toxicology of metals* (3rd ed.). (pp. 599–643). Amsterdam: Academic Press.

Tylenda, C. A., & Fowler, B. A. (2007). Antimony. In G. Nordberg, B. Fowler, M. Nordberg, et al. (Eds.), *Handbook on the toxicology of metals* (3rd ed.). (pp. 353–365). Amsterdam: Academic Press.

33

Diabetes Mellitus

David B. Sacks

OBJECTIVES

1. Define the following:
   - Advanced glycation end product
   - Albuminuria
   - C-peptide
   - Diabetes mellitus
   - Epinephrine
   - Fatty acid
   - Fructosamine
   - Gestational diabetes mellitus
   - Glucagon
   - Glucose tolerance
   - Glycated albumin
   - Glycated hemoglobin
   - Hemoglobin A_{1c}
   - Hyperglycemia
   - Insulin
   - Insulin-like growth factor 1
   - Insulin resistance syndrome
   - Ketoacidosis
   - Ketone bodies
   - Proinsulin
2. Compare and contrast type 1 and type 2 diabetes mellitus including the following:
   - Causes
   - Symptoms
   - Insulin concentration in blood
   - Appearance of symptoms versus clinical diagnosis
   - Development of chronic complications
   - Genetic factors
3. Compare impaired glucose tolerance with impaired fasting glucose and the usefulness of these classifications as risk factors for diabetes and cardiovascular disease.
4. List the three important functions of insulin.
5. Summarize how the following hormones specifically affect blood glucose concentration: insulin, glucagon, epinephrine, cortisol, somatostatin, and growth hormone.
6. Briefly describe the function of the facilitative glucose transporters, including the importance of GLUT4 in glucose uptake by skeletal muscle.
7. List five antibodies that are involved in the pathogenesis of type 1 diabetes mellitus.
8. Describe insulin resistance and the role it plays in the pathogenesis of type 2 diabetes mellitus; explain how loss of β-cell function results in development of type 2 diabetes mellitus.
9. State how diet and exercise are related to development of type 2 diabetes mellitus.
10. State the basic criteria for the diagnosis of diabetes mellitus including fasting and random plasma glucose concentration, oral glucose tolerance test, and glycated hemoglobin results.
11. Outline the procedure for and interpretation of an oral glucose tolerance test for the diagnosis of type 2 diabetes mellitus, impaired glucose tolerance, and gestational diabetes mellitus.
12. Summarize the role of the clinical laboratory in the diagnosis and short- and long-term management of diabetes mellitus.
13. Describe the assay principles of a whole blood glucose meter; calculate the difference between whole blood and plasma glucose concentrations.
14. List five variables that affect the accuracy and reproducibility of blood glucose meters; state how dehydration affects the reliability of a blood glucose meter.
15. Describe a typical implanted sensor for monitoring blood glucose including assay method, calibration, and usefulness in a person with type 1 diabetes.
16. Discuss the metabolic relationships among glucose, ketones, fatty acids, and metabolic acids including how they are altered, the effects of increased counterregulatory hormones, and the clinical significance of ketonemia and ketonuria in uncontrolled diabetes mellitus.
17. Describe the glycation of hemoglobin; state the percentage of hemoglobin A_1 constituted by A_{1c}.

18. Explain how measurement of glycated hemoglobin is useful in the diagnosis and monitoring of diabetes; state the cutoff value of hemoglobin A_{1c} used to diagnose diabetes.
19. Specify the effects that young and old erythrocytes have on hemoglobin A_{1c} values.
20. List and discuss three techniques used to determine hemoglobin A_{1c} values including principles of analysis, possible interference by hemoglobin variants, and specimen requirements.
21. Describe the effects of hyperglycemia on advanced glycation end products and how these end products contribute to certain complications of diabetes.
22. State how urinary albumin excretion is useful in determining diabetic nephropathy; list the specimen collection requirements for this urine test.
23. Describe three semiquantitative analyses of urinary albumin.
24. State the healthy reference intervals used for fasting plasma glucose, oral glucose tolerance, and hemoglobin A_{1c} analyses.
25. Analyze and solve case studies related to the different types of diabetes and impaired glucose tolerance using descriptions of symptoms and results of laboratory analyses.

KEY WORDS AND DEFINITIONS

Advanced glycation end products (AGEs) Proteins that have been irreversibly modified by nonenzymatic attachment of glucose; may contribute to the chronic complications of diabetes.

Diabetes mellitus A group of metabolic disorders of carbohydrate metabolism in which glucose is underutilized, producing hyperglycemia.

Gestational diabetes mellitus (GDM) Carbohydrate intolerance with onset during pregnancy.

Glucose A six-carbon simple sugar that is the premier fuel for most organisms and an important precursor of other body constituents.

Glycated hemoglobin Hemoglobin that has a sugar residue attached; HbA_{1c} is the major fraction of glycated hemoglobin.

Glycogen A polysaccharide with a formula of $(C_6H_{10}O_5)$ used by muscle and liver for carbohydrate storage.

Hyperglycemia Increased glucose concentrations in the blood.

Hypoglycemia Decreased glucose concentrations in the blood.

Insulin A protein hormone produced by β-cells of the pancreas that decreases blood glucose concentrations.

Ketones Compounds that arise from free fatty acid breakdown; insulin deficiency leads to increased serum ketones, which are the major contributors to the metabolic acidosis that occurs in individuals with diabetic ketoacidosis.

Diabetes mellitus is a group of metabolic disorders characterized by **hyperglycemia** resulting from defects in **insulin** secretion, insulin action, or both. Some patients may experience acute life-threatening hyperglycemic episodes, such as ketoacidosis or hyperosmolar coma. Acute life-threatening hypoglycemic episodes may occur as a result of therapy. As the disease progresses, patients are at increased risk for the development of specific complications, including *retinopathy* leading to blindness, *nephropathy* leading to renal failure, and *neuropathy* (nerve damage), collectively known as microvascular complications, as well as *atherosclerosis,* which is considered a *macrovascular complication*. The last may result in stroke, gangrene, or coronary artery disease.

Diabetes is a common disease, although the exact prevalence is unknown. The number of people with diabetes has increased dramatically worldwide. It is estimated that approximately 380 million people currently have diabetes, and by 2035 this number is predicted to reach 592 million, 80% of whom will live in low- and middle-income countries. In the United States, the prevalence in 1999–2002 was 9.3% and by 2005–2006, this had increased to 12.9% (equivalent to ≈40 million people). Similarly, the prevalence of diabetes in Asian populations has increased rapidly in recent decades, with China and India ranked first and second, respectively, among countries with the largest diabetic populations. Recent analysis in Chinese adults suggests that in 2010 the estimated prevalence of diabetes and prediabetes was 11.6% and 50.1%, respectively, which is equivalent to 113.9 million persons with diabetes and 493.4 million with prediabetes. The prevalence varies widely among countries, reaching as high as 25% in the Middle East and 35% in the Western Pacific. Further statistics about individual countries, compiled by the International Diabetes Federation (IDF), can be found at http://www.diabetesatlas.org/. These data led to a description of diabetes as "one of the main threats to human health in the twenty-first century." It is estimated that approximately 50% of individuals with diabetes worldwide remain undiagnosed. The prevalence of undiagnosed diabetes in the United States was reduced to 11% between 2006 and 2010 and has remained fairly stable in recent years. The prevalence of diabetes mellitus increases with age, and approximately half of all cases occur in people older than 55 years. In the United States, more than 20% of the population older than 65 years have diabetes. A racial predilection has been noted, and by the age

of 65, 33%, 25%, and 17% of Hispanics, blacks, and whites, respectively, in the United States have diabetes. In 2012, diabetes mellitus was estimated to be responsible for $245 billion in healthcare expenditures in the United States. The estimated economic burden for undiagnosed diabetes, prediabetes, and **gestational diabetes mellitus (GDM)** in 2012 was $78 billion, producing a total economic burden of more than $322 billion. Worldwide diabetes caused at least $612 billion in health expenditures and an estimated 4.9 million deaths in 2014. Acute and chronic complications make diabetes the fourth most common cause of death in the developed world.

CLASSIFICATION

Diabetes was initially diagnosed by the oral glucose tolerance test (OGTT). The 1979 classification scheme by the National Diabetes Data Group recognized two major forms of diabetes: type 1 (insulin-dependent) diabetes mellitus (IDDM) and type 2 (non–insulin-dependent) diabetes mellitus (NIDDM). The terms *juvenile-onset* and *adult-onset diabetes* were abolished. To base the classification on cause rather than on treatment, the American Diabetes Association (ADA) reexamined the classification and diagnosis of diabetes mellitus in 1995. The revised classification, published in 1997, eliminates the terms *insulin-dependent diabetes mellitus* and *non–insulin-dependent diabetes mellitus,* which now are termed *type 1* and *type 2 diabetes,* respectively (Box 33.1).

Type 1 Diabetes Mellitus

Approximately 5% to 10% of all cases of diabetes mellitus are included in this category. Patients usually have abrupt onset of symptoms (e.g., polyuria, polydipsia, rapid weight loss). They have insulinopenia (a deficiency of insulin) caused by destruction of pancreatic islet β-cells and are dependent on insulin to sustain life and prevent ketosis. Most patients have antibodies that identify an autoimmune process (see later discussion); some have no evidence of autoimmunity and are classified as type 1 idiopathic. The peak incidence occurs in childhood and adolescence. Approximately 75% acquire the disease before the age of 18, but onset in the remainder may occur at any age. Age at presentation is not a criterion for classification.

Type 2 Diabetes Mellitus

This group accounts for approximately 90% of all cases of diabetes. Patients have minimal symptoms, are not prone to ketosis, and *are not dependent on insulin* to prevent ketonuria.

BOX 33.1 Classification of Diabetes Mellitus

I. Type 1 diabetes
II. Type 2 diabetes
III. Gestational diabetes mellitus (GDM)
IV. Specific types of diabetes due to other causes

From the American Diabetes Association. (2017). Classification and diagnosis of diabetes. *Diabetes Care, 40* (Suppl 1), S11–S24.

Insulin concentrations may be normal, decreased, or increased, and most people with this form of diabetes have impaired insulin action. *Obesity* is commonly associated, and weight loss alone usually improves hyperglycemia in these persons. However, many individuals with type 2 diabetes may require dietary intervention, oral antihyperglycemic agents, or insulin to control hyperglycemia. Most patients acquire the disease after age 40, but it may occur in younger people. Type 2 diabetes in children and adolescents is an emerging, significant problem. Among children in Japan, type 2 diabetes is now more common than type 1.

Gestational Diabetes Mellitus

This is defined as any degree of **glucose** intolerance (i.e., hyperglycemia) *with onset or first recognition during pregnancy* (i.e., diabetic women who become pregnant are not included in this category). Estimates of the frequency of abnormal glucose tolerance during pregnancy range from less than 1% to 28%, depending on the population studied and the diagnostic tests employed. In the United States, GDM occurs in 6% to 8% of pregnancies (≈240,000 cases annually). The prevalence of GDM is increasing, at least in part, due to the considerable increase in obesity. Women with GDM are at significantly greater risk for the subsequent development of type 2 diabetes mellitus, which occurs in 15% to 60%. The risk is particularly high in women who have marked hyperglycemia during or soon after pregnancy, women who are obese, and those whose GDM was diagnosed before 24 weeks' gestation. At 6 to 12 weeks postpartum, all patients who had GDM should be evaluated for diabetes using nonpregnant OGTT criteria. If diabetes is not present, patients should be reevaluated for diabetes at least every 3 years.

Specific Types of Diabetes Due to Other Causes

This subclass includes uncommon patients in whom hyperglycemia is due to a specific underlying disorder, such as genetic defects of β-cell function; genetic defects in insulin action; diseases of the exocrine pancreas (e.g., cystic fibrosis); endocrinopathies (e.g., Cushing syndrome, acromegaly, glucagonoma); administration of hormones or drugs known to induce β-cell dysfunction (e.g., Dilantin, pentamidine) or to impair insulin action (e.g., glucocorticoids, thiazides, β-adrenergics); infection; uncommon forms of immune-mediated diabetes; or other genetic conditions (e.g., Down syndrome, Klinefelter syndrome, porphyria). This was formerly termed *secondary diabetes.*

CATEGORIES OF INCREASED RISK FOR DIABETES

People who have blood glucose concentrations above normal, but less than those required for a diagnosis of diabetes mellitus, fall into an intermediate category. In 1979, this was termed impaired glucose tolerance (IGT), and was defined as a 2 hours postload plasma glucose following an OGTT of 140 to 199 mg/dL (7.8 to 11.1 mmol/L). An OGTT is required to assign a patient to this class. To avoid an OGTT, the category

of impaired fasting glucose (IFG) was added in 1997 by the ADA and by the World Health Organization (WHO) in 1999. IFG is diagnosed by a fasting glucose value between 100 and 125 mg/dL (5.6 and 6.9 mmol/L). (Note that the WHO and a number of other diabetes organizations define the cutoff for IFG at 110 mg/mL [6.1 mmol/L]).

In 2009, hemoglobin A_{1c} (HbA_{1c}) was added as a criterion to diagnose diabetes. People with HbA_{1c} values below the cutoff for diabetes, that is, 6.5% (48 mmol/mol), but above the upper reference limit, that is, 5.6% (38 mmol/mol), are at high risk of developing diabetes. For example, the incidence of diabetes in people with HbA_{1c} of 5.5% to 6.0% (37 to 42 mmol/mol) is 3- to 8-fold higher than the general population, while in those with HbA_{1c} between 6.0 and less than 6.5% (42 and 48 mmol/mol) it is increased more than 10-fold.

Individuals with IFG and/or IGT and/or intermediate HbA_{1c} (5.7% to 6.4%; 39 to 46 mmol/mol) have been referred to as having "prediabetes," because they are at high risk for progressing to diabetes. Moreover, they are at increased risk for the development of cardiovascular disease. Nevertheless, there is considerable controversy surrounding prediabetes. For example, IFG, IGT, and HbA_{1c} do not always identify the same individuals. There is also lack of agreement on the cutpoints for IFG and HbA_{1c}. This is due, in large part, to the continuous nature of the risk for the development of diabetes and the concentration of glucose or HbA_{1c}.

HORMONES THAT REGULATE BLOOD GLUCOSE CONCENTRATION

During a brief fast, a precipitous decline in the concentration of blood glucose is prevented by breakdown of **glycogen** stored in the liver and synthesis of glucose in the liver. Some glucose is derived from gluconeogenesis in the kidneys. These organs contain glucose-6-phosphatase, which is necessary to convert glucose 6-phosphate (derived from gluconeogenesis or glycogenolysis) to glucose. Skeletal muscle lacks this enzyme; muscle glycogen therefore cannot contribute directly to blood glucose. With more prolonged fasting (>42 hours), gluconeogenesis accounts for essentially all glucose production. In contrast, after a meal, the absorbed glucose is converted to glycogen (for storage in the liver and skeletal muscle) or fat (for storage in adipose tissue). Despite large fluctuations in the supply and demand of carbohydrates, the concentration of glucose in the blood is normally maintained within a fairly narrow range by hormones that modulate the movement of glucose into and out of the circulation. These include insulin, which decreases blood glucose, and the counterregulatory hormones (glucagon, epinephrine, cortisol, and growth hormone), which increase blood glucose concentrations (Fig. 33.1). Normal glucose disposal depends on (1) the ability of the pancreas to secrete insulin, (2) the ability of insulin to promote uptake of glucose into peripheral tissue, and (3) the ability of insulin to suppress hepatic glucose production. The major insulin target organs are liver, skeletal muscle, and adipose tissue. These organs exhibit some differences in their responses to insulin. For example, the hormone stimulates glucose uptake through a specific glucose transporter—GLUT4—into muscle and fat cells, but not into liver cells.

Insulin

Insulin is a protein hormone produced by the β-cells of the islets of Langerhans in the pancreas. Insulin was the first protein hormone to be sequenced, the first substance to be measured by radioimmunoassay (RIA), and the first compound produced by recombinant DNA technology for clinical use. It is an anabolic hormone that stimulates the uptake of glucose

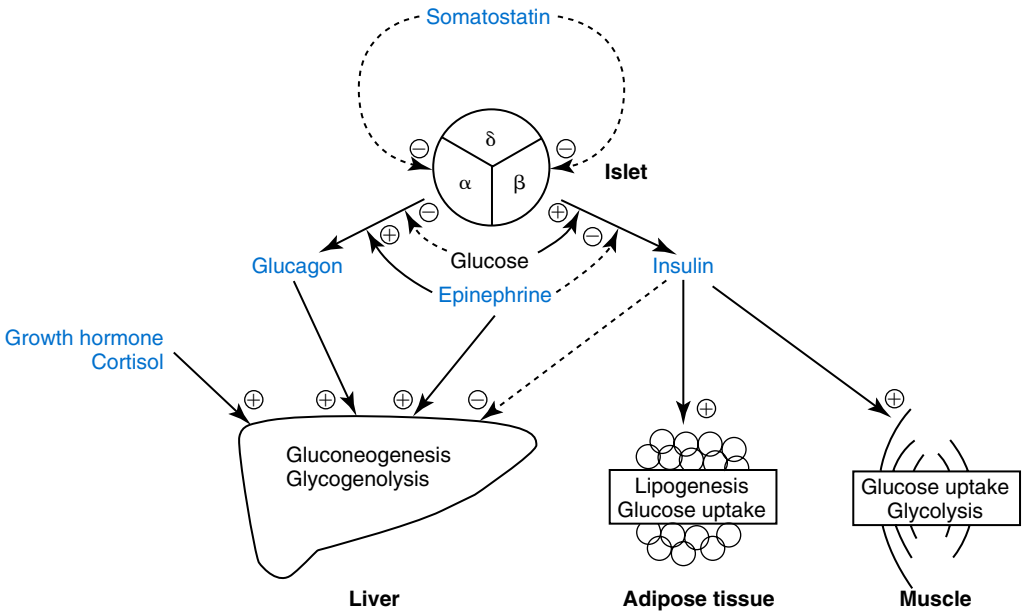


Fig. 33.1 Hormonal regulation of blood glucose. Key: +, stimulation; –, inhibition. Cortisol, growth hormone, and epinephrine antagonize the effects of insulin.

into fat and muscle, promotes the conversion of glucose to glycogen or fat for storage, inhibits glucose production by the liver, stimulates protein synthesis, and inhibits protein breakdown. Human insulin (molecular mass 5808 Da) consists of 51 amino acids in two chains (A and B) joined by two disulfide bridges, with a third disulfide bridge within the A chain. Insulin from most animals is immunologically and biologically similar to human insulin, and in the past, patients were treated with insulin purified from beef or pig pancreas. Virtually all patients are now treated with recombinant human insulin.

Preproinsulin, a protein of about 100 amino acids, is not detectable in the circulation under normal conditions because it is rapidly converted by cleaving enzymes to proinsulin. Proinsulin is stored in secretory granules in the Golgi complex of the β-cells, where *proteolytic cleavage to insulin and connecting peptide (C-peptide) occurs.* At the cell membrane, insulin and C-peptide are released into the portal circulation in equimolar amounts. In addition, small amounts of proinsulin and intermediate cleavage forms enter the circulation.

Proinsulin, which has relatively low biological activity (approximately 10% of insulin potency), is the major storage form of insulin. Normally, only small amounts (about 3% of the amount of insulin, on a molar basis) of proinsulin enter the circulation.

C-peptide is devoid of biological activity but appears necessary to ensure the correct structure of insulin. Fasting C-peptide concentrations are 5- to 10-fold higher than those of insulin because of the longer half-life of C-peptide (≈35 minutes). The liver does not extract C-peptide, which is removed from the circulation by the kidneys and is degraded, with a fraction excreted unchanged in the urine.

Glucose Transport

One of the fundamental effects of insulin is to increase glucose uptake into cells. The molecular mechanism of this action of insulin is extremely complex. The transport of glucose into cells is modulated by two families of proteins. The sodium-dependent glucose transporter promotes the uptake of glucose and galactose from the lumen of the small bowel and their reabsorption from urine in the kidney. The second family of glucose carriers, termed *facilitative glucose transporters (GLUTs),* is located on the surface of all cells (Table 33.1). These transporters are designated GLUT1 to GLUT14, according to the order in which they were identified. Eleven have been shown to mediate glucose transport. On the basis of sequence similarities, they can be divided into three subfamilies: class I (GLUT 1 to 4), class II (GLUT 5, 7, 9, and 11), and class III (GLUT 6, 8, 10, 12, and 14). GLUT1 is widely expressed and provides many cells with their basal glucose requirement. GLUT1 in the blood-brain barrier and GLUT3 in neuronal cells provide the constant high concentrations of glucose required by the brain. GLUT2 is expressed in hepatocytes, β-cells of the pancreas, and basolateral membranes of intestinal and renal epithelial cells. It is a low-affinity, high-capacity transport system that allows non–rate-limiting movement of glucose into and from these cells. GLUT4 catalyzes the rate-limiting step for glucose uptake and metabolism in skeletal muscle, the major organ of glucose consumption.

When circulating insulin concentrations are low, most of the GLUT4 is localized in intracellular compartments and is inactive. After a meal, the pancreas releases insulin, which stimulates the translocation of GLUT4 to the plasma membrane, thereby promoting glucose uptake into skeletal muscle and fat. Insulin-stimulated glucose transport into skeletal muscle is impaired in individuals with type 2 diabetes, but the mechanism of the defect has not been established. GLUT5 is responsible for fructose uptake in the intestine. Less is known about the other GLUTs. Glucose transport has been reported for GLUT 6, 8, 11, and 12.

Insulin-Like Growth Factors

Insulin-like growth factors 1 and 2 (IGF-1 and IGF-2) are polypeptides structurally related to insulin. These hormones (previously referred to as *nonsuppressible insulin-like activity*

TABLE 33.1 Facilitative Human Glucose Transporters

Name	Class	Tissue	Function
GLUT1	I	Erythrocytes, brain, blood-brain barrier, fetal tissues	Basal glucose transport, particularly in erythrocytes, brain, and placenta
GLUT2	I	Liver, β-cells of pancreas, small intestine, kidney, brain	Non–rate-limiting glucose transport
GLUT3	I	Brain (neurons), testis	Glucose transport in neurons
GLUT4	I	Skeletal muscle, cardiac muscle, adipose tissue (white and brown)	Insulin-stimulated glucose transport
GLUT5	II	Small intestine, kidney	Transports fructose (not glucose)
GLUT6	III	Brain, spleen, leukocytes	
GLUT7	II	Small intestine, colon, testis, prostate	
GLUT8	III	Testis, brain, adrenal gland, liver, spleen, brown adipose tissue, lung	
GLUT9	II	Kidney, liver, small intestine, placenta, lung, leukocytes	Transports urate
GLUT10	III	Heart, lung, brain, liver, pancreas, placenta, kidney	
GLUT11	II	Heart, skeletal muscle	
GLUT12	III	Heart, prostate, skeletal muscle, placenta	
HMIT (GLUT13)	III	Brain, adipose tissue	Transports *myo*-inositol (not glucose)
GLUT14	I	Testis	

or *somatomedin*) exhibit metabolic and growth-promoting effects similar to those of insulin. Accumulating evidence implicates the IGF axis in the development of several common cancers. IGF-1 (previously known as *somatomedin C*) is an important mediator of growth hormone action and is one of the major regulators of cell growth and differentiation. The physiological role of IGF-2 is not known. Synthesis of IGF-1 depends on growth hormone and occurs predominantly in the liver. In addition, many other cells produce IGF-1 that does not enter the circulation but acts locally. Circulating IGF concentrations are approximately 1000-fold higher than insulin concentrations, and the hormone is kept inactive by binding to a family of at least six specific binding proteins. These proteins regulate IGF by protecting the ligands in the circulation and delivering them to their target tissue. In contrast to insulin, which is unbound in the circulation, less than 10% of total serum IGF-1 is free. The biological actions of IGF are exerted through specific IGF receptors or the insulin receptor. The IGF-1 receptor is closely related to the insulin receptor in structure and biochemical properties. In contrast, the IGF-2 receptor is quite different, and its physiological relevance is not understood.

The significance of IGFs in normal carbohydrate metabolism is not known. Exogenous administration produces **hypoglycemia**, whereas a deficiency of IGF-1 results in dwarfism (pygmies and Laron dwarfs). Clinically, IGF-1 testing is primarily used to evaluate growth hormone deficiency and to detect and monitor treatment of growth hormone excess (acromegaly). IGF-2 may be produced in excess by extrapancreatic neoplasms, and patients may have fasting hypoglycemia.

Counterregulatory Hormones

Several hormones have actions opposite to those of insulin. These counterregulatory hormones are catabolic and increase hepatic glucose production initially by enhancing the breakdown of glycogen to glucose (glycogenolysis), and later by stimulating the synthesis of glucose (gluconeogenesis). The initial response (within minutes) to low blood glucose is an increase in glucose production, stimulated by glucagon and epinephrine. Over time (3 to 4 hours), growth hormone and cortisol increase glucose mobilization and decrease glucose use (see Fig. 33.1). Evidence suggests that glucose production by the liver is an inverse function of ambient glucose concentration, independent of hormonal factors *(glucose autoregulation)*. The role of other hormones or neurotransmitters is not clear but appears relatively unimportant.

Glucagon

Glucagon is a 29–amino acid polypeptide secreted by α-cells of the pancreas. It is derived from proglucagon, which is hydrolyzed by prohormone convertase 2 (PC2). The major target organ for glucagon is the liver, where it binds to a specific G protein–coupled receptor, which is expressed abundantly in liver and kidney, and to a lesser extent in other tissues, including heart, adipose, pancreas, and brain. Glucagon stimulates the production of glucose in the liver predominantly by glycogenolysis. Gluconeogenesis is also activated, and glycogenesis is inhibited. Glucagon secretion is regulated primarily by plasma glucose concentrations, with low and high plasma glucose being stimulatory and inhibitory, respectively. Long-standing diabetes mellitus impairs the glucagon response to hypoglycemia, resulting in an increased incidence of hypoglycemic episodes. Stress, exercise, and amino acids induce glucagon release. Insulin inhibits glucagon release from the pancreas and decreases glucagon gene expression, thereby attenuating its biosynthesis. Increased glucagon concentrations, secondary to insulin deficiency, are believed to contribute to the hyperglycemia and ketosis of diabetes. In addition to its effects on glycemia, glucagon directly regulates triglyceride, free fatty acid, and bile metabolism. For example, it enhances fatty acid oxidation and ketogenesis in the liver.

Proglucagon is also produced in the distal gut by enteroendocrine L cells and neurons in the nucleus of the solitary tract. In the intestinal L cells, proglucagon is cleaved by PC1/3 to glucagon-like peptide-1 (GLP-1), GLP-2, oxyntomodulin, glicentin, and intervening peptide 2 (IP2). Food ingestion stimulates secretion of GLP-1, which acts on β-cells of the pancreas to stimulate insulin gene transcription, potentiate glucose-induced insulin secretion, and inhibit glucagon secretion. GLP-1 and gastric inhibitory peptide (GIP; synthesized in and secreted from the duodenum and proximal jejunum) are incretin hormones that are responsible for up to 70% of postprandial insulin secretion. GLP-1 receptors are widely expressed and GLP-1 also regulates glucose concentration through extrapancreatic mechanisms, including inhibiting glucose production and food intake. For these reasons, GLP-1 analogs are used in the treatment of type 2 diabetes, and four (exenatide, liraglutide, albiglutide, and dulaglutide) have received US Food and Drug Administration (FDA) approval for use in the United States.

Epinephrine

Epinephrine (also called adrenaline), a catecholamine secreted by the adrenal medulla, stimulates glucose production (glycogenolysis) and decreases glucose use, thereby increasing blood glucose concentrations. It also stimulates glucagon secretion and inhibits insulin secretion by the pancreas (see Fig. 33.1). Epinephrine appears to have *a key role in glucose counterregulation when glucagon secretion is impaired* (e.g., in type 1 diabetes mellitus). Physical or emotional stress increases epinephrine production, releasing glucose for energy. Tumors of the adrenal medulla, known as *pheochromocytomas,* secrete excess epinephrine or norepinephrine and produce moderate hyperglycemia as long as glycogen stores are available in the liver. For more details, see Chapter 26.

Growth Hormone

Growth hormone is a polypeptide secreted by the anterior pituitary gland. It stimulates gluconeogenesis, enhances lipolysis, and antagonizes insulin-stimulated glucose uptake. For more details, see Chapter 40.

Cortisol

Cortisol, secreted by the adrenal cortex in response to adrenocorticotropic hormone (ACTH), stimulates gluconeogenesis and increases the breakdown of protein and fat. Patients with Cushing syndrome have increased cortisol production as a result of tumor or hyperplasia of the adrenal cortex and may become hyperglycemic. In contrast, people with Addison disease have adrenocortical insufficiency caused by destruction or atrophy of the adrenal cortex and may exhibit hypoglycemia. For more details, refer to Chapter 41.

Other Hormones Influencing Glucose Metabolism

Thyroxine

Thyroxine, secreted by the thyroid gland, is not directly involved in glucose homeostasis, but it stimulates glycogenolysis and increases the rates of gastric emptying and intestinal glucose absorption. These factors may produce glucose intolerance in thyrotoxic individuals, but patients usually have a fasting plasma glucose (FPG) concentration in the reference interval. For more details, refer to Chapter 42.

Somatostatin

Somatostatin, also called growth hormone–inhibiting hormone, is a 14–amino acid peptide found in the gastrointestinal tract, the hypothalamus, and the δ-cells of the pancreatic islets. Although somatostatin does not appear to have a direct effect on carbohydrate metabolism, it inhibits the release of growth hormone from the pituitary. In addition, somatostatin inhibits secretion of glucagon and insulin by the pancreas, thus modulating the reciprocal relationship between these two hormones.

MEASUREMENT OF INSULIN, PROINSULIN, C-PEPTIDE, AND GLUCAGON

Although interest in the possible clinical value of measurement of the concentrations of insulin and its precursors continues, the assays are useful primarily for research purposes. There is no role for routine testing for insulin, proinsulin, or C-peptide in most patients with diabetes mellitus. Measurement of C-peptide is sometimes necessary in the United States for patients to obtain insurance coverage for continuous subcutaneous insulin infusion pumps. Occasionally, C-peptide measurements may help distinguish type 1 from type 2 diabetes in ambiguous cases (e.g., patients who have type 2 phenotype, but present with ketoacidosis). It must be emphasized that the diagnostic criteria for diabetes mellitus do not include hormone measurements.

Various methods are used to measure insulin, proinsulin, C-peptide, and glucagon. A brief overview is provided here. For additional details, readers are referred to the expanded version of this chapter in the 6th edition of *Tietz Textbook of Clinical Chemistry and Molecular Diagnostics.*

Insulin

The primary clinical application of insulin measurement is in the evaluation of patients with fasting hypoglycemia (discussed in more detail in Chapter 22). Measurement of circulating insulin could be helpful in evaluating insulin resistance and insulin secretion. Increased concentrations of insulin in nondiabetic individuals are an independent predictor of the development of coronary artery disease. Nevertheless, it is not clear whether the increased insulin is responsible for the risk of coronary disease, and the clinical value of measuring it in this context is questionable.

In the past, measurement of insulin was advocated by some in the evaluation and management of patients with polycystic ovary syndrome. Women with this condition have insulin resistance and abnormal carbohydrate metabolism that may respond to oral antihyperglycemic agents. However, it is not clear whether assessing insulin resistance by measuring insulin concentrations affords any advantage over clinical signs of insulin resistance (body mass index, acanthosis nigricans), and the American College of Obstetricians and Gynecologists (ACOG) does not recommend routine measurements of insulin. Although a few investigators have recommended measuring insulin along with glucose during an OGTT as an aid to the early diagnosis of diabetes mellitus, this approach is not recommended.

Although insulin has been assayed for more than 50 years, no highly accurate, precise, and reliable procedure is available to measure insulin. The results of the most widely used commercial immunoassays are considerably inconsistent with within-assay coefficients of variation (CVs) ranging from 3.7% to 39% and among-assay CVs of 12% to 66%. The term *immunoreactive insulin* is used in reference to assays that may recognize, in addition to insulin, substrates that share antigenic epitopes with insulin. Examples include proinsulin, proinsulin conversion intermediates, and insulin derivatives produced by glycation or dimerization. Antisera raised against insulin show some cross-reactivity with proinsulin but not with C-peptide. Specificity is not a problem in healthy individuals because the low proinsulin concentrations do not appreciably affect the absolute values of insulin. In certain situations (e.g., patients with diabetes or islet cell tumors), proinsulin is present at higher concentrations, and direct assay of plasma may falsely overestimate the true insulin concentration. Because proinsulin has very low activity, incorrect conclusions regarding the availability of biologically active insulin may be reached in patients with diabetes. The magnitude of the error depends on the concentration of proinsulin and the extent of cross-reactivity of the antiserum with proinsulin. The various immunoassays also show different cross-reactivities with pharmacological insulins.

A stable isotope dilution mass spectrometry (IDMS) assay has been developed to measure the concentrations of insulin, proinsulin, and C-peptide. The difference in mass among the three analytes allows specific measurement of each protein. Comparison of patient samples revealed that most, but not all, results were higher by immunoassay than by mass spectrometry. Thus, immunoassays may overestimate insulin, particularly at low concentrations. The high protein concentration in the serum requires extraction of proteins (e.g., by immunoaffinity) and purification by high-performance liquid chromatography (HPLC) before quantification by mass

spectrometry. This method is not suitable for routine laboratory analysis, but it is the best higher-order measurement procedure available and can be used as a candidate reference measurement procedure.

Reference intervals vary among assays. After an overnight fast, insulin concentrations in healthy, normal, nonobese people range from 12 to 150 pmol/L (2 to 25 μIU/mL). (Fasting is defined as no caloric intake for at least 8 hours.) More specific assays that have minimal cross-reactivity with proinsulin reveal a fasting plasma insulin concentration less than 60 pmol/L (10 μIU/mL). Concentrations up to 1200 pmol/L (200 μIU/mL) can be reached during a glucose tolerance test. Fasting insulin values are higher in obesity and lower in trained athletes.

Proinsulin

High proinsulin concentrations are usually noted in patients with benign or malignant β-cell tumors of the pancreas. Most patients with β-cell tumors have increased insulin, C-peptide, and proinsulin concentrations, but occasionally only proinsulin is increased. Despite its low biological activity, proinsulin production may be adequate to produce hypoglycemia. In addition, a rare form of familial hyperproinsulinemia, produced by impaired conversion to insulin, has been described. Measurement of proinsulin can be useful to determine the amount of proinsulin-like material that cross-reacts in an insulin assay. Patients with type 2 diabetes have increased proportions of proinsulin and proinsulin conversion intermediates, high concentrations of which are associated with cardiovascular risk factors. Even relatively mild hyperglycemia produces hyperproinsulinemia, with values greater than 40% of insulin concentration in type 2 diabetes. Similarly, women with GDM have higher concentrations of proinsulin and split-32, 33-proinsulin than pregnant normoglycemic control subjects. Increased proinsulin concentrations may also be detected in patients with chronic renal failure, cirrhosis, or hyperthyroidism.

Accurate measurement of proinsulin has been difficult for several reasons: the blood concentrations are low; antibody production is difficult; most antisera cross-react with insulin and C-peptide, which are present in much higher concentrations; the assays measure intermediate cleavage forms of proinsulin, and reference preparations of pure proinsulin are not readily available. Biosynthetic proinsulins have more recently allowed the production of monoclonal antibodies to proinsulin and have been used in the manufacture of proinsulin calibrators and reference preparations.

Reference intervals for proinsulin are highly dependent on the method of analysis, the degree of cross-reactivity of the antisera, and the purity of proinsulin calibrators. Each laboratory should establish its own reference intervals. Reference intervals in healthy, fasting individuals reported in the literature vary from 1.1 to 6.9 pmol/L to 2.1 to 12.6 pmol/L.

C-Peptide

Measurement of C-peptide has several advantages over insulin measurement. Because hepatic metabolism is negligible, C-peptide concentrations are better indicators of β-cell function than is peripheral insulin concentration. Furthermore, C-peptide assays do not measure exogenous insulin and do not cross-react with insulin antibodies, which interfere with the insulin immunoassay.

The primary reason for measuring C-peptide is to evaluate fasting hypoglycemia. Some patients with insulin-producing β-cell tumors, particularly if hyperinsulinism is intermittent, may exhibit increased C-peptide concentrations with normal insulin concentrations. When hypoglycemia is due to surreptitious insulin injection, *insulin* concentrations will be high, but C-peptide values will be low; this occurs because C-peptide is not found in commercial insulin preparations and exogenous insulin suppresses β-cell function.

Basal or stimulated (by glucagon or glucose) C-peptide concentrations provide estimates of a patient's insulin secretory capacity and rate. Although valuable in clinical research, C-peptide measurement has a negligible role in the routine management of patients with diabetes. Medicare patients in the United States must have low C-peptide concentrations to be eligible for insurance coverage of insulin pumps.

Measurement of C-peptide is used to monitor patients' response to pancreatic surgery. C-peptide should be undetectable after a radical pancreatectomy and should increase after a successful pancreas or islet cell transplant. In addition, C-peptide concentration is used as an endpoint in immunomodulatory trials for the prevention of type 1 diabetes.

Measurements of urine C-peptide are useful when continuous assessment of β-cell function is desired, or when frequent blood sampling is not practical. The 24-hour urine C-peptide content (in the absence of renal failure, which produces increased concentrations) correlates well with fasting serum C-peptide concentration. However, the fraction of secreted C-peptide that is excreted in the urine exhibits high between-subject and within-subject variability. Some have advocated a C-peptide–to-creatinine ratio in a single urine sample to avoid these issues.

Assays are not affected by anti-insulin antibodies, but methodological problems produce large between-method variation. These difficulties include variable specificity among different antisera, variable cross-reactivity with proinsulin, and various types of C-peptide preparations used as a calibrator. A comparison of 40 serum samples using nine commercial C-peptide assay methods showed within- and between-run CVs ranging from less than 2% to greater than 10%, and from less than 2% to greater than 18%, respectively. Some methods had high imprecision, with between-run CVs exceeding 15%. Two IDMS methods for measuring C-peptide have been developed. Calibrating C-peptide measurements to an IDMS reference method increased comparability among laboratories.

Fasting serum concentrations of C-peptide in healthy people range from 0.78 to 1.89 ng/mL (0.25 to 0.6 nmol/L). After stimulation with glucose or glucagon, values range from 2.73 to 5.64 ng/mL (0.9 to 1.87 nmol/L), or three to five times the prestimulation value. Urinary C-peptide is usually in the range of 74 ± 26 μg/L (25 ± 8.8 μmol/L).

Glucagon

Very high concentrations of glucagon are seen in patients with α-cell tumors of the pancreas called *glucagonomas*. Patients with this tumor frequently have weight loss, necrolytic migratory erythema, diabetes, stomatitis, and diarrhea. Skin lesions often occur first and are frequently overlooked. Most tumors have metastasized when finally diagnosed. Low glucagon concentrations are associated with chronic pancreatitis and long-term sulfonylurea therapy.

A competitive RIA is available for measuring glucagon. Calibrator values are assigned at the manufacturer using the WHO glucagon international standard (69/194). Immunoassays that do not use radioactivity have been developed. A recent comparison of 10 commercially available assays reveals that some show cross-reactivity with other peptides from proglucagon, and all lack sensitivity at low glucagon concentrations.

Fasting plasma concentrations of glucagon vary depending on the method, ranging from 0 to 20 ng/L to up to 135 ng/L (0 to 5.7 pmol/L to 38.7 pmol/L). Values up to 500 times the upper reference limit may be found in patients with autonomously secreting α-cell neoplasms.

PATHOGENESIS OF TYPE 1 DIABETES MELLITUS

Type 1 diabetes mellitus results from autoimmune destruction of the insulin-secreting cells of pancreatic β-cells by chronic mononuclear cell infiltration, called insulitis. In the vast majority of patients, destruction is mediated by T cells. The α-cells, δ-cells, and other islet cells are preserved. The autoimmune process begins months or years before the clinical presentation, and an 80% to 90% reduction in the volume of β-cells is required to induce symptomatic type 1 diabetes. The rate of islet cell destruction is variable and is usually faster in children than in adults.

Autoimmune Antibodies

Circulating antibodies can be detected in the serum years before the onset of hyperglycemia. The best characterized islet autoantibodies are as follows:

1. Islet cell cytoplasmic antibodies (ICAs) react with a sialoglycoconjugate antigen present in the cytoplasm of all endocrine cells of the pancreatic islets. These antibodies are detected in the serum of 0.5% of normal subjects and 75% to 85% of patients with newly diagnosed type 1 diabetes. The antibodies are detected by immunofluorescence microscopy on frozen sections of human pancreatic tails. The ICA assay is cumbersome, labor intensive, and difficult to standardize. Few clinical laboratories are likely to implement this assay, which has marked interlaboratory variability in sensitivity and specificity.
2. *Insulin autoantibodies* (IAAs) are present in more than 90% of children who develop type 1 diabetes before age five, but in less than 40% of individuals who develop diabetes after age 12. Their frequency in healthy people is similar to that of ICA. An important caveat is that insulin antibodies develop after insulin therapy, even in those persons who use human insulin.
3. *Antibodies to the 65 kDa isoform of glutamic acid decarboxylase* (GAD65) have been found up to 10 years before the onset of clinical type 1 diabetes and are present in approximately 60% of patients with newly diagnosed type 1 diabetes.
4. *Insulinoma-associated antigens* (IA-2A and IA-2βA), directed against two tyrosine phosphatases, have been detected in more than 50% of newly diagnosed type 1 diabetes patients.
5. *Zinc transporter ZnT8* was identified in 2007 as a major autoantigen in type 1 diabetes. Initial analysis identified ZnT8 in 60% to 80% of patients with new-onset type 1 diabetes compared with less than 2% of controls and less than 3% of individuals with type 2 diabetes.

Autoantibody markers of immune destruction are present in 85% to 90% of individuals with immune-mediated diabetes when fasting hyperglycemia is initially detected. Approximately 5% to 10% of white adult patients who have the type 2 diabetes phenotype also have islet cell autoantibodies, particularly to GAD65. This condition has been termed latent autoimmune diabetes of adulthood (LADA) (previously termed type 1.5 diabetes or slow-onset diabetes in adults). Up to 1% to 2% of healthy individuals have a single autoantibody and are at low risk of developing immune-mediated diabetes. Because the prevalence of immune-mediated diabetes is low (~0.3% in the general population in most countries), the positive predictive value of a single autoantibody will be extremely low. Screening for islet autoantibodies is controversial because no acceptable therapy has been documented to reliably prevent or delay the clinical onset of diabetes in islet cell autoantibody-positive individuals.

Genetics

Susceptibility to type 1 diabetes is inherited, but the mode of inheritance is complex and has not been defined. It is a multigenic trait, and the major locus is the major histocompatibility complex on chromosome 6. At least 11 other loci on 9 chromosomes also contribute, with the regulatory region of the insulin gene *INS* on chromosome 11p15 being an important locus. The human leukocyte antigen (HLA)-DQ and -DR genetic factors are by far the most important determinants for risk of type 1 diabetes. The concordance rate between identical twins is approximately 30%. Approximately 95% of whites with type 1 diabetes express HLA-DR3 or HLA-DR4 histocompatibility antigens. However, up to 40% of the nondiabetic population also express these alleles. In contrast, the HLA-DQB1*0602 allele significantly decreases the risk of type 1 diabetes. HLA typing can indicate absolute risk of diabetes. The risk of a sibling developing diabetes is 1%, 5%, and 10% to 20% if the number of haplotypes shared is none, one, and two, respectively. However, only 10% of patients with type 1 diabetes have an affected first-degree relative. More than 40 non-HLA-susceptibility gene markers have been confirmed, but these have smaller effects than HLA. Non-HLA genetic factors that increase risk include the insulin gene *(INS)*,

cytotoxic T-lymphocyte-associated protein 4 *(CTLA4)*, and protein tyrosine phosphatase non-receptor type 22 (lymphoid) *(PTPN22)*. The multiplicity of independent chromosomal regions associated with a predisposition to type 1 diabetes suggests that other susceptibility genes will be identified. Routine measurement of genetic markers is not of value at this time for the diagnosis or management of patients with type 1 diabetes.

Environment

Reports describe that environmental factors are involved in initiating diabetes. Viruses, such as rubella, mumps, enterovirus, and Coxsackie virus B, have been implicated. It seems likely that autoimmunity to β-cells is initiated by a viral protein (that shares an amino acid sequence with a β-cell protein) or some other environmental insult. Genetic susceptibility and other host factors (e.g., HLA type) determine the progression of the β-cell destruction. Epidemiological studies have implicated early exposure to cows' milk as a trigger of type 1 diabetes. This model is contentious and has been debated for decades.

PATHOGENESIS OF TYPE 2 DIABETES MELLITUS

At least two major identifiable pathological defects have been reported in patients with type 2 diabetes. One is a decreased ability of insulin to act on peripheral tissue. This is called *insulin resistance*. The other is β-*cell dysfunction,* which is an inability of the pancreas to produce sufficient insulin to compensate for the insulin resistance. Thus, a relative deficiency of insulin occurs early in the disease and absolute insulin deficiency late in the disease. The debate over whether type 2 diabetes is due primarily to a defect in β-cell secretion or to peripheral resistance to insulin, or to both, has been raging for decades. Nevertheless, type 2 diabetes is an extremely heterogeneous disease, and no single cause is adequate to explain the progression from normal glucose tolerance to diabetes. The fundamental molecular defects in insulin resistance and insulin secretion result from a combination of environmental and genetic factors.

Loss of β-Cell Function

Increased β-cell demand induced by insulin resistance is ultimately associated with progressive loss of β-cell function that is necessary for the development of fasting hyperglycemia. The major defect is a loss of glucose-induced insulin release, which is termed *selective glucose unresponsiveness.* Hyperglycemia appears to render the β-cells increasingly unresponsive to glucose (called glucotoxicity), and the degree of dysfunction correlates with both glucose concentration and duration of hyperglycemia. Restoration of euglycemia rapidly resolves the defect. Increased free fatty acids in serum have also been implicated in β-cell failure (lipotoxicity). Other insulin secretory abnormalities in type 2 diabetes include disruption of the normal pulsatile release of insulin and an increased ratio of plasma proinsulin to insulin. The number of β-cells is also reduced in patients with type 2 diabetes.

Insulin Resistance

Insulin resistance is defined as "a decreased biological response to normal concentrations of circulating insulin"; it is found in obese, nondiabetic individuals and in patients with type 2 diabetes. The underlying pathophysiologic defect(s) has (have) not been identified, but insulin resistance is usually attributed to a defect in insulin action. Measurement of insulin resistance in a routine clinical setting is difficult, and surrogate measures, namely, fasting insulin concentration or the euglycemic insulin clamp, are used to provide an indirect assessment of insulin function. The euglycemic clamp is performed in hospital under close supervision. The subject receives a constant intravenous infusion of insulin in one arm with concurrent intravenous infusion of variable amounts of glucose in the other arm to maintain blood glucose at a normal fasting concentration. A broad clinical spectrum of insulin resistance ranges from euglycemia (with marked increase in endogenous insulin) to hyperglycemia (despite large doses of exogenous insulin). Several rare clinical syndromes are also associated with insulin resistance. The prototype is the type A insulin resistance syndrome, which is characterized by hyperinsulinemia, acanthosis nigricans, and ovarian hyperandrogenism (see more on polycystic ovarian syndrome in Chapter 43).

The insulin resistance syndrome (also known as syndrome X, or the metabolic syndrome) is a constellation of associated clinical and laboratory findings, consisting of insulin resistance, hyperinsulinemia, obesity, dyslipidemia (high triglyceride and low high-density lipoprotein [HDL] cholesterol), and hypertension. Individuals with this syndrome are at increased risk for cardiovascular disease. The metabolic syndrome is diagnosed if three or more of the following criteria are met:

- Increased waist circumference (population- and country-specific), for example, Europe, Canada, and the United States: greater than 35 inches (>88 cm) (women) or greater than 40 inches (>102 cm) (men).
- Triglycerides greater than 150 mg/dL (>1.7 mmol/L).
- HDL cholesterol less than 51 mg/dL (<1.3 mmol/L) (women) or less than 39 mg/dL (<1.0 mmol/L) (men).
- Blood pressure greater than or equal to 130/85 mm Hg.
- FPG ≥100 mg/dL (≥5.6 mmol/L).

The "metabolic syndrome" concept has been questioned by several experts, including the person who first described it and major clinical diabetes organizations. A WHO Expert Consultation concluded that although the metabolic syndrome may be useful as an educational concept, it has limited practical utility for diagnosis and management, and further efforts to redefine it are inappropriate.

Environment

Environmental factors, such as diet and exercise, are important determinants in the pathogenesis of type 2 diabetes. Convincing evidence links obesity to the development of type 2 diabetes, but the association is complex. Although 60% to 80% of patients with type 2 diabetes are obese, diabetes develops in greater than 15% of obese

individuals. In contrast, virtually all obese subjects, even those with normal carbohydrate tolerance, have hyperinsulinemia and are insulin resistant. Other factors, such as family history of type 2 diabetes (genetic predisposition), the duration of obesity, and the distribution of fat are important. Nevertheless, the rising prevalence of diabetes is believed to be a consequence of the increase in obesity (defined as a body mass index ≥30 kg/m^2), which was reported to be 37.7% in US adults in 2014.

An inverse relationship has been noted between the degree of physical activity and the prevalence of type 2 diabetes. For every 500 kcal increase in daily energy expenditure, a 6% decrease in age-adjusted risk of type 2 diabetes occurs. This effect is independent of both body weight and a parental history of diabetes. The mechanism of the protective effect of exercise is thought to be increased sensitivity to insulin in skeletal muscle and adipose tissue.

Type 2 Diabetes Susceptibility Genes

Genetic factors contribute to the development of type 2 diabetes. For example, the concordance rate for type 2 diabetes in identical twins approaches 100%. Type 2 diabetes is 10 times more likely to occur in an obese person with a parent who has diabetes than in an equally obese person without a diabetic family history. However, the mode of inheritance is unknown, and type 2 diabetes has been described as a "geneticist's nightmare." It is genetically more complex than Mendelian disorders and is not inherited according to simple Mendelian rules. Multiple genetic factors interact with exogenous influences (such as environmental factors) to produce the phenotype.

Multiple factors complicate the search for susceptibility genes in type 2 diabetes. A variety of approaches have produced several genes that are associated with type 2 diabetes. Recent genome-wide association studies (GWAS) have substantially contributed to our understanding of the genetic architecture of type 2 diabetes, with more than 60 genetic loci now identified. Most of these genetic loci are associated with the insulin secretion pathway, rather than with insulin resistance. Despite considerable effort to identify the genetic basis of type 2 diabetes mellitus, genetic defects identified to date account for only approximately 5% of patients with type 2 diabetes. Therefore, the gene or genes causing common forms of type 2 diabetes remain unknown. Moreover, the risk alleles in these loci all have relatively small effects (odds ratios, 1.1 to 1.3). Combined analysis of 48 different type 2 risk loci did not significantly affect the time from diagnosis to prescription of the first drug.

DIAGNOSIS OF DIABETES

The diagnostic criteria of diabetes mellitus are shown in Box 33.2. For type 1 diabetes, the diagnosis is usually easy because hyperglycemia appears abruptly, is severe, and is accompanied by serious metabolic derangements. Diagnosis of type 2 diabetes may be difficult because hyperglycemia often is not severe enough for the patient to notice symptoms of diabetes. Nevertheless, the risk of complications makes it important to identify people early with the disease.

BOX 33.2 Criteria for the Diagnosis of Diabetes Mellitus

Any one of the following is diagnostic:

1. Hemoglobin A_{1c} (HbA_{1c}) ≥6.5% (≥48 mmol/mol)[a] OR
2. Fasting plasma glucose (FPG) ≥126 mg/dL (≥7.0 mmol/L)[b] OR
3. Symptoms of hyperglycemia and random plasma glucose ≥200 mg/dL (≥11.1 mmol/L)[c] OR
4. 2-hour plasma glucose ≥200 mg/dL (≥11.1 mmol/L) during an oral glucose tolerance test (OGTT)[d]

In the absence of unequivocal hyperglycemia, these criteria should be confirmed by repeating the same test on a different day. Mixing different methods to diagnose diabetes should be avoided.

[a]The test should be performed in a laboratory using a method that is National Glycohemoglobin Standardization Program (NGSP)-certified and standardized to the DCCT assay. Point-of-care assays should not be used for diagnosis.
[b]Fasting is defined as no calorie intake for at least 8 h.
[c]The classic symptoms of hyperglycemia include polyuria, polydipsia, and unexplained weight loss. Random is defined as any time of day without regard to time since last meal.
[d]The OGTT should be performed as described by the World Health Organization (WHO), using a glucose load containing the equivalent of 75 g of anhydrous glucose dissolved in water.
From the American Diabetes Association. (2017). Classification and diagnosis of diabetes. *Diabetes Care, 40* (Suppl 1), S11–S24.

Fasting Plasma Glucose Concentrations

FPG concentrations of 126 mg/dL (7.0 mmol/L) or greater on more than one occasion are diagnostic of diabetes mellitus (see Box 33.2). However, some investigators believe that fasting hyperglycemia may be a relatively late development in the course of type 2 diabetes, delaying the diagnosis and leading to underestimation of the prevalence of diabetes mellitus in the population. Complications, such as retinopathy, proteinuria, and neuromuscular disease, are present in approximately 30% of patients at clinical diagnosis of type 2 diabetes, and onset of type 2 diabetes probably occurs at least 4 to 7 years before clinical diagnosis. Screening of high-risk individuals for diabetes is now recommended by some clinical organizations. Fasting glucose (or HbA_{1c}) should be measured in all asymptomatic persons at age 45 (or younger in subjects at increased risk), with follow-up testing every 3 years (see discussion later in this chapter). However, no published evidence indicates that treatment based on screening is efficacious.

Oral Glucose Tolerance Test

Although more sensitive than FPG determinations, glucose tolerance testing is affected by multiple factors that result in *poor reproducibility* (Box 33.3). Moreover, approximately 20% of OGTTs fall into the nondiagnostic category (e.g., only one blood sample exhibits increased glucose concentration).

BOX 33.3 Factors Other Than Diabetes That Influence the Oral Glucose Tolerance Test

Patient Preparation
Duration of fast
Prior carbohydrate intake
Medications (e.g., thiazides, oral contraceptives, corticosteroids)
Trauma
Intercurrent illness
Age
Activity
Weight

Administration of Glucose
Form of glucose (anhydrous or monohydrate)
Quantity of glucose ingested
Volume of glucose administered
Rate of ingestion

During the Test
Posture
Anxiety
Caffeine
Smoking
Activity
Time of day
Sample preservation

Unless results are grossly abnormal initially, the OGTT should be performed on two separate occasions to establish the diagnosis of diabetes.

The following conditions should be met before an OGTT is performed: discontinue, when possible, medications known to affect glucose tolerance; perform in the morning after 3 days of unrestricted diet (containing at least 150 g of carbohydrate per day) and activity and perform the test after a 10- to 16-hour fast only in ambulatory outpatients (bed rest impairs glucose tolerance), who should remain seated during the test without smoking cigarettes. Glucose tolerance testing should not be performed on hospitalized, acutely ill, or inactive patients. The test should begin between 7.00 a.m. and 9.00 a.m.. Venous plasma glucose should be measured after fasting, then 2 hours after an oral glucose load. For adults, the recommended glucose load is 75 g, which may not be a maximum stimulus; for children, 1.75 g/kg, up to 75 g maximum is given. The glucose should be dissolved in 300 mL of water and ingested over 5 minutes. A commercial, more palatable form of glucose may be ingested, but whether the anhydrous or monohydrate form of glucose should be used is still in question.

An OGTT is not widely used in clinical practice for the diagnosis of diabetes in the United States because of its lack of reproducibility and inconvenience. The sensitivity of FPG concentrations is lower than the OGTT for diagnosing diabetes, and some authors claim that the OGTT better identifies patients at risk for developing complications of diabetes. An FPG value less than 100 mg/dL (<5.6 mmol/L) or a random glucose concentration less than less than 140 mg/dL (<7.8 mmol/L) is sufficient to rule out the diagnosis of diabetes mellitus.

DIAGNOSIS OF GESTATIONAL DIABETES MELLITUS

Normal pregnancy is associated with increased insulin resistance, especially in the late second and third trimesters. Euglycemia is maintained by increased insulin secretion, with GDM developing in those women who fail to augment insulin sufficiently. Risk factors for GDM include a family history of diabetes in a first-degree relative, obesity, advanced maternal age, glycosuria, and selected adverse outcomes in a previous pregnancy (e.g., stillbirth, macrosomia). Recommendations for laboratory diagnosis of GDM remain controversial and lack consensus. The Hyperglycemia in Pregnancy and Adverse Outcome (HAPO) study was a large (≈25,000 pregnant women), prospective, multinational study published in 2008. The study revealed strong, graded, linear associations between maternal glycemia and adverse outcomes. Based on these data, it was recommended that all pregnant women not previously known to have diabetes should be evaluated for GDM by a 75-g OGTT at 24 to 28 weeks' gestation. Diagnostic cutpoints for fasting, 1- and 2-hour plasma glucose concentrations have been established (Table 33.2).

These new criteria, which increase the incidence of GDM by approximately 2.5-fold, have not been accepted by all clinical organizations around the globe. For example, in the United States, ACOG continues to advocate prior diagnostic criteria (see Table 33.2). These state that all pregnant women should be evaluated by a two-step method:

1. Screen by a 50-g oral glucose load (the patient does not need to be fasting). A plasma glucose value greater than or equal to 140 mg/dL (7.7 mmol/L) at 1 hour after glucose ingestion indicates the necessity for definitive testing. Approximately 15% of pregnant women meet this criterion and require a full OGTT. This subgroup includes approximately 80% of all women with GDM. Some experts recommend a cutoff of 130 mg/dL, which increases sensitivity for GDM to 90% but includes approximately 25% of all pregnant women.
2. The diagnosis of GDM can be made on the basis of the result of the 100-g, 3-hour OGTT. This should be performed on a different day from the 50-g OGTT.

Although usually asymptomatic and not life-threatening to the mother, GDM is associated with increased neonatal mortality and morbidity, including hypocalcemia, hypoglycemia, and macrosomia. Maternal hyperglycemia causes the fetus to secrete more insulin, resulting in stimulation of fetal growth and macrosomia. Recognition is important because therapy can reduce perinatal morbidity and mortality. Maternal complications include a high rate of cesarean delivery and

TABLE 33.2 Screening for and Diagnosis of Gestational Diabetes Mellitus

I One-step
1. Perform between 24 and 28 weeks' gestation in the morning after an overnight fast of at least 8 h.
2. Measure fasting venous plasma glucose.
3. Give 75 g of glucose orally.
4. Measure plasma glucose hourly for 2 h after glucose is given.
5. At least one value must meet or exceed the following:

	Glucose Concentration
Fasting	92 mg/dL (5.1 mmol/L)
1 h	180 mg/dL (10.0 mmol/L)
2 h	153 mg/dL (8.5 mmol/L)

II Two-step
A. Screening (step 1)
1. Perform between 24 and 28 weeks' gestation[a] in all pregnant women not previously diagnosed with diabetes.
2. Give 50 g oral glucose load without regard to time of day or time of last meal.
3. Measure venous plasma glucose at 1 h.
4. If glucose is ≥140 mg/dL (7.8 mmol/L)[b], proceed to Step 2 and perform glucose tolerance test.

B. Diagnosis (step 2)
1. Perform in the morning after an overnight fast of at least 8 h.
2. Measure fasting venous plasma glucose.
3. Give 100 g of glucose orally.
4. Measure plasma glucose hourly for 3 h after glucose is given.
5. At least two values must meet or exceed the following:

	Carpenter & Coustan	NDDG
Fasting	95 mg/dL (5.3 mmol/L)	105 mg/dL (5.8 mmol/L)
1 h	180 mg/dL (10.0 mmol/L)	190 mg/dL (10.6 mmol/L)
2 h	155 mg/dL (8.6 mmol/L)	165 mg/dL (9.2 mmol/L)
3 h	140 mg/dL(7.8 mmol/L)	145 mg/dL (8.0 mmol/L)

NDDG, National Diabetes Data Group.
[a]The WHO states that this test can be performed at any time during pregnancy.
[b]Some experts recommend a cutoff of 135 mg/dL (7.5 mmol/L).
Modified from the American Diabetes Association. (2017). Classification and diagnosis of diabetes. *Diabetes Care, 40* (Suppl 1), S11–S24.

hypertension. In addition, mothers with GDM are at significantly increased risk of subsequent type 2 diabetes.

Distinct from GDM is pregnancy in a patient with preexisting diabetes. This is associated with an increased incidence of congenital malformation, but meticulous glycemic control during the first 8 weeks of pregnancy can significantly decrease the risk of congenital malformation. Tight control results in an increased incidence of maternal hypoglycemia, which is teratogenic in animals but does not cause malformation in humans.

Women with GDM should be screened for diabetes 6 to 12 weeks postpartum using nonpregnant OGTT criteria (see Box 33.2). (HbA_{1c} is not recommended because of the antepartum therapy for hyperglycemia.) If glucose values are normal, glycemia should be reassessed at least every 3 years using glucose or HbA_{1c}.

CHRONIC COMPLICATIONS OF DIABETES MELLITUS

Patients with type 1 or type 2 diabetes are at high risk for the development of chronic complications. Diabetes-specific microvascular pathology in the retina, renal glomeruli, and peripheral nerves produces retinopathy, nephropathy, and neuropathy, respectively. As a result, diabetes is the most frequent cause of new cases of blindness in the industrialized world in persons between 25 and 74 years of age and is the leading cause of end-stage renal disease. Diabetes is also associated with a marked increase in atherosclerotic macrovascular disease involving cardiac, cerebral, and peripheral large vessels. Therefore, patients with diabetes have a high rate of myocardial infarction (the major cause of mortality in diabetes), stroke, and limb amputation.

Type 1 Diabetes

Although it was theorized for many years that better glycemic control would decrease rates of long-term complications of diabetes, it was not until the publication of the Diabetes Control and Complications Trial (DCCT) in 1993 that this hypothesis was verified. The DCCT was a multicenter, randomized trial that compared the effects of intensive and conventional therapy on the development and progression of complications in 1441 patients with type 1 diabetes, all of whom required insulin. During the study period, which averaged 6.5 years, intensively managed patients maintained significantly lower mean blood glucose concentrations. Compared with conventional therapy, intensive therapy reduced the risks of retinopathy, nephropathy, and neuropathy by 40% to 75%. Intensive therapy delayed the onset and slowed the progression of these complications, regardless of age, gender, or duration of diabetes. Absolute risks of retinopathy and nephropathy were proportional to the mean HbA_{1c} (discussed later in the chapter). Although intensive therapy also reduced the development of hypercholesterolemia, macrovascular complications were not significantly decreased in the initial assessment. However, analysis after 17 years of follow-up documented a 42% lower incidence of cardiovascular disease in the intensively treated group. This landmark study has had a considerable impact on therapeutic goals and comprehension of the pathogenesis of complications of diabetes.

At the conclusion of the DCCT, 95% of participants enrolled in the long-term follow-up study, termed the Epidemiology of Diabetes Interventions and Complications (EDIC). Five years after the end of the DCCT, no difference in metabolic control (as assessed by HbA_{1c} measurements) was noted between the former conventional and intensively treated groups. Nevertheless, further progression of retinopathy and neuropathy was significantly lower in the former

intensive group, demonstrating that the beneficial effects of intensive treatment persisted for at least several years beyond the period of strictest intervention.

Type 2 Diabetes

The role of hyperglycemia in the development of complications in individuals with type 2 diabetes was established in the United Kingdom Prospective Diabetes Study (UKPDS). The UKPDS was a major, randomized, multicenter clinical study that included 5102 patients with newly diagnosed type 2 diabetes who were followed for an average of 10 years. Analogous to the findings of the DCCT, the UKPDS demonstrated that in patients with type 2 diabetes, intensive treatment diminishes the development of microvascular complications by approximately 10% to 40%. Intensive treatment also decreased the rate of occurrence of macrovascular complications. Although the reduction was not statistically significant initially, follow-up 10 years after the study ended showed a significant reduction in myocardial infarction among patients who had received intensive therapy. Analogous to the EDIC findings, long-term benefits for microvascular complications were observed with follow-up of patients in the UKPDS despite loss of glycemic separation between intensive and standard cohorts after the study ended. An important caveat of both the DCCT and the UKPDS was that intensive therapy produced a 3-fold increase in the incidence of severe hypoglycemia.

ROLE OF THE CLINICAL LABORATORY IN DIABETES MELLITUS

The clinical laboratory has a vital role in both the diagnosis and management of diabetes (Table 33.3). The National Academy of Clinical Biochemistry (NACB) published evidence-based guidelines for laboratory analysis in diabetes. These guidelines were reviewed and endorsed by the ADA in 2011 as a Position Statement. A brief overview is presented.

Diagnosis

Preclinical (Screening)

Several large clinical trials are underway to assess a variety of therapeutic strategies designed to delay or prevent the onset of type 1 diabetes. Until effective intervention becomes available and cost-effective screening strategies are developed for young children, screening for antibodies and determining HLA type is not currently warranted, except in clinical research studies. A decrease in glucose-stimulated insulin secretion is the first functional abnormality in both type 1 and type 2 diabetes. Nevertheless, tests of insulin secretion are not currently recommended for routine clinical use.

Screening asymptomatic individuals for type 2 diabetes has been the subject of much controversy. The ADA, which previously did not support screening, now advocates screening in all asymptomatic individuals over the age of 45 years. If the HbA_{1c} is less than 5.7% (39 mmol/mol) or the FPG is less than 100 mg/dL (5.6 mmol/L), testing should be repeated at 3-year intervals.

TABLE 33.3 Role of the Laboratory in the Management of Diabetes Mellitus

Diagnosis	
Preclinical (Screening)	Immunologic markers
	ICA
	IAA
	GAD antibodies
	Protein tyrosine phosphatase antibodies (IA-2)
	Zinc transporter ZnT8 antibodies
	Genetic markers (e.g., HLA)
	Insulin secretion
	Fasting
	Pulses
	In response to a glucose challenge
	Blood glucose
	OGTT
	HbA_{1c}
Clinical	Blood glucose
	OGTT
	HbA_{1c}
	Ketones (urine and blood)
	Other (e.g., insulin, C-peptide, stimulation tests)
Management	
Acute	Glucose
	Blood
	Urine
	Ketones
	Blood
	Urine
	Acid-base status (pH), bicarbonate
	Lactate
	Other abnormalities related to cellular dehydration or therapy (e.g., potassium, sodium, phosphate, osmolality)
Chronic	Glucose
	Blood (fasting, random)
	Urine
	Glycated proteins
	HbA_{1c}
	Fructosamine
	Glycated serum albumin
	1,5-AG
	Urinary protein
	Albuminuria (previously termed "microalbuminuria")
	Proteinuria
	Evaluation of complications (e.g., creatinine, cholesterol, triglycerides)
	Evaluation of pancreas transplant (C-peptide, insulin)
	Eligibility for insulin pump (C-peptide)

1,5-AG, 1,5-Anhydroglucitol; *HbA_{1c},* hemoglobin A_{1c}; *HLA,* human leukocyte antigen; *IAA,* insulin autoantibodies; *ICA,* islet cell cytoplasmic antibodies; *OGTT,* oral glucose tolerance test.

Clinical

The laboratory diagnosis of diabetes is made exclusively by the demonstration of hyperglycemia, as evidenced by measurements of venous plasma glucose or HbA_{1c}. Although other tests (e.g., C-peptide, insulin analysis) have been proposed to assist in the diagnosis and classification of the disease, they do not at present have a role outside of research studies.

Management

Acute

In diabetic ketoacidosis, hyperosmolar nonketotic coma, and hypoglycemia, the clinical laboratory has an essential role in both diagnosis and monitoring of therapy. Several analytes are frequently measured to guide clinicians in treatment regimens to restore euglycemia and correct other metabolic disturbances. The metabolic abnormalities of these conditions are beyond the scope of this book, and interested readers are referred to a standard textbook of medicine. The NACB guidelines also provide information on the tests that are used.

Chronic

The DCCT and UKPDS studies documented a correlation between blood glucose concentrations and the development of long-term complications of diabetes. Measurement of glucose and glycated proteins provides an index of short- and long-term glycemic control, respectively (see section on glycated proteins later in the chapter). Detection of and monitoring for complications are achieved by assaying creatinine, urine albumin, and serum lipids. The success of newer therapies, such as islet cell or pancreas transplantation, can be monitored by measuring serum C-peptide or insulin concentrations.

SELF-MONITORING OF BLOOD GLUCOSE

Patients with diabetes, especially those who need insulin, require careful monitoring to maintain control of blood glucose. Therapeutic regimens include multiple daily insulin injections, insulin pumps, and continuous subcutaneous insulin injections. *Testing urine for glucose is not adequate for monitoring patients on insulin therapy.* Although some evidence suggests that it may be effective for monitoring type 2 diabetes, the ADA states that limitations of urine testing make blood glucose measurements the preferred method for assessing glycemic control.

Portable meters for measurement of blood glucose concentrations are used in three major settings: in acute and chronic care facilities (at the patient's bedside and in clinics or hospitals); in physicians' offices, and by patients or their caregivers at home, work, and school. Patients measure their own blood glucose concentrations and modify their insulin doses based on the result. In 2006, the overall rate of daily self-monitoring of blood glucose (SMBG) was 63.4% among all adults with diabetes in the United States and 86.7% among those treated with insulin.

A large number of simple test strips permit rapid glucose measurements on a drop of whole blood. More than 80 different blood glucose meters are commercially available. These meters vary in size, weight, calibration method, and other features, and use the same enzyme-catalyzed reactions, predominantly glucose oxidase or glucose dehydrogenase, as described in Chapter 22 for glucose analysis. The reagents are combined in dry form on a small surface area of a test strip.

To perform the measurement, a sample of blood (usually from a fingerstick, but anticoagulated whole blood collected in ethylenediaminetetraacetic acid [EDTA] or heparin may be used) is placed on the test pad, which is attached to a plastic support. The test strip is then inserted into the meter. (In some devices, the strip is inserted into the meter before the sample is applied.) After a fixed time, the result appears on a digital display screen. These meters use reflectance photometry or electrochemistry to measure the rate of the reaction or the final concentration of the products. Reflectance photometry measures the amount of light reflected from a test pad containing reagent. In electrochemical systems, the enzymatic reaction produces a flow of electrons in an electrode incorporated on the test strip. The current, which is directly proportional to the amount of glucose in the sample, is converted to a digital readout. Large variability has been noted among meters as to the test time (5 to 45 seconds) and the claimed reading range (30 to 500 mg/dL to 0 to 600 mg/dL: 1.7 to 28 mmol/L to 0 to 33 mmol/L). Calibration is automatic on some devices, whereas others use lot-specific code chips or strips. All manufacturers supply control solutions. Recent advances in technology facilitate data analysis and sharing. Some meters have Bluetooth capabilities (enabling transmission of data to a smartphone or computer), cellular connections that automatically send data to the "cloud," USB ports, and/or communication with an insulin pump. Strict adherence to the instructions is necessary to obtain accurate results. Some meters have a porous membrane that separates erythrocytes, and analysis is performed on the resultant plasma. Whole blood glucose concentrations are approximately 10% to 15% lower than plasma or serum concentrations, but meters can be calibrated to report plasma glucose values, even when the sample is whole blood. An International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) working group recommended that glucose meters be harmonized using a factor of 1.11× to report the concentration of glucose in plasma, irrespective of the sample type or technology.

Analytical Goals for Glucose Meters

Multiple analytical goals have been proposed for the performance of glucose meters. The recommendations promulgated in 2002 by the Clinical and Laboratory Standards Institute (CLSI) (previously called the National Committee for Clinical Laboratory Standards [NCCLS]) and in 2003 by the International Organization for Standardization (ISO) are that 95% of results should fall within 20% of laboratory-measured glucose concentrations when greater than 75

mg/dL (>4.2 mmol/L) and within 15 mg/dL (0.83 mmol/L) of a laboratory glucose measurement if the glucose concentration is less than 75 mg/dL (<4.2 mmol/L). In both CSLI and ISO guidelines, 5% of these results can be considerably outside these limits. Note that the CLSI guideline is for meter use in acute and chronic care facilities (mainly hospitals), while the ISO guidelines pertain to meters for self-testing (i.e., SMBG). Many experts believe these acceptance criteria are too wide. The CLSI and ISO documents were revised in 2013, with tightening of acceptance criteria. The current CLSI guideline POCT12-A3 indicates that for 95% of the samples, the difference between meter and laboratory measurement should be less than 12.5% when the laboratory glucose value is greater than 100 mg/dL (≥5.6 mmol/L) and less than 12 mg/dL (<0.67 mmol/L) when the glucose concentration is less than 100 mg/dL (<5.6 mmol/L). In addition, no more than 2% of results can differ by more than 20% at greater than 75 mg/dL (≥4.2 mmol/L) and by greater than 15 mg/dL (>0.83 mmol/L) when the glucose concentration is less than 75 mg/dL (<4.2 mmol/L). The revised ISO goals, also issued in 2013, are that for 95% of the samples, the difference between meter and laboratory measurement should be less than 15% when the laboratory glucose value is greater than 100 mg/dL (≥5.6 mmol/L) and less than 15 mg/dL (<0.83 mmol/L) when the glucose concentration is less than 100 mg/dL (<5.6 mmol/L). Moreover, 99% of results must be within zones A and B of the consensus error grid (see following paragraph). In the United States, the FDA issued guidance in 2016 for glucose meters. Two separate documents were released, one for home (SMBG) use and the other for hospitals. The analytic goals are more stringent than either the CLSI or ISO guidelines.

A different method uses an error grid to try to define clinically important errors by identifying fairly broad target ranges. An approach using simulation modeling concluded that meters that achieve both a CV and a bias less than 5% rarely lead to major errors in insulin dosage. Lack of consensus on quality goals for glucose meters reflects the absence of agreed-upon objective criteria. When biological variation criteria are used, a goal for total error (including both bias and imprecision) of 6.9% or less has been proposed.

Glucose meters are also used to calculate insulin dose in patients without diabetes on tight glucose control protocols in intensive care units (ICUs). Evidence in 2001 showed that intensive insulin therapy significantly reduced mortality and morbidity of critically ill patients in the surgical ICU. Although some subsequent studies were unable to replicate these findings, aggressive glucose control with intravenous infusion of insulin is used extensively in hospital ICUs. Many factors, such as hypoxia, shock, and low hematocrit, are common in these patients and can compromise glucose analysis in capillary blood samples. The use of glucose meters in these settings has been questioned by some experts. Some meters may have adequate analytical performance (as noted above), but the use of skin puncture (fingerstick) samples introduces serious error in patients who have conditions such as shock.

Performance of Glucose Meters

The most common errors in SMBG, such as proper application, timing, and removal of excess blood, have been reduced by advances in technology but can still occur. Additional innovations that reduce operator error include (1) systems that abort testing if the sample volume is inadequate, (2) built-in programs that simplify quality control, and (3) increased memory that allows the instrument to store up to several hundred glucose readings that can be downloaded into a computer.

Several factors affect the accuracy and reproducibility of SMBG. These include (1) user variability—up to 50% of values may vary by more than 20% from reference values; (2) hematocrit—the presence of anemia (false increase) or polycythemia (false depression) may result in up to 30% variability; and (3) defective reagent strips or instrument malfunction (rare). Other variables include changes in altitude, environmental temperature, or humidity; hypotension; hypoxia; and high triglyceride concentrations. In addition, *these assays are unreliable at very high and very low glucose concentrations* (<60 and >500 mg/dL; <3.3 and >28 mmol/L). Because dehydration, a common feature of diabetic ketoacidosis, greatly increases blood viscosity, inaccurately low blood glucose results may be obtained. Several drugs interfere, but not with all meters. Another important factor is the lack of correlation among meters, even from a single manufacturer, caused by different assay methods and architecture. Moreover, results from 2 m of the same brand have been observed to differ substantially. Patient factors are also important, particularly adequate training when the meter is used for SMBG. Recurrent education at clinic visits and comparison of SMBG with concurrent laboratory glucose analysis improved the accuracy of patients' blood glucose readings. It is important to evaluate the patient's technique at regular intervals.

Meter performance varies widely. Under carefully controlled conditions in which all assays were performed by a single medical technologist, approximately 50% of analyses met the ADA criterion of less than 5% deviation from reference values. Performance of older meters was substantially worse. Note that medical technologists perform better than patients. Comparison with laboratory values of almost 22,000 measurements of capillary glucose by patients using meters revealed no significant improvement in meter performance between 1989 and 1999.

An analysis in 2012 revealed that only 18 of 34 (53%) glucose meters approved for use in Europe fulfilled the minimum accuracy requirements of the 2013 ISO standard. The imprecision of most meters precludes their use in the diagnosis of diabetes.

ALTERNATIVES TO METERS FOR MONITORING OF BLOOD GLUCOSE

A major limitation to performing SMBG is that it is painful and inconvenient. Since the 1960s, attempts have been made to develop a painless method for monitoring blood glucose

concentrations. Three general approaches have been used, namely, implanted sensors, minimally invasive monitoring, and noninvasive monitoring.

Implanted Sensors

Several implanted biosensors have been developed. Detection systems are based on enzymes, electrodes, or fluorescence. The most widely studied method is an electrochemical sensor that is usually implanted subcutaneously. All monitoring devices use glucose oxidase to measure glucose every 1 to 5 minutes. The values are sent to a monitor. Results are recorded, can be downloaded later in the physician's office, and are usually available to the patient in real time. These devices are subject to some limitations. Implantation of a needle type of sensor into the subcutaneous tissue induces inflammatory responses in the host, requiring sensors to be replaced every 7 to 14 days. The sensors require calibration by the user initially and at least twice a day with a glucose meter and are subject to the imprecision of the meter. New sensors have been developed that eliminate calibration by users. In addition, changes in glucose concentration in the interstitial fluid occur 4 to 20 minutes later than changes in the blood. Nevertheless, real-time continuous glucose monitoring improves long-term glycemic control in a subset of patients with type 1 diabetes. Newer features, such as automated suspension of insulin delivery for up to 2 hours when glucose concentrations reach a preset low threshold, have recently been shown to significantly reduce the rate of hypoglycemia and improve HbA_{1c} in patients with type 1 diabetes. Automated closed-loop insulin delivery (also termed the "artificial pancreas") are a focus of intense research. The system employs a control algorithm that modulates insulin delivery according to real-time interstitial glucose measured by a sensor. Closed-loop systems have been shown to be superior to conventional insulin pump therapy.

Minimally Invasive Glucose Monitoring

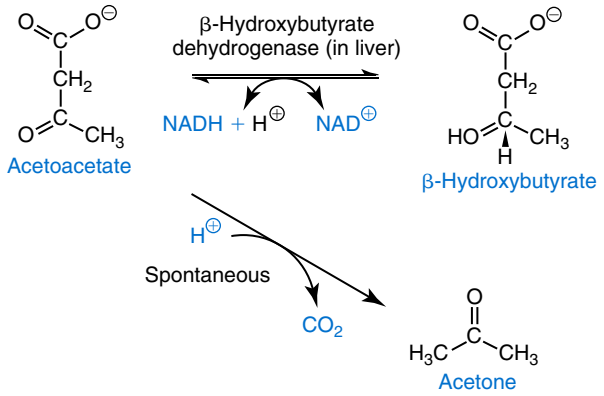
The concept underlying these methods is that the concentration of glucose in the interstitial fluid correlates with the blood glucose concentration. The principle of the GlucoWatch Biographer involves the application of a low-level electrical current to the skin. This induces movement by electroosmosis of glucose across the skin, where it is measured by a glucose oxidase detector. Glucose concentrations in transdermal fluid and plasma are highly correlated and studies revealed reasonable correlation of the GlucoWatch with SMBG. The clearest application of the GlucoWatch appears to be in the detection of unsuspected hypoglycemia. This device has been withdrawn from the market, but it is likely to stimulate enhanced efforts to bring other technologies into clinical use.

Noninvasive Glucose Monitoring

Noninvasive in vivo monitoring of glucose has been an area of active investigation for many years. Near-infrared spectroscopic devices measure absorption or reflection of light from subcutaneous tissue. Although glucose has a specific absorption at 1035 nm, many substances interfere. A computer, individually calibrated, screens out interfering information to obtain the glucose result. Alternative approaches include Raman scattering spectroscopy and photoacoustic spectroscopy. Notwithstanding the investment of considerable resources, no noninvasive sensing technology is approved by the FDA for glucose measurement in patients.

KETONE BODIES

The development of ketosis requires changes in both adipose tissue and the liver. *The primary substrates for ketone body formation are free fatty acids* from adipose stores. Normally, long-chain fatty acids are taken up by the liver, reesterified to triglycerides, and stored in the liver or incorporated in very low-density lipoproteins and returned to the plasma. In contrast to other tissues, the brain cannot use free fatty acids for energy. When glucose is unavailable, ketone bodies supply the vast majority of the brain's energy. After a 3-day fast, ketone bodies provide 30% to 40% of the body's energy requirements. In uncontrolled diabetes, the low insulin concentrations result in increased lipolysis and decreased reesterification, thereby increasing plasma free fatty acids. In addition, the increased glucagon/insulin ratio enhances fatty acid oxidation in the liver. Increased counterregulatory hormones also augment lipolysis and ketogenesis in fat and liver, respectively. Thus, increased hepatic ketone production and decreased peripheral tissue metabolism lead to acetoacetate accumulation in the blood. A small fraction undergoes spontaneous decarboxylation to form acetone, but most of it is converted to β-hydroxybutyrate.



The relative proportions in which the three ketone bodies are present in blood vary, depending on the redox state of the cell. In healthy people, β-hydroxybutyrate and acetoacetate—which are present at approximately equimolar concentrations—constitute virtually all the serum ketones. Acetone is a minor component. In severe diabetes, the ratio of β-hydroxybutyrate to acetoacetate may increase to 6:1 because of the presence of a high concentration of nicotinamide adenine dinucleotide (NADH), which favors β-hydroxybutyrate production.

Clinical Significance of Ketones

Excessive formation of ketone bodies results in increased blood concentrations of **ketones** *(ketonemia)* and increased excretion of ketones in the urine *(ketonuria)*. This process is observed in conditions associated with decreased availability of carbohydrates (such as starvation or frequent vomiting) or decreased use of carbohydrates (such as diabetes mellitus, glycogen storage disease [von Gierke disease], and alkalosis). The popular high-fat, low-carbohydrate diets are ketogenic and increase ketone bodies in the circulation. Diabetes and alcohol consumption are the most common causes of ketoacidosis. Urine ketone test results are positive in more than 30% of first-morning-void specimens from pregnant women. Semiquantitative determination of ketone bodies in blood is more accurate than determination of these compounds in urine in the treatment of diabetic ketoacidosis. Although not always excreted in proportion to blood ketone concentrations, urine ketones are widely used for monitoring control in patients with type 1 diabetes because of convenience. Patients with type 1 diabetes should test for ketones during acute illness or stress, with consistent increases in blood glucose (>300 mg/dL; 16.7 mmol/L), during pregnancy, or when symptoms of ketoacidosis are present.

Most of the tests for ketone bodies described in Chapter 22 detect or measure acetoacetate only. This may produce a paradoxical situation. When a patient initially presents in ketoacidosis, the test results for ketones may be only weakly positive. With therapy, β-hydroxybutyrate is converted to acetoacetate, and the ketosis appears to worsen.

GLYCATED PROTEINS

Measurement of glycated proteins, primarily glycated hemoglobin (GHb), is effective in monitoring long-term glucose control in people with diabetes mellitus. It provides a retrospective index of integrated plasma glucose values over an extended period of time and is not subject to the wide fluctuations observed when blood glucose concentrations are assayed. GHb concentrations are therefore a valuable and widely used adjunct to blood glucose determinations for monitoring long-term glycemic control. In addition, GHb has recently been recommended for the diagnosis of diabetes and is a measure of risk for the development of microvascular complications of diabetes.

Glycated Hemoglobins

Glycation is the nonenzymatic addition of a sugar residue to amino groups of proteins. Human adult hemoglobin (Hb) usually consists of HbA (≈97% of the total), HbA_2 (2.5%), and HbF (0.5%). HbA is made up of four polypeptide chains, two α-chains and two β-chains. Chromatographic analysis of HbA identifies several minor hemoglobins, namely, HbA_{1a}, HbA_{1b}, and HbA_{1c}, which are collectively referred to as HbA_1, *fast hemoglobins* (because they migrate more rapidly than HbA in an electrical field), *glycohemoglobins,* or **glycated hemoglobins**. The Joint Commission on Biochemical Nomenclature of the International Union of Pure and Applied Chemistry recommends the term *neoglycoprotein* for such derivatives and the term *glycation* to describe this process. Therefore, although *glycosylated* and *glucosylated* have been widely used in the literature, the term *glycated* is preferred. HbA_{1c} is formed by the condensation of glucose with the N-terminal valine residue of each β-chain of HbA to form an unstable Schiff base (aldimine, pre-HbA_{1c}; Fig. 33.2). The Schiff base may dissociate or may undergo an Amadori rearrangement to form a stable ketoamine, HbA_{1c}. HbA_{1a1} and HbA_{1a2}, which make up HbA_{1a}, have fructose-1,6-diphosphate and glucose-6-phosphate, respectively, attached to the amino terminal of the β-chain. HbA_{1b} contains pyruvic acid linked to the amino terminal valine of the β-chain, probably by a ketamine bond. *HbA_{1c} is the major fraction,* constituting approximately 80% of HbA_1.

Glycation may also occur at sites other than the end of the β-chain, such as lysine residues or the α-chain. The sum of all GHbs, referred to as total glycated hemoglobin, cannot be separated from nonglycated hemoglobin by methods based on charge, but are measured by boronate affinity chromatography.

Formation of GHb is essentially irreversible, and the concentration in the blood depends on both the life span of the red blood cell (RBC; average 120 days) and the blood glucose concentration. Because the rate of formation of GHb is directly proportional to the concentration of glucose in the blood, *the GHb concentration represents integrated values for glucose over the preceding 8 to 12 weeks.* This provides an additional criterion for assessing glucose control because GHb values are free of day-to-day and hour-to-hour glucose fluctuations and are unaffected by exercise or food ingestion immediately

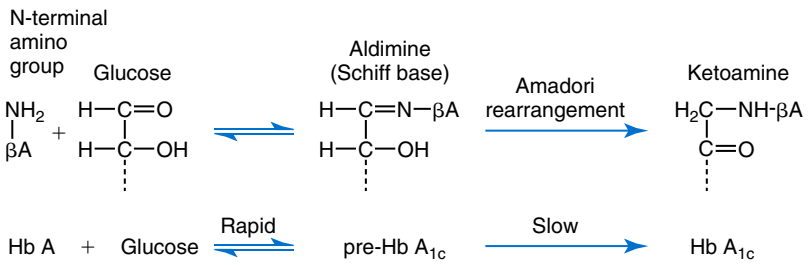


Fig. 33.2 Formation of hemoglobin (Hb)A_{1c}.

before collection of the blood sample. The contribution of the plasma glucose concentration to GHb depends on the time interval, with more recent values (such as during the preceding month) providing a larger contribution than earlier values (such as 3 months earlier). The plasma glucose in the preceding 1 month determines 50% of the HbA_{1c}, whereas days 60 to 120 determine only 25%. After a sudden alteration in blood glucose concentrations, the rate of change in HbA_{1c} is rapid during the initial 2 months, followed by a more gradual change approaching steady state 3 months later.

Interpretation of GHb depends on RBCs having a normal life span. Patients with hemolytic disease or other conditions with shortened RBC survival exhibit a substantial reduction in HbA_{1c}. HbA_{1c} can still be used to monitor these patients, but values must be compared with previous values from the same patient—not with published reference intervals. Similarly, individuals with recent significant blood loss have falsely low values because of a higher fraction of young erythrocytes. High HbA_{1c} concentrations have been reported in iron-deficiency anemia. The effects of hemoglobin variants (such as HbF, HbS, and HbC) depend on the specific method of analysis used (discussed later). Depending on the particular hemoglobinopathy and assay, results may be spuriously increased or decreased. Most manufacturers of HbA_{1c} assays have modified their assays to eliminate interference from many of the common hemoglobin variants. Therefore, HbA_{1c} can be accurately measured by selecting an appropriate instrument, provided the RBC life span is not altered (see www.NGSP.org for additional information). Another source of error in selected methods is *carbamylated hemoglobin,* formed by attachment of urea. It is present in large amounts in renal failure, which is common in people with diabetes. Carbamylated hemoglobin does not interfere with most modern methods. Most interferents produce small effects, and HbA_{1c} can be measured accurately in the vast majority of patients with diabetes.

Hemoglobin A_{1c} in Diagnosis of Diabetes

In 2009, an International Expert Committee advised that HbA_{1c} could be used for the diagnosis of diabetes (see Box 33.2). An HbA_{1c} value greater than 6.5% was selected as the decision point, based on the prevalence of retinopathy. This recommendation has been endorsed by both the ADA and the WHO. HbA_{1c} concentrations 5.7% to 6.4% indicate individuals at high risk of developing diabetes. HbA_{1c} was also recommended as an alternative to glucose for screening for diabetes. This last recommendation has been accepted by the ADA.

Hemoglobin A_{1c} in Monitoring of Diabetes

HbA_{1c} is firmly established as an index of long-term blood glucose concentrations and a measure of the risk for developing microvascular complications in patients with diabetes. Absolute risks of retinopathy and nephropathy are directly proportional to the mean HbA_{1c} concentration. In persons without diabetes, HbA_{1c} is directly related to risk of cardiovascular disease.

Methods for Determination of Glycated Hemoglobins

More than 150 different methods have been described for the determination of GHbs. Most methods separate GHb from nonglycated hemoglobin using techniques based on charge differences (ion-exchange chromatography, HPLC, electrophoresis, and isoelectric focusing) or structural differences (affinity chromatography and immunoassay). Chemical analysis (enzymatic, photometry, and spectrophotometry) has also been used. Recently, methods have become commercially available that use capillary electrophoresis or an enzymatic assay that specifically measure HbA_{1c}. The result is expressed as a percentage of total hemoglobin. Most laboratories in the United States use immunoassay or HPLC and report HbA_{1c}. The selection of method by a laboratory is influenced by several factors, including sample volume, patient population, and cost. It is advisable to consult clinicians in this process. The ADA recommends that laboratories use only HbA_{1c} assays that are certified by the NGSP (previously termed the National Glycohemoglobin Standardization Program) as traceable to the DCCT reference. These assays are listed on the NGSP website (www.NGSP.org/ accessed) and are updated several times a year.

High-performance liquid chromatography. HbA_{1c} and other hemoglobin fractions are separated by HPLC, with cation exchange chromatography. Several fully automated systems are commercially available. Assays require only 5 μL of whole blood, and fingerstick samples can be collected in a capillary tube for analysis. Anticoagulated blood is diluted with a hemolysis reagent containing borate. Samples are incubated at 37°C for 30 minutes to remove the Schiff base and are inserted into the autosampler. (Some instruments have a shorter preincubation step, and others separate labile GHb chromatographically, eliminating the step to remove the Schiff base.) A step gradient using three phosphate buffers of increasing ionic strength is passed through the column. Detection is performed at 415 and 690 nm, and results are quantified by integrating the areas under the peaks. Analysis time is as short as 3 minutes. HbA_{1c} by HPLC was used for analysis of all patient samples in the DCCT.

Immunoassay. Assays for HbA_{1c} have been developed using antibodies raised against the Amadori product of glucose (ketoamine linkage) plus the first few (four to eight) amino acids at the N-terminal end of the β-chain of hemoglobin. A widely used assay measures HbA_{1c} in whole blood by inhibition of latex agglutination. The agglutinator, a synthetic polymer containing multiple copies of the immunoreactive portion of HbA_{1c}, binds the anti-HbA_{1c} monoclonal antibody that is attached to latex beads. This agglutination produces light scattering, measured as an increase in absorbance. HbA_{1c} in the patient's sample competes for the antibody on the latex, inhibiting agglutination, thereby decreasing light scattering. Immunoassays are generally calibrated to give values that match HPLC values. The antibodies do not recognize labile intermediates or other GHbs (such as HbA_{1a} or HbA_{1b})

because both the ketoamine with glucose and specific amino acid sequences are required for binding. Similarly, several hemoglobin variants, such as HbF, HbA_2, HbS, and carbamylated hemoglobin, are not detected. The procedure has been adapted for capillary blood samples using a bench-top analyzer with reagent cartridges designed for use in physicians' office laboratories.

Affinity chromatography. Affinity gel columns are used to separate GHb, which binds to *m*-aminophenylboronic acid on the column, from the nonglycated fraction. The boronic acid reacts with the *cis*-diol groups of glucose that is bound to hemoglobin to form a reversible five-member ring complex, thus selectively holding the GHb on the column. Nonglycated hemoglobin does not bind. Sorbitol is added to elute the GHb. Absorbance of bound and nonbound fractions, measured at 415 nm, is used to calculate the percentage of GHb.

The major advantages of affinity chromatography are as follows: no interference from nonglycated hemoglobins and negligible interference from the labile intermediate form of HbA_{1c}. It is unaffected by variations in temperature and has reasonably good precision. Hemoglobin variants such as HbS, HbC, HbD, or HbE produce little effect.

Affinity methods measure total GHb. This includes components other than HbA_{1c} because the assay detects ketoamine structures on lysine and valine residues on both α- and β-chains of hemoglobin.

Although the method detects all GHbs, most commercial systems are calibrated to report a standardized HbA_{1c} value. The value is derived from an equation obtained from linear regression between total GHb and HbA_{1c} analysis by HPLC. A linear relationship has been demonstrated, and standardized HbA_{1c} values are thus comparable with values obtained by methods specific for HbA_{1c}. Columns and reagents are commercially available.

Capillary electrophoresis. Advantages of capillary electrophoresis include high resolving ability (due to the high voltage that can be applied) and small sample volume (see detail in Chapter 11). Briefly, charged molecules are separated by their electrophoretic mobility in an alkaline buffer (pH 9.4), and by electrolyte pH and electroosmotic flow. Hemoglobins are detected by absorption spectroscopy at the cathodic end of the capillary. An automated liquid-flow capillary electrophoresis method to measure HbA_{1c} is commercially available and has been approved by the FDA in the United States.

Enzymatic. Enzymatic assays to measure HbA_{1c} have been developed recently based on a colorimetric method in which fructosyl peptide oxidase catalyzes the oxidative deglycation of *N*-(deoxyfructosyl)-Val-His. The procedure has been adapted for analysis on a high-throughput automated analyzer and has been approved by the FDA for use in the United States.

Assay Standardization

Clinical laboratories measure GHb with diverse assays that use multiple methods and quantify different components. The DCCT results accentuated the need for accurate GHb measurement and provided a strong impetus for standardization of GHb assays. Committees were established under the auspices of the American Association for Clinical Chemistry (AACC) in 1993 and the IFCC in 1995 to standardize GHb assays.

The NGSP was established in 1996 to implement the protocol developed by the AACC to calibrate GHb results to DCCT-equivalent values. Employing a network of reference laboratories, the NGSP interacts with manufacturers of GHb methods to help them calibrate their methods and trace values to the DCCT. Manufacturers apply for certification by performing precision testing and report results in DCCT-equivalent HbA_{1c} values. This calibration effort has markedly improved harmonization of results and has reduced imprecision. Results obtained using NGSP-certified assays can be compared directly with results of the DCCT and UKPDS, allowing alignment with clinical outcomes data. The ADA recommends that clinical laboratories in the United States use only assays certified by the NGSP and participate in proficiency testing offered by the College of American Pathologists (CAP). The CAP GHb survey uses pooled whole blood specimens at several HbA_{1c} concentrations. Target values are assigned by the NGSP network. Thus, individual laboratories can directly compare their HbA_{1c} results with those of the DCCT and UKPDS.

A different approach was adopted by the IFCC. A working group was established to devise a reference system for standardization based on HbA_{1c}. The IFCC group developed a mixture of purified HbA_{1c} and HbA_0 as primary reference material. Two candidate reference methods, namely, electrospray ionization mass spectrometry (ESI-MS) and capillary electrophoresis, were proposed. These specifically measure the glycated N-terminal valine of the β-chain of hemoglobin. Analysis is performed by digesting the hemoglobin molecule with endoproteinase Glu-C, which cleaves the β-chain between Glu-6 and Glu-7, releasing the N-terminal hexapeptide. Glycated and nonglycated hexapeptides are separated and quantified by HPLC-ESI-MS or by HPLC and capillary electrophoresis. HbA_{1c} is measured as the ratio between glycated and nonglycated N-terminal hexapeptides. The IFCC Working Group has established a network of laboratories to implement and maintain the reference system. Comparisons between IFCC and NGSP reference methods (and reference systems from Japan and Sweden) indicate a close and stable relationship and allow manufacturers to calibrate their instruments to a higher-level reference method. However, HbA_{1c} results obtained using IFCC reference methods are 1.5% to 2% absolute HbA_{1c} units lower than those of the NGSP (and lower than other reference systems). The IFCC method is not suitable for routine analysis of patient samples.

Units for Reporting of HbA_{1c}

HbA_{1c} is reported in the NGSP system as a percentage of total hemoglobin. These values, which are equivalent to those reported in the DCCT and the UKPDS, represent the most widely used reporting system in patient care and in the published literature. The IFCC method reports HbA_{1c} in System

International (SI) units, namely, mmol HbA_{1c}/mol total Hb. Comparison between the IFCC and NGSP networks produced a master equation that permits conversion between the two reference systems. For example, an HbA_{1c} result of 7% (in NGSP/DCCT/UKPDS units) is equivalent to 53 mmol/mol (in IFCC units). A calculator to convert between units is available at http://www.ngsp.org/convert1.asp. Some countries have elected to report HbA_{1c} exclusively in SI units.

In some countries, an estimated average glucose (eAG) is reported together with HbA_{1c}. A conversion table is available at http://www.ngsp.org/A1ceAG.asp. A large, prospective clinical study revealed a significant correlation between HbA_{1c} and mean plasma glucose concentrations. The concept of reporting eAG with HbA_{1c} is not accepted by all and is controversial.

Performance Goals

Some expert groups have proposed goals for HbA_{1c} assay accuracy and precision and these have tightened over the years. Within-subject biological variation (CVi) of HbA_{1c} is less than 2%. Recent ADA guidelines recommend an intralaboratory CV of less than 2% and an interlaboratory CV of less than 3.5%. For a single method, the goal should be an interlaboratory CV of less than 3.5%.

Specimen Collection and Storage

Patients need not be fasting. Venous blood should be collected in tubes containing EDTA, oxalate, or fluoride. Sample stability depends on the assay method used. Whole blood may be stored at 4°C for up to 1 week. Above 4°C, HbA_{1a+b} increases in a time- and temperature-dependent manner, but HbA_{1c} is only slightly affected. Samples are not stable at −20°C. For most methods, whole blood samples stored at −70°C or colder are stable for at least 18 months.

Reference Intervals

Values for GHbs are expressed as a proportion of total blood hemoglobin. HbA_{1c} should be reported. The reference interval (using an NGSP-certified method) is 4% to 5.6% (20 to 38 mmol/mol).

Results are not affected by acute illness. Intraindividual variability is minimal (CV ≈1%). HbA_{1c} increases slightly with age and differs among racial groups (e.g., higher in African Americans and Hispanics than Caucasians). It is not known whether this has clinical significance. In patients with poorly controlled diabetes, values rarely exceed 15%. HbA_{1c} greater than 15% or below 4% should prompt additional studies to determine the possible presence of variant hemoglobin. Note that target values derived from DCCT and UKPDS, not the reference values, are used to evaluate metabolic control in patients with diabetes.

There is no specific value of HbA_{1c} below which the risk of diabetic complications is eliminated completely. The ADA states that for most patients the goal of treatment should be to maintain HbA_{1c} at less than 7% (53 mmol/mol). (Some organizations recommend an HbA_{1c} target of less than 6.5% [48 mmol/mol].) These goals are applicable only if the assay method is certified as traceable to the DCCT reference. Each laboratory should establish its own reference interval. Assay precision is important because each 1% change in HbA_{1c} represents an approximate 30-mg/dL (1.7 mmol/L) change in average blood glucose.

No consensus has been reached on optimum frequency of testing. The ADA recommends for patients with type 1 or type 2 diabetes that *HbA_{1c} should be routinely monitored at least every 6 months in patients meeting treatment goals (and who have stable glycemic control).*

Glycated Serum Proteins

In selected patients with diabetes (e.g., GDM), assays may be needed that are more sensitive than HbA_{1c} to shorter-term alterations in average blood glucose concentrations. Nonenzymatic attachment of glucose to amino groups of proteins other than hemoglobin (e.g., serum proteins, membrane proteins, lens crystallins) to form ketoamines also occurs. Because serum proteins turn over more rapidly than erythrocytes (the circulating half-life for albumin is about 20 days), *the concentration of glycated serum albumin reflects glucose control over a period of 2 to 3 weeks.* Therefore deterioration of control and improvement with therapy are evident earlier than with HbA_{1c}.

Fructosamine

Fructosamine is the generic name for plasma protein ketoamines. The name refers to the structure of the ketoamine rearrangement product formed by the interaction of glucose with the ε-amino group on lysine residues of albumin. Analogous to HbA_{1c}, fructosamine may be used as an index of the average concentration of blood glucose over an extended time.

Because all glycated serum proteins are fructosamines and albumin is the most abundant serum protein, measurement of fructosamine is thought to be largely a measure of glycated albumin (GA), but this has been questioned by some investigators. Although the fructosamine assay can be automated and is cheaper and faster than HbA_{1c}, *there is a lack of consensus on its clinical utility*. The clinical value of fructosamine has not been firmly established, and no convincing evidence relates its concentration to the chronic complications of diabetes. An important limitation is the lack of long-term prospective studies with clinical outcomes.

Fructosamine may be useful in circumstances where HbA_{1c} is of little value, such as in patients with hemoglobin variants that are associated with decreased erythrocyte life span. Gross changes in protein concentration and half-life may have large effects on the proportion of protein that is glycated. Thus, fructosamine results may be invalid in patients with nephrotic syndrome, cirrhosis of the liver, or dysproteinemias, or after rapid changes in acute-phase reactants. *There is no role for fructosamine in the diagnosis of diabetes.*

Determination of Fructosamine

Methods for measuring glycated proteins include (1) affinity chromatography using immobilized phenylboronic acid (similar to the GHb assay); (2) HPLC of glycated lysine

residues after hydrolysis of the glycated proteins; (3) a photometric procedure in which mild acid hydrolysis releases 5-hydroxymethylfurfural—proteins are precipitated with trichloroacetic acid and the supernatant is reacted with 2-thiobarbituric acid; and (4) other procedures using phenylhydrazine and ε-*N*-(2-furoylmethyl)-L-lysine (furosine). None of these assays is popular because they are not suitable for routine clinical laboratories. Prolonged storage at ultralow temperatures (−96°C) prevents in vitro glycation of serum proteins.

An alternative method of measuring fructosamine is based on the principle that under alkaline conditions, fructosamine undergoes an Amadori rearrangement. The resultant compounds have reducing activity that can be differentiated from other reducing substances. In the presence of carbonate buffer, fructosamine rearranges to the eneaminol form, which reduces nitroblue tetrazolium (NBT) to a formazan. Absorbance at 530 nm is measured at two time points, and the absorbance change is proportional to the fructosamine concentration. A 10-minute preincubation period is necessary to allow fast-reacting interfering reducing substances to react. The assay is easily automated and has excellent between-batch analytical precision. Hemoglobin (>100 mg/dL; 1 g/L) and bilirubin (>4 mg/dL; ~68 μmol/L) may interfere; therefore moderate to grossly hemolyzed and icteric samples should not be used. Ascorbic acid concentrations greater than 5 mg/dL may cause negative interference.

Reference Intervals for Fructosamine

Values in a nondiabetic population are 205 to 285 μmol/L using a colorimetric assay. The reference interval corrected for albumin is 191 to 265 μmol/L.

Glycated Albumin

Albumin, which comprises approximately 60% of total serum protein, makes up more than 80% of total glycated serum proteins. The N-terminus and 59 lysine residues are potential glycation sites and it is not known how many of these are glycated in vivo. Analysis of human plasma by HPLC tandem mass spectrometry and [$^{13}C_6$]-glucose labeling identified 35 different glycation sites on albumin. Assays that measure only glycated albumin (GA), rather than all glycated serum proteins (i.e., fructosamine), are commercially available.

The clinical use of GA is limited by the same caveats that apply to fructosamine: limited evidence relating it to the complications of diabetes and lack of long-term prospective studies with clinical outcomes. Further studies are required to determine whether it is useful for routine monitoring of patients' glycemic control.

Determination of Glycated Albumin

Methods that have been used to quantify GA include: a colorimetric procedure in which mild acid hydrolysis releases 5-hydroxymethylfurfural—proteins are precipitated with trichloroacetic acid and the supernatant is reacted with 2-thiobarbituric acid; RIA using beads coated with antibody to albumin, and ^{125}I-labeled antibody directed against glucitollysine epitopes of GA previously reduced by sodium borohydride ($NaBH_4$) to reduce the Schiff base; enzyme-linked immunosorbent assay (ELISA) in which GA binds to a monoclonal antibody coated on a plate, followed by incubation with an enzyme-linked antihuman albumin antibody; enzyme-linked boronate immunoassay where boronic acid-HRP conjugate binds to the cis-diols of GA, which is immobilized by an antihuman albumin antibody coated onto a microtiter plate; affinity chromatography using immobilized phenylboronic acid, followed by elution and measurement of albumin; boronate affinity chromatography; HPLC with anion exchange chromatography to separate albumin, followed by boronate affinity chromatography to separate glycated from nonglycated albumin; enzymatic assay using ketoamine oxidase; and mass spectrometry.

Probably the most widely used method globally is enzymatic. The assay has two steps. In the first, glycated amino acids are eliminated by oxidation with ketoamine oxidase. In the second step, GA is hydrolyzed by an albumin-specific proteinase to glycated amino acids, which are subsequently oxidized by ketoamine oxidase to glucosone, producing hydrogen peroxide. This is quantified with a chromogen by measuring absorbance at 546/700 nm. Total albumin is measured with bromocresol purple and GA is expressed as a percentage of total albumin. This assay is commercially available in a few countries and has been used in numerous published studies. It was approved in 2017 by the FDA for use in the United States.

Reference Intervals for Glycated Albumin

Reference intervals vary considerably, depending on the method, ranging from 0.8% to 1.4% to 18% to 22%. The reference interval for the enzymatic assay, which is expressed as a percentage of total albumin, is 11.9% to 15.8%.

No significant gender differences have been observed, but values in blacks are significantly higher than in whites. Within-subject biological variation is low (CVi, 2.1%), but between-subject variation is reported to be 10.6%. In patients with poorly controlled diabetes, values may increase by up to 5-fold. Factors that influence albumin metabolism have been reported to alter GA independently of glycemia. These include the nephrotic syndrome, thyroid disease, cirrhosis of the liver, smoking, hyperuricemia, and hypertriglyceridemia. Samples can be stored as long as 23 years at −70°C.

Advanced Glycation End Products

The molecular mechanism by which hyperglycemia produces toxic effects is unknown, but glycation of tissue proteins may be important. Nonenzymatic attachment of glucose to long-lived proteins, lipids, or nucleic acids produces stable Amadori early-glycated products. These undergo a series of additional rearrangements, dehydration, and fragmentation reactions, resulting in stable **advanced glycation end products (AGEs)**. The amounts of these products do not return to normal when hyperglycemia is corrected, and they accumulate continuously over the life span of the protein. Hyperglycemia accelerates the formation of protein-bound AGE, and patients with diabetes thus have more AGE than healthy subjects. Through

effects on the functional properties of protein and the extracellular matrix, AGEs may contribute to the microvascular and macrovascular complications of diabetes.

Measurement of AGEs in the circulation has also been used as a biomarker to monitor the complications of diabetes. However, the diverse structures and composition of AGEs has resulted in assay difficulties. Analysis by ELISA has lacked standardization, yielding variable results. The development of stable isotope dilution analysis liquid chromatography-tandem mass spectrometry, in conjunction with careful preanalytic sample preparation, shows potential to resolve these problems. Some AGE products fluoresce, which forms the basis of noninvasive measurement of skin autofluorescence with a portable reader. Some studies have revealed a positive association of skin autofluorescence with complications of diabetes, but adjustment for HbA_{1c} rendered associations nonsignificant. Limitations of skin autofluorescence measurements include lack of specificity for AGE and most AGEs are not fluorescent. The clinical role of AGE is undefined.

1,5-Anhydroglucitol

Another marker of long-term glycemia is 1,5-anhydroglucitol (1,5-AG), which reflects glucose concentrations over the preceding 2 to 14 days. It is a 1-deoxy form of glucose that originates predominantly from the diet, with the vast majority (>99.9%) normally being reabsorbed from the glomerular filtrate by the SGLT4 sodium-dependent glucose transporter. When blood glucose concentrations exceed the renal threshold (usually ~180 mg/dL [~10.0 mmol/L]), reabsorption of 1,5-AG decreases, leading to a rapid reduction in serum 1,5-AG concentrations. Therefore, low 1,5-AG indicates hyperglycemia, correlating particularly with postprandial blood glucose concentration. An automated colorimetric assay is commercially available (and FDA approved for use in the United States). The reference interval is 10.2 to 33.8 μg/mL (males) and 5.9 to 31.8 μg/mL (females). Several factors unrelated to glycemia may alter 1,5-AG values, including diet, medications, renal disease, and liver disease. The evidence linking it to outcomes is limited and the clinical value of measuring 1,5-AG remains to be established.

ALBUMINURIA

Clinical Significance

Patients with diabetes mellitus are at high risk of developing renal damage. End-stage renal disease requiring dialysis or transplantation develops in approximately one-third of patients with type 1 diabetes, and diabetes is the most common cause of end-stage renal disease in the United States and Europe. Although nephropathy is less common in patients with type 2 diabetes, approximately 60% of all cases of diabetic nephropathy occur in these patients because of the higher incidence of this form of diabetes. Persistent proteinuria detectable by routine screening tests (equivalent to a urinary albumin excretion rate [AER] >200 μg/min or >300 mg/24 h) indicates overt diabetic nephropathy. This is usually associated with long-standing disease and is unusual less than 5 years after the onset of type 1 diabetes. Once diabetic nephropathy occurs, renal function deteriorates rapidly and renal insufficiency evolves. Treatment at this stage can retard the rate of progression without stopping or reversing the renal damage. Preceding this stage is a period of increased AER not detected by routine dipstick methods. This range of 20 to 200 μg/min (or 30 to 300 mg/24 h) of increased AER has been called microalbuminuria. The term microalbuminuria, although widely used, is misleading. It implies a small version of the albumin molecule rather than an excretion rate of albumin greater than normal but less than that detectable by routine methods. Current nomenclature has eliminated the terms microalbuminuria and macroalbuminuria.

The presence of increased AER denotes an increase in the transcapillary escape rate of albumin and is therefore a marker of microvascular disease. Persistent AER greater than 20 μg/min represents a 20-fold greater risk for the development of clinically overt renal disease in patients with type 1 and type 2 diabetes. Prospective studies have demonstrated that increased urinary albumin excretion precedes and is highly predictive of diabetic nephropathy, end-stage renal disease, cardiovascular mortality, and total mortality in patients with diabetes mellitus. Intensive glucose-lowering therapy can significantly reduce the risk of development of increased AER and overt nephropathy in individuals with diabetes. In addition, increased AER identifies a group of nondiabetic subjects at increased risk for coronary artery disease. Interventions, such as control of blood glucose concentrations and blood pressure, particularly with angiotensin-converting enzyme (ACE) inhibitors, slow the rate of decline in renal function.

Specimen Collection and Storage

Variations in urine flow rate in a person may be corrected by expressing albumin as a ratio to creatinine (i.e., ACR). AER is increased by physiological and other factors (e.g., exercise within 24 hours, posture, diuresis), infection, fever, marked hyperglycemia, and marked hypertension. Samples should not be collected after exertion, in the presence of urinary tract infection, during acute illness, immediately after surgery, or after an acute fluid load. All the following urine samples are currently acceptable: 24-hour collection; overnight (8 to 12 hours, timed) collection; 1- to 2-hour timed collection (in laboratory or clinic), and first morning sample for simultaneous albumin and creatinine measurement. The AER is more practical and convenient for the patient than timed specimens and is the recommended method. A first morning void sample is best because it has lower within-person variation than a random urine sample. At least three separate specimens, collected on different days, should be assayed because of high within-subject biological variation (CVi of 30% to 50%) and diurnal variation (50% to 100% higher during the day). Urine should be stored at 4°C after collection. Alternatively, 2 mL of 50 g/L sodium azide can be added per 500 mL of urine, but preservatives are not recommended for some assays. Bacterial contamination and glucose have no effect. Specimens are stable in untreated urine for 1 week at 4°C and for 5 months at −80°C. The albumin concentration decreases by 0.27%/day at −20°C.

An estimated glomerular filtration rate (eGFR) should also be calculated from serum creatinine in patients who have a positive screening test. Serum creatinine and eGFR should be performed at least annually in all adults with diabetes because some patients have decreased GFR without albuminuria.

Semiquantitative Assays

Several semiquantitative assays are available for screening for albuminuria. These test strips, most of which are optimized to read "positive" at a predetermined albumin concentration, have been recommended for screening programs. In view of the wide variability in AER, a "normal" value does not rule out renal disease. Because these assays measure albumin concentration, dilute urine may yield a false negative test result. Refrigerated urine samples should be allowed to reach at least 10°C before analysis.

Quantitative Assays

All sensitive, specific assays for urine albumin use immunochemistry with antibodies to human albumin. Four methods are available: RIA, ELISA, radial immunodiffusion, and immunoturbidimetry. Each method has advantages and disadvantages, and the choice depends on local experience and technical support. In general, these *methods have similar imprecisions, detection limits, and reference intervals.* Details of these methods are found in an expanded version of this chapter in the 6th edition of *Tietz Textbook of Clinical Chemistry and Molecular Diagnostics.*

Reference Intervals

Albuminuria[a]

	μg/min	mg/24 h	mg/g Creatinine	mg/mmol Creatinine
Normal	<20	<30	<30	<3.5
High albuminuria	20–200	30–300	30–300	3.5–30
Very high albuminuria[b]	>200	>300	>300	>30

[a]Laboratories should verify that these ranges are appropriate for use in their own settings.
[b]Also termed overt nephropathy. Previously called clinical albuminuria.

The ADA recommends initial albuminuria measurement in patients with type 1 diabetes who have had diabetes for 5 years or longer, and in all type 2 diabetic patients. Because of the difficulty involved in dating the onset of type 2 diabetes, screening should commence at diagnosis. Analysis should be performed annually in all patients who have a negative screening result. Screening may be performed with a semiquantitative assay. If the screening result is positive, albuminuria should be evaluated by a quantitative assay. Diagnosis requires the demonstration of albuminuria in at least two of three samples measured within a 3- to 6-month period. If the confirmatory test result is positive, treatment should be initiated.

POINTS TO REMEMBER

HbA_{1c}

- Glycated hemoglobin is formed by nonenzymatic attachment of glucose to hemoglobin.
- HbA_{1c} has glucose attached to the N-terminal Val of the beta chain of hemoglobin.
- The concentration of HbA_{1c} depends on the concentration of glucose in the blood and the erythrocyte lifespan.
- The average erythrocyte lifespan is 120 days, and HbA_{1c} therefore reflects the average blood glucose concentration over the preceding 8–12 weeks.
- Any condition that substantially changes erythrocyte lifespan will alter HbA_{1c}.
- HbA_{1c} is used to diagnose diabetes, monitor glycemic control, evaluate the need to change therapy, and predict the development of microvascular complications.

Glucose

- Hyperglycemia results from defects in insulin secretion and/or insulin action.
- Blood glucose homeostasis is regulated by several hormones, including insulin, glucagon, epinephrine, and cortisol.
- Blood glucose concentrations fluctuate widely during the day, depending on food ingestion, exercise, and other factors (e.g., stress).
- Self-monitoring of blood glucose using portable meters in patients with diabetes who require insulin has been shown to improve patient outcomes.
- Self-monitoring of blood glucose in noninsulin treated patients is not yet proven to be effective.
- The use of continuous glucose monitoring systems (CGMS), which use subcutaneously implanted glucose sensors, by patients on insulin is increasing.

Type 2 diabetes

- Type 2 diabetes is the most common form of diabetes, accounting for ~90% of all cases.
- The onset is insidious, and patients have minimal symptoms.
- Many patients have irreversible complications at the time of diagnosis.
- Patients exhibit both insulin resistance and inadequate insulin secretion.
- Insulin resistance is very difficult to measure, and at diagnosis, the patient may have normal, increased, or decreased insulin concentrations.
- The molecular defects are a consequence of both genetic and environmental factors.
- The gene (or genes) that causes the common forms of type 2 diabetes have not been identified.
- Obesity is linked to the development of type 2 diabetes and lifestyle changes (weight loss and exercise) can delay the onset of the disease.

AT A GLANCE

Diagnosis of Diabetes

Diabetes is diagnosed if at least one of the following criteria is met:

- Fasting plasma glucose (FPG) ≥126 mg/dL (7.0 mmol/L)
- 2-hour plasma glucose ≥200 mg/dL (11.1 mmol/L) during an oral glucose tolerance test
- HbA_{1c} ≥6.5% (48 mmol/mol)

The same test should be repeated on a different day to confirm the diagnosis.

Glucose should be measured in venous plasma.

Glycolysis should be minimized by placing the tube immediately after collection in an ice-water slurry and separating plasma from cells within 30 min. If that cannot be achieved, blood should be collected in a tube containing a rapidly effective glycolysis inhibitor, for example, citrate buffer.

Plasma glucose and HbA_{1c} should be measured in an accredited laboratory; point-of-care devices are not suitable for screening or diagnosis.

In the United States, HbA_{1c} analysis should be performed using a method that is National Glycohemoglobin Standardization Program (NGSP)–certified and standardized to the Diabetes Control and Complications Trial (DCCT) assay.

REVIEW QUESTIONS

1. Hemoglobin A_{1c} (HbA_{1c}) indicates compliance of a patient with diabetes with his or her insulin-taking regimen by monitoring glucose control. HbA_{1c} concentration represents the integrated glucose value in the blood over what period?
   a. 8 to 12 days
   b. 8 to 12 weeks
   c. 8 to 12 months
   d. 1 day
2. Which of the following values obtained during an oral glucose tolerance test (OGTT) is diagnostic of diabetes mellitus?
   a. 2-Hour specimen = 125 mg/dL (6.9 mmol/L)
   b. Fasting glucose = 138 mg/dL (7.7 mmol/L)
   c. Fasting glucose = 110 mg/dL (6.1 mmol/L)
   d. 2-Hour specimen = 80 mg/dL (4.4 mmol/L)
3. Type 2 diabetes:
   a. is associated with resistance to the action of insulin.
   b. is caused by destruction of pancreatic β-cells.
   c. is also known as insulin-dependent diabetes mellitus.
   d. occurs less frequently than type 1 diabetes.
4. All of the following results are confirmatory and diagnostic laboratory values for diabetes mellitus *except*:
   a. nonfasting blood glucose >200 mg/dL. (11.1 mmol/L)
   b. 2-hour oral glucose tolerance values >200 mg/dL. (11.1 mmol/L)
   c. urine glucose >250 mg/dL. (13.9 mmol/L)
   d. fasting plasma glucose >126 mg/dL. (7.0 mmol/L)
5. Release of glucose from its storage form is referred to as:
   a. glycogenesis.
   b. glycogenolysis.
   c. glycolysis.
   d. glyconeogenesis.
6. Which of the following hormones produces *hyperglycemia*?
   a. Epinephrine
   b. Glucagon
   c. Thyroid hormone
   d. All of the above hormones produce hyperglycemia
7. Whole blood glucose values are approximately what percent different from plasma glucose values?
   a. 20% higher
   b. 11% lower
   c. 11% higher
   d. There is no difference between whole blood glucose and plasma glucose values
8. The development of ketosis in uncontrolled diabetes is a result of:
   a. increased lipolysis of fatty acids from adipose stores and decreased reesterification of these fatty acids to triglycerides.
   b. increased nonenzymatic addition of glucose to proteins, lipids, and nucleic acids that form ketoamines.
   c. increased formation of advanced glycation end products that do not return to normal levels when diabetes is controlled.
   d. formation of circulating antibodies that are formed against the excess adipose tissue present in a person with diabetes.
9. Which of the following hormones promotes *decreased* blood glucose?
   a. Epinephrine
   b. Glucagon
   c. Cortisol
   d. Insulin
10. The purpose of examining urinary albumin excretion in an individual with type 1 or type 2 diabetes is to:
   a. assess the ability of the pancreas to synthesize sufficient insulin.
   b. determine the rate of formation of advanced glycation end products.
   c. assess the possibility of overt diabetic nephropathy.
   d. examine the health of the liver in its ability to synthesize albumin.

SUGGESTED READINGS

American Diabetes Association. (1997). Report of the expert committee on the diagnosis and classification of diabetes mellitus. *Diabetes Care*, *20*, 1183–1197.

American Diabetes Association. (2017). Classification and diagnosis of diabetes. *Diabetes Care*, *40*(Suppl. 1), S11–S24.

Atkinson, M. A., Eisenbarth, G. S., & Michels, A. W. (2014). Type 1 diabetes. *Lancet*, *383*, 69–82.

Bry, L., Chen, P. C., & Sacks, D. B. (2001). Effects of hemoglobin variants and chemically modified derivatives on assays for glycohemoglobin. *Clinical Chemistry*, *47*, 153–163.

Brownlee, M. (2001). Biochemistry and molecular cell biology of diabetic complications. *Nature*, *414*, 813–820.

Czech, M. P. (2017). Insulin action and resistance in obesity and type 2 diabetes. *Nature Medicine*, *23*, 804–814.

DCCT. (1993). The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *New England Journal of Medicine*, *329*, 977–986.

Gerich, J. E. (2000). Physiology of glucose homeostasis. *Diabetes Obesity Metabolism*, *2*, 345–350.

Kahn, S. E., Cooper, M. E., & Del Prato, S. (2014). Pathophysiology and treatment of type 2 diabetes: Perspectives on the past, present, and future. *Lancet*, *383*, 1068–1083.

Little, R. R., Rohlfing, C. L., & Sacks, D. B. (2011). Status of hemoglobin A_{1c} measurement and goals for improvement: from chaos to order for improving diabetes care. *Clinical Chemistry*, *57*, 205–214.

Levey, A. S., Becker, C., & Inker, L. A. (2015). Glomerular filtration rate and albuminuria for detection and staging of acute and chronic kidney disease in adults: a systematic review. *JAMA: The Journal of the American Medical Association*, *313*, 837–846.

Mueckler, M., & Thorens, B. (2013). The slc2 (GLUT) family of membrane transporters. *Molecular Aspects Medicine*, *34*, 121–138.

Metzger, B. ,E., Lowe, L. P., Dyer, A. R., Trimble, E. R., Chaovarindr, U., Coustan, D. R., et al. (2008). Hyperglycemia and adverse pregnancy outcomes. *New England Journal of Medicine*, *358*, 1991–2002.

Prasad, R. B., & Groop, L. (2015). Genetics of type 2 diabetes-pitfalls and possibilities. *Genes (Basel)*, *6*, 87–123.

Reaven, G. M. (2005). The metabolic syndrome: requiescat in pace. *Clinical Chemistry*, *51*, 931–938.

Sandoval, D. A., & D'Alessio, D. A. (2015). Physiology of proglucagon peptides: role of glucagon and GLP-1 in health and disease. *Physiological Review*, *95*, 513–548.

Sacks, D. B. (2011). A1c versus glucose testing: a comparison. *Diabetes Care*, *34*, 518–523.

Sacks, D. B., Arnold, M., Bakris, G. L., Bruns, D. E., Horvath, A. R., Kirkman, M. S., et al. (2011). Guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes mellitus. *Clinical Chemistry*, *57*, e1–e47.

U.K. Prospective Diabetes Study (UKPDS) Group. (1998). Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). UK prospective diabetes study (UKPDS) group. *Lancet*, *352*, 837–853.

van den Berghe, G., Wouters, P., Weekers, F., Verwaest, C., Bruyninckx, F., Schetz, M., et al. (2001). Intensive insulin therapy in the critically ill patients. *New England Journal of Medicine*, *345*, 1359–1367.

34

Cardiovascular Disease

Fred S. Apple, Jens Peter Goetze, Allan S. Jaffe

OBJECTIVES

1. Define the following terms:
   - Acute coronary syndrome (ACS)
   - Acute myocardial infarction (AMI)
   - Angina
   - Atherosclerosis
   - Cardiac biomarker
   - Congestive heart failure (CHF)
   - Coronary artery disease (CAD)
   - High-sensitivity cardiac troponin assay
   - Electrocardiogram (ECG)
   - Ischemia
   - Myocardium
   - Myocardial infarction
   - Myocardial injury
   - Natriuretic peptide (NP)
   - Plaque
   - Reperfusion
   - Troponin (cTn)
2. Describe the anatomy of the heart, including layers, chambers, and protein makeup of muscle.
3. List the events in the process of atherosclerotic plaque formation.
4. Describe an ideal cardiac biomarker, including necessary characteristics, analytical considerations, and persistence in blood following an AMI.
5. Compare cardiac troponin I and T, including structural differences, physiological function, localization, and usefulness in diagnosing an AMI.
6. Define the two analytical criteria required to designate a cTn assay as highly sensitive.
7. For the following cardiac biomarkers, list and describe their location within heart tissue, physiological function if known, clinical utility for diagnosing disease, tissue specificity, other pathophysiology conditions that cause increased values, and the specimen requirements needed for laboratory analysis:
   - B-type natriuretic peptide (BNP)
   - cTn I and T
   - N-terminal portion of proBNP (NT-proBNP)
8. List four preanalytical considerations that must be assessed by a clinical laboratory that uses NP assays and state what units must be used to report these analytes.
9. List four preanalytical considerations that must be assessed by a clinical laboratory that uses NP assays; state what units must be used to report these analytes.
10. Evaluate and analyze case studies related to cardiovascular disease and the use of cardiac biomarkers in the diagnosis of cardiovascular disease.

KEY WORDS AND DEFINITIONS

99th percentile upper reference limit (URL) A statistically derived cTn concentration determined from a minimum number of a normal, apparently healthy reference population.

Acute coronary syndrome (ACS) A sudden cardiac disorder that varies from angina (chest pain on exertion with reversible tissue injury), to unstable angina (with minor myocardial injury), and to myocardial infarction (with extensive tissue necrosis, which is irreversible).

Acute myocardial infarction (AMI) An acute obstruction of circulation of the heart muscle occurring during the period when circulation to a region of the heart is obstructed and necrosis is occurring.

Angina A condition marked by severe pain in the chest, often also spreading to the shoulders, arms, and neck, caused by an inadequate blood supply to the heart.

Arrhythmia Any variation from the normal rhythm of the heartbeat; alternative (and broader) term: dysrhythmia, especially to indicate an abnormally slow or fast heartbeat that may have rhythmic beating.

Atherosclerosis Any of a group of diseases characterized by thickening and loss of elasticity of arterial walls.

Atherosclerotic plaque A pearly white area within the wall of an artery that causes the intimal (interior) surface to bulge into the lumen; composed of lipid, cell debris, smooth muscle cells, collagen, and sometimes calcium; also known as an atheroma; vulnerable to rupture that causes the formation of a platelet- and fibrin-rich thrombus leading to myocardial infarction and ischemic stroke.

Cardiac biomarker A biological compound whose measurement is useful in the diagnosis/detection of

cardiac disease; used to (1) detect cardiac disorders, (2) detect risk of developing cardiac disorders, (3) monitor the disorder, or (4) predict the response of a disorder to a treatment.

Congestive heart failure (CHF) A clinical syndrome due to heart disease, characterized by breathlessness and abnormal sodium and water retention, often resulting in edema; also called heart failure.

Coronary arteries The two main arteries that provide blood to the heart, surrounding the heart like a crown, coming out of the aorta, arching down over the top of the heart, and dividing into two branches.

Electrocardiogram (ECG) A graphic recording of the electrical activity produced by the heart.

Myocardial injury Any mechanism of injury to the heart that results in an increase in cTnI or cTnT blood concentrations above the 99th percentile upper reference limit.

Myocardial ischemia Deficiency of blood supply to the heart muscle due to obstruction or constriction of the coronary arteries.

Myocardium The middle and thickest layer of the heart wall, composed of cardiac muscle.

Necrosis The sum of the morphological changes indicative of cell death and caused by the progressive degradative action of enzymes.

Non–ST segment elevation myocardial infarction (NSTEMI) A myocardial infarction in which the ST segment is not elevated in one lead or several leads of the ECG.

Percutaneous coronary intervention (PCI) The management of coronary artery occlusion by any of various catheter-based techniques.

Risk stratification A statistical process used to determine detectable characteristics of a biomarker associated with an increased chance of experiencing adverse outcomes.

ST segment elevation myocardial infarction (STEMI) Any type of myocardial infarction in which the ST segment is elevated in one lead or several leads of the ECG.

Thrombolysis Lysis of a thrombus or thrombi.

Unstable angina Angina that occurs unpredictably or suddenly increases in severity or frequency.

Ventricles (right and left) The two lower chambers of the heart, responsible, respectively, for pumping blood into the lungs via the pulmonary artery and into the systemic circulation via the aorta.

Acute ischemic disease and heart failure are the two most common cardiovascular diseases that rely on a biochemical diagnosis. The cardiac biomarkers cardiac troponin (cTn) and natriuretic peptides (NP), respectively, used in their diagnosis and management, are the major focus of this chapter.

The most serious form of ischemic heart disease is **acute myocardial infarction (AMI)**. AMI occurs when there is an imbalance between supply and demand for oxygen in the **myocardium** (heart muscle) resulting in injury to and the eventual death of muscle cells (myocytes). When the blood supply to the muscle in a region of the heart is blocked for more than a few minutes, many or most of the muscle cells in the affected region die. This is called **necrosis** (cell death) of the myocardium. Other events of lesser severity may be missed entirely or be called **angina**, which ranges from stable to **unstable angina**. The ischemic events in the heart, ranging from angina (no cell death) to AMI (cell death), are known as **acute coronary syndromes (ACS)**.

In the United States, approximately 92.1 million (1 in 3) individuals have some form of heart disease. Of those, 7.4 million have or have had a myocardial infarction (MI) and approximately 8.7 million have angina pectoris. Cardiovascular disease accounts for more deaths (31%) than any other disease. It also leads to more hospitalizations that any other disease, with between 5 and 6 million admissions yearly. Approximately 695,000 patients will have a coronary artery event, 580,000 will have an AMI annually, and another 325,000 will have a recurrent event. Blacks have a greater number of events than whites. It is estimated that another 165,000 MIs occur silently. Coronary heart disease causes more than 1 in 7 deaths in the United States. Deaths that occur acutely result from ventricular **arrhythmias** or pump dysfunction and **congestive heart failure (CHF)** with or without cardiogenic shock. Death rates are sharply age dependent, both during hospitalization and in the year after infarction. The US yearly economic burden of cardiovascular disease is in excess of $316 billion for coronary heart disease and stroke and by 2030 is estimated to be $918 billion.

In acute ischemic heart disease, the clinical laboratory plays an important role in detection of **myocardial injury**. The measurement of cTn is an important test for this purpose. These proteins are found almost exclusively in heart muscle cells (cTnT has been shown to be expressed in diseased skeletal muscle from some patients with neuromuscular diseases) and released into the circulation when cells die. Increased concentrations of cTn in the blood are sensitive signs of damage to heart muscle. Conversely, reference concentrations (≤**99th percentile upper reference limit [URL]**) provide powerful evidence for the healthcare provider that a patient's symptoms are not related to cardiac injury.

Heart failure is often termed CHF. In the United States, CHF is the only cardiovascular disease with an increasing incidence. The National Heart, Lung, and Blood Institute estimates current prevalence at 4.8 million Americans and 23 million worldwide with CHF. There are approximately 580,000 new cases each year, with approximately 1 million admissions to hospitals for CHF per year. CHF is the leading cause of hospitalization in individuals 65 years of age and older.

Clinical chemistry testing has become important in detection of CHF. The key blood tests are B-type natriuretic peptide (BNP) and the N-terminal portion of proBNP

(designated NT-proBNP), which are breakdown products of proBNP. BNP and NT-proBNP are released from the stressed heart following upregulation of proBNP. Unlike cTn, which are intracellular proteins that escape from heart muscle cells only because the cells are dead or seriously injured (irreversible injury), BNP is a hormone that is secreted into the blood, stimulated by the stretch of the heart wall that occurs in heart failure. (Note that the general use of the term NP in this chapter refers to either BNP or NT-proBNP unless specifically indicated.)

ANATOMY AND PHYSIOLOGY OF THE HEART

The average human adult heart weighs approximately 325 g in men and 275 g in women. It is enclosed in a sac called the pericardium. The cardiac wall is composed of three layers: the epicardium (the outer most layer), a middle layer, and an inner layer called the endocardium. The heart has four chambers. The two upper chambers are termed the right and left atria, and the two lower chambers are termed the **right and left ventricles** (Fig. 34.1). The endocardium is the layer most

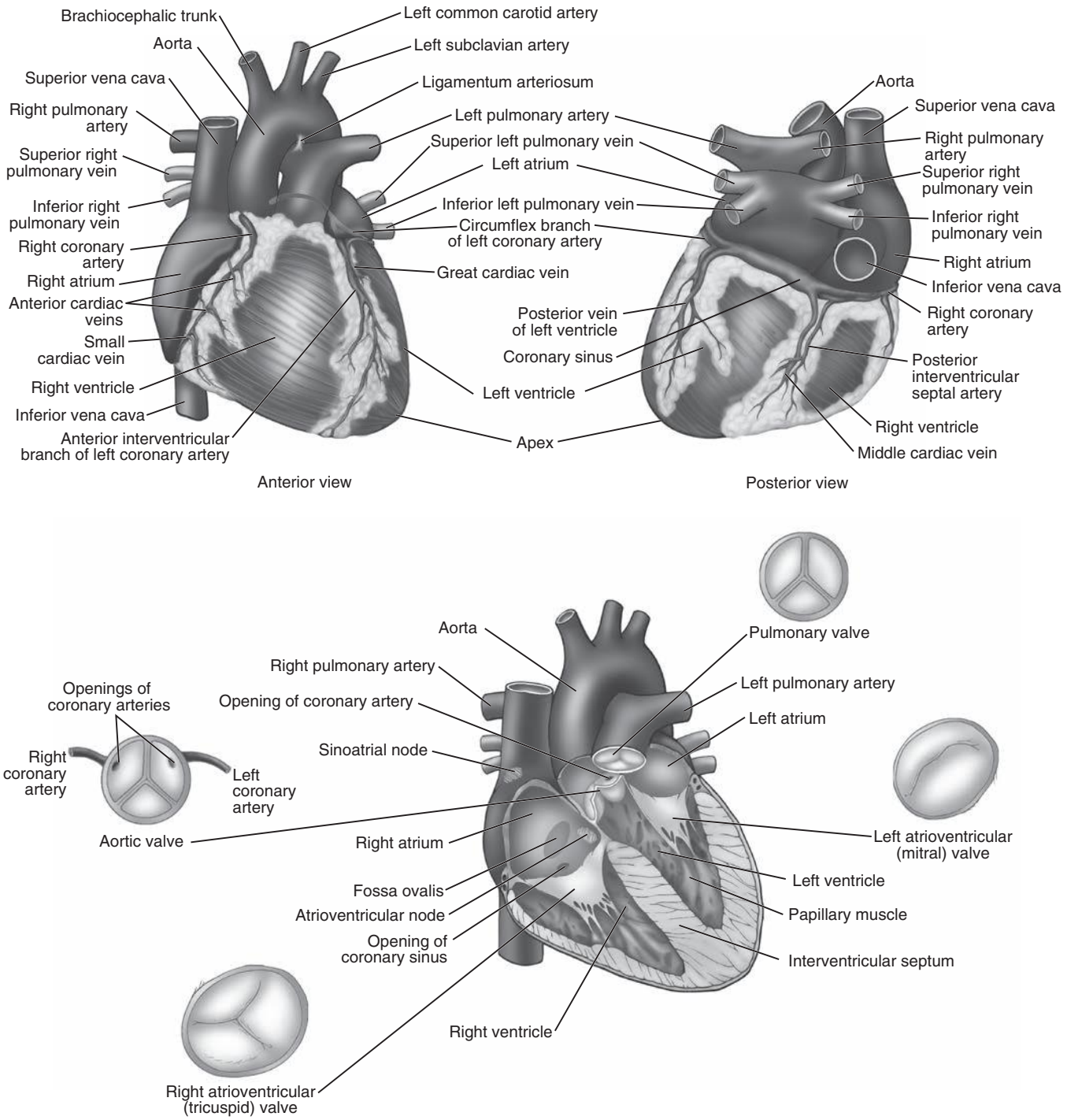


Fig. 34.1 Anatomy of the heart. (From Dorland, W. A. [2003]. *Dorland's Illustrated Medical Dictionary* [30th ed.]. Philadelphia: WB Saunders: Panel 20.)

susceptible to **myocardial ischemia**, a condition where the heart tissue is slowly or suddenly starved of oxygen and other nutrients. Note that the **coronary arteries**, which supply the blood to the wall of the heart, are on the epicardium. The work of the heart is generated by the alternating contraction and relaxation of striated muscle fibers. The fibers contain the contractile proteins actin, myosin, and troponins that regulate contraction. Two of the troponins, the cardiac forms of troponins I and T, have become the definitive biomarkers of cardiac injury.

A typical cardiac cycle consists of two intervals known as systole and diastole (Fig. 34.2). During systole, the blood

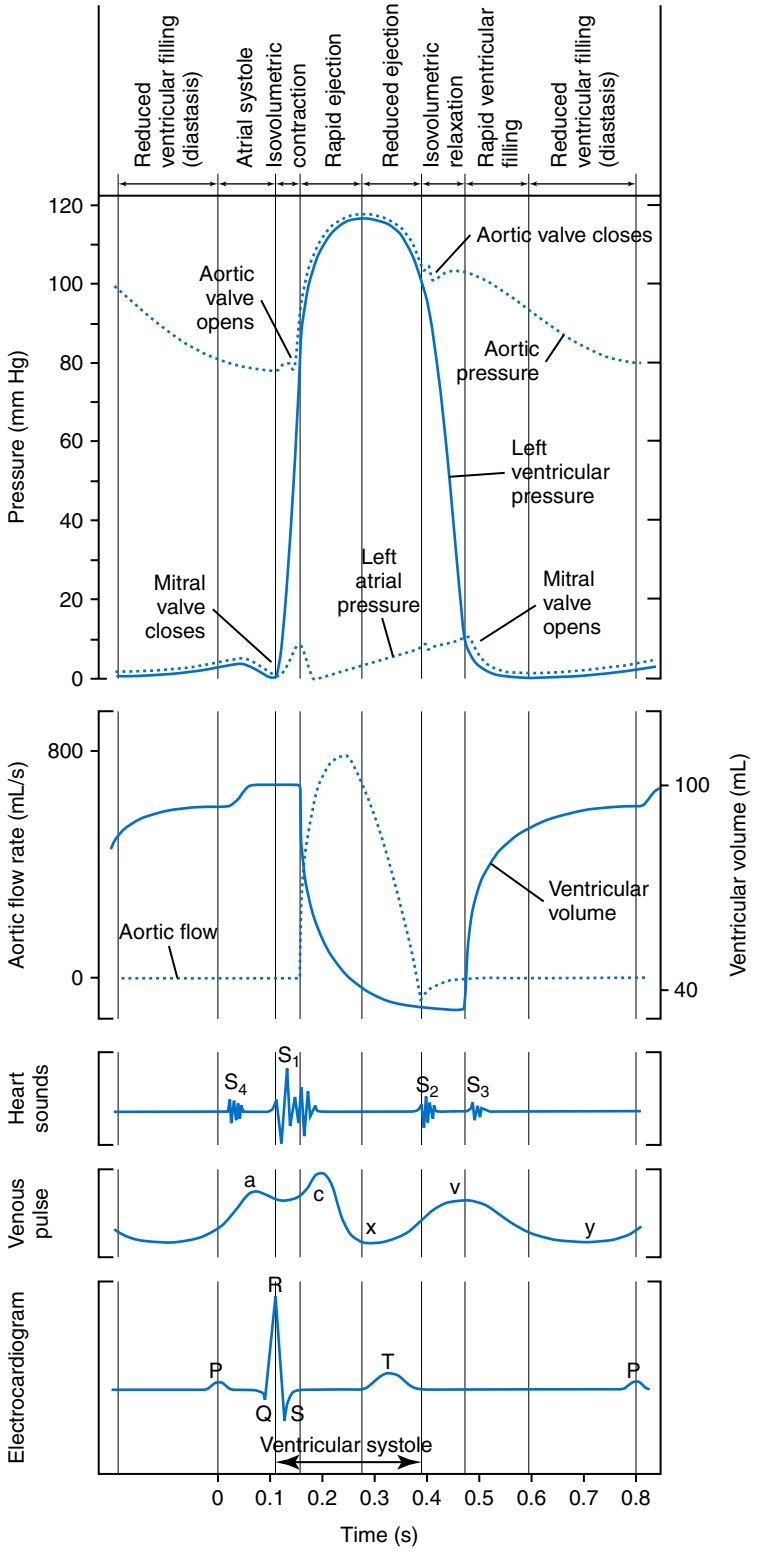


Fig. 34.2 The cardiac cycle. (From Dorland, W. A. [2003]. *Dorland's Illustrated Medical Dictionary* [30th ed.]. Philadelphia: WB Saunders, with permission from the National Kidney Foundation.)

pressure (BP) in the aorta is typically about 120 mm Hg. During diastole, the BP falls to about 70 mm Hg. At rest, the heart pumps between 60 and 80 times per minute. The cardiac cycle is tightly controlled by the cardiac conducting system, which initiates electrical impulses and carries them, via a specialized conducting system, to the myocardium. The **electrocardiogram (ECG)** records changes in electrical potential and is a graphic tracing of the variations in electrical potential caused by the excitation of the heart muscle. Clinically, the ECG is used to identify (1) anatomic, (2) metabolic, (3) ionic, and (4) hemodynamic changes. The clinical sensitivity and specificity of ECG abnormalities for detecting ACSs are influenced by a wide spectrum of physiological and anatomical changes and the clinical situation.

Under normal circumstances, each cardiac cycle's electrical potential changes, with each being similar to that of every other cycle, and includes three major components (Fig. 34.3): the atrial depolarization (the P wave), ventricular depolarization (the QRS complex), and repolarization (the ST segment and T wave). A routine ECG is composed of 12 leads. Each lead records the same electrical impulse but in a different position relative to the heart. Areas of pathology shown on the ECG are localized by analyzing differences between the tracing in question and what is considered normal in the 12 different leads.

Cardiac Disease

Acute Coronary Syndrome

ACS includes patients who have a variety of forms of unstable ischemic heart disease. In the severe form of AMI, the ECG typically shows elevation of a portion called the ST segment. The associated clinical picture is known as **ST segment elevation myocardial infarction (STEMI)**. Partial loss of coronary perfusion, if severe, also leads to necrosis, but the magnitude of cell death is generally less and the ECG does not show elevation of the ST segment. The condition is known as **non–ST segment elevation myocardial infarction (NSTEMI)**, or non-STEMI. Patients with STEMI will usually develop Q waves on their ECGs (see Fig. 34.3), hence the term Q-wave MI. If they do not have STE but have biochemical evidence of cardiac injury (e.g., an increased cTnI or cTnT in blood), they are called NSTEMI. Those who have unstable ischemia and do not show evidence of cardiac necrosis (cell death indicated by increased blood concentrations of a cTn) are classified as having unstable angina. Most of these syndromes occur in response to an acute event in the coronary artery that obstructs circulation to a region of the heart. If the obstruction blocks much of the passageway for blood in the vessel and persists, then necrosis usually results. Because necrosis is known to take some time to develop, it is apparent that opening the blocked coronary artery in a timely fashion will often prevent some of the death of myocardial tissue.

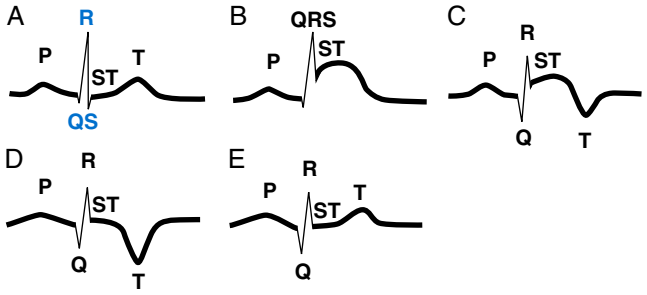


Fig. 34.3 Electrocardiograms from a patient with an acute myocardial infarction. The sequence is (A) normal; (B) hours after infarction, the ST segment becomes elevated; (C) hours to days later, the T wave inverts and the Q wave becomes larger; (D) days to weeks later, the ST segment returns to near normal; and (E) weeks to months later, the T wave becomes upright again, but the large Q wave may remain.

The major cause of ACS is **atherosclerosis**. This is a disease caused by plaque (a deposit of fatty material) being formed on the inner lining of the coronary arteries that feed the surface of the heart, which contributes to significant narrowing of the artery's lumen. Such plaques are vulnerable to rupture that results in the formation of a platelet- and fibrin-rich thrombus leading to MI and ischemic stroke. **Atherosclerotic plaque** has a tendency to break open (plaque disruption) and form blood clots (thrombi) within the vessel, further blocking or completely stopping blood flow. Myocardial ischemia and subsequent infarction usually begin in the endocardium and spread toward the epicardium. Irreversible cardiac injury consistently occurs when the occlusion is complete for at least 15 to 20 minutes. Most of the damage occurs within the first 2 to 3 hours.

Restoration of coronary blood flow within the first 60 to 90 minutes results in the maximal salvage of tissue, but treatment even up to 4 to 6 hours is associated with increased survival. For patients with STEMI, earlier opening of the vessel can be achieved with clot-dissolving agents (**thrombolysis**) and/or **percutaneous coronary intervention (PCI)**, which is a nonsurgical method used to open narrowed arteries that supply the heart muscle with blood. Administration of clot-dissolving medications without the use of PCI still plays a major role in treatment. In addition, it is now apparent that urgent invasive revascularization also benefits those with NSTEMI.

Precipitating Factors

A precipitating factor is defined as an element that causes or contributes to the occurrence of a disorder. In many patients with AMI, the following precipitating factors have been observed for patient activities at the onset of AMI: (1) heavy physical exertion, 13%; (2) modest or usual exertion, 18%; (3) surgical procedure, 6%; (4) rest, 51%; and (5) sleep, 8%.

Role of Clinical History in Diagnosis of Acute Coronary Syndrome

The clinical history remains of substantial value in establishing a diagnosis. A prodromal history of angina is elicited in 40% to 50% of patients with AMI. Approximately one-third have had symptoms from 1 to 4 weeks before hospitalization. In the remaining two-thirds, symptoms predate admission by a week or less, with one-third of these patients having had symptoms for 24 hours or less. In most patients, the pain of an ACS is severe but rarely intolerable. Such pain is described as (1) constricting, (2) crushing, (3) oppressing, or (4) compressing. In addition, the patient often complains of something sitting

on or squeezing the chest. The pain is usually felt behind the sternum (retrosternal), spreading frequently to both sides of the chest, favoring the left side. Often the pain radiates down the left arm. In some instances, the pain of AMI may begin in the upper abdomen (epigastrium) and simulate a variety of abdominal disorders, which may lead to a misdiagnosis of indigestion. In other patients, the discomfort of AMI radiates to the (1) shoulders, (2) upper extremities, (3) neck, and (4) jaw, again usually favoring the left side. In patients with preexisting angina, the pain of infarction usually resembles that of angina with respect to features and location, but it is generally (1) much more severe, (2) lasts longer (more than 30 minutes), and/or (3) is not relieved by rest and nitroglycerin. Older individuals, diabetics, and women are more likely to present atypically, without pain or with nonspecific symptoms. Sometimes, the pain of AMI may have disappeared by the time a physician first encounters the patient, or it may persist for a few hours.

Role of Cardiac Biomarkers in Acute Coronary Syndrome

A **cardiac biomarker** measurement is useful in detecting cardiac injury (irreversible structural damage), most commonly for detecting myocardial injury, including AMI. In the latter setting, they are most useful when patients have nondiagnostic ECGs. For a biomarker to be clinically useful in detecting AMI, it must be released rapidly from the heart into the circulation and provide clinically sensitive and specific diagnostic (rule-out and rule-in) information. The analytical assays must be rapid and able to measure low concentrations of the biomarker in serum, plasma, or whole blood samples. Also, the ideal biomarker (cTn) of myocardial injury would persist in the circulation for several days to provide a late diagnostic time window for patients who arrive late after the event (for example, those with minimal pain). Understanding the kinetic rise and fall of a biomarker after an acute or chronic injury is important for clinical interpretation.

Diagnosis of Acute Myocardial Infarction

Historically, the use of cardiac biomarkers in the diagnosis of AMI began in the 1950s; with the measurement of serum (1) aspartate aminotransferase (AST), (2) lactate dehydrogenase (LD), (3) total creatine kinase (CK), and (4) α-hydroxybutyrate. In 1986 the World Health Organization (WHO) produced criteria for the diagnosis of AMI. These criteria required that at least two of the following be met for diagnosing AMI: (1) a history of chest pain, (2) evolutionary changes on the ECG, and/or (3) elevations of serial cardiac biomarkers to a level two times the normal value. Superseding these criteria are the 2000 European Society of Cardiology/American College of Cardiology (ESC/ACC) consensus conference updated in 2007 and 2012 (Global Task Force; pending the Fourth Universal Definition update in 2018) codified the role of cardiac biomarkers by advocating that diagnosis should be based on biomarkers of cardiac damage in the appropriate clinical situation (Box 34.1). The recommendations of the Biochemistry Panel of the ESC/ACC Committee for the use of biomarkers in the diagnosis of an established AMI are listed in Box 34.2.

Congestive Heart Failure

CHF is a syndrome characterized by ineffective pumping of the heart, often leading to an accumulation of fluid in the lungs. At least half of such cases are a result of the loss of cardiac tissue function and are called heart failure with reduced ejection fraction (HFREF). The other half are due to increased stiffness of the cardiac muscle. This type of HF is referred to as heart failure with preserved ejection fraction, or HFPEF.

BOX 34.1 Criteria for the Definition of Acute Myocardial Infarction

1. Detection of rise and/or fall of cardiac biomarkers (preferably troponin) above the 99th percentile of the upper reference limit, together with evidence of ischemia with at least one of the following:
   a. Ischemic symptoms
   b. Electrocardiogram changes of new ischemia (new ST-T changes or new left bundle branch block)
   c. Development of pathologic Q waves on the electrocardiogram
   d. Imaging evidence of new loss of viable myocardium or new regional wall motion abnormality
   e. Identification of an intracoronary thrombus by angiography or autopsy

From Thygesen, K., Alpert, J. S., Jaffe, A. S., et al., the Writing Group on behalf of the Joint ESC/ACCF/AHA/WHF Task Force for the Universal Definition of Myocardial Infarction. (2012). Third universal definition of myocardial infarction. ESC/ACCF/AHA/WHF expert consensus document. *Circulation*, *126*, 2020–2035.

BOX 34.2 European Society of Cardiology/American College of Cardiology Recommendations for Use of Cardiac Biomarkers for Detection of Myocardial Injury and Myocardial Infarction

- Increases in biomarkers of cardiac injury are indicative of injury to the myocardium, but not an ischemic mechanism of injury.
- cTn (I or T) is a preferred biomarker for diagnosis of myocardial injury.
- Increases in cardiac biomarker proteins reflect irreversible injury.
- Improved quality control of cTn assays is essential, especially at the 99th percentile for the hs-cTn assays.
- Myocardial infarction is present when there is cardiac damage, as detected by an increase of cTn above the 99th percentile URL in a clinical setting consistent with myocardial ischemia.
- For patients with an ischemic mechanism of injury, prognosis is related to the extent of cTn increases.
- If an ischemic mechanism is unlikely, other causes of cardiac injury should be pursued.
- Samples must be obtained at least at 0–3 h for hs-cTn assays and at 6–9 h for contemporary assays after the symptoms begin.

cTn, Cardiac troponin; *hs*, high sensitivity; *URL*, upper reference limit.

Other forms include those related to valvular heart disease and so-called high-output HF. The condition is one in which there is an abnormality of cardiac function such that the heart cannot pump sufficient blood to satisfy the requirements of metabolizing tissues, which are abnormally high. With too little blood being delivered, the organs and other tissues do not receive enough oxygen and nutrients to function properly.

To functionally stage CHF patients, the New York Heart Association (NYHA) classification system is often used (Table 34.1). In this system, Class I patients are generally considered asymptomatic with no restrictions on physical activity. In the highest class (IV), patients are often symptomatic at rest with severe limitations on physical activity. The clinical manifestations of heart failure vary considerably and many are nonspecific. The findings depend on many factors, including (1) the clinical characteristics of the patient, (2) the extent and rate at which the heart's performance becomes abnormal, (3) the cause of the heart disease, (4) concomitant comorbidities, and (5) the part of the heart that is affected by abnormal functioning. The severity of impairment ranges from mild, manifested clinically only during stress, to advanced, in which cardiac pump function is unable to sustain life without external support.

Because the symptoms and signs of heart failure are nonspecific, an objective biomarker test for heart failure (such as natriuretic peptide [NP] do increase progressively with increasing severity of disease, but should not increase or decrease in conditions that mimic CHF.

CARDIAC BIOMARKERS

Numerous biomarkers have been monitored to assess myocardial injury and dysfunction (Fig. 34.4, Tables 34.2 and 34.3). Most are myocardial proteins and differ in their (1) location within the myocyte, (2) release kinetics after damage, and (3) clearance from the circulation. In this section, cTns (used as biomarkers of myocardial injury and in diagnosis of AMI) and NPs (used in CHF) are described.

TABLE 34.1 New York Heart Association Functional Classification Used to Classify the Extent of Heart Failure

NYHA Class	Symptoms
I	Cardiac disease, but no symptoms and no limitation in ordinary physical activity (e.g., shortness of breath when walking, climbing stairs, etc.)
II	Mild symptoms (mild shortness of breath and/or angina) and slight limitation during ordinary activity
III	Marked limitation in activity due to symptoms, even during less-than-ordinary activity (e.g., walking short distances [20–100 m]). Comfortable only at rest
IV	Severe limitations; experiences symptoms even while *at rest* (mostly bedbound patients)

NYHA, New York Heart Association.

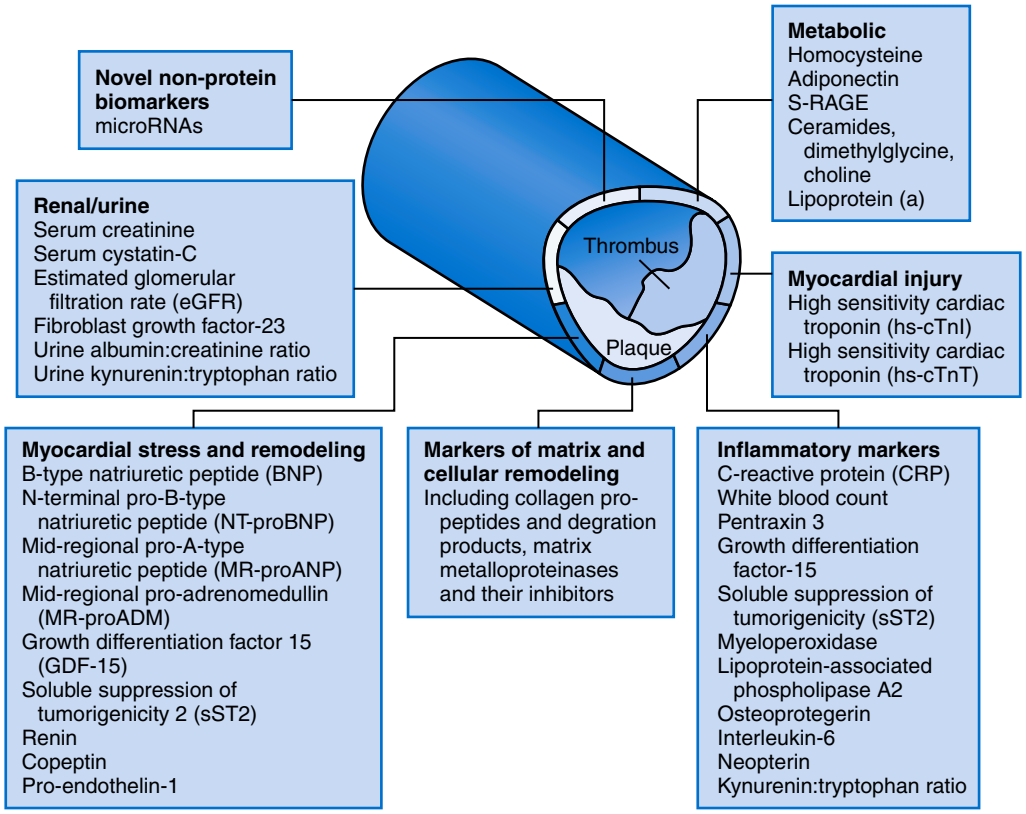


Fig. 34.4 Selected cardiovascular biomarkers associated with prognosis in stable coronary artery disease. (From Omland, T., & White, H. D. [2017]. State of the art: blood biomarkers for risk stratification in patients with stable ischemic heart disease. *Clin Chem, 63*, 165–176.)

TABLE 34.2 Biomarkers in Heart Failure Divided by Pathophysiologic Pathway

Myocardial Stretch	
ANP, BNP, NT-proBNP, MR-proBNP	Neuregulin
GDF-15	sST2
Extracellular Matrix Remodeling	
MMPs (2, 3, 4, 8, and 9)	Myostatin
TIMP1	Syndecan-4
IL-6	Galectin-3
Collagen propeptides	sST2
N-terminal collagen type III peptide	Osteopontin
Myocyte Injury	
TnT, TnI, hsTn	Heart shock protein 60
CK-MB	sTRAIL
Myosin light-chain kinase I	Pentraxin 3
Heart-type fatty acid-binding protein	sFAS
Oxidative Stress	
MPO	Urinary and plasma isoprostanes
Oxidized LDLs	Urinary 8-hydpoxy-2′-deoxyguanosine
Urinary biopyrrins	Plasma malondialdehyde
Inflammation	
CRP	Leucine-rich 2-glycoprotein
TNF-α	Cytokines
LP-PLA2	Prolactin
TWEAK	Adiponectin
IL-6 (1, 10, 18)	Soluble endoglin
Adipokines	Serine protease PR3
FAS (APO-1)	S100A8/A9 complex
Osteoprotegerin	CA-125
Soluble TNF receptors 1 and 2	Pentraxin-3
YKL-40	Midkine
IL-1 receptor antagonist	
Neurohumoral Activation	
Norepinephrine	Adrenomedullin
Renin	MR-proADM
Angiotensin II	Chromogranin A and B
Aldosterone	Urocortin
Arginine vasopressin	Endothelin
Copeptin	
Renal Dysfunction	
Neutral gelatinase-associated lipocalin	FGF-23
NAG	Cystatin C
Kidney injury molecule-1	B-trace protein
Quiescin Q6	
Other	
RDW	MicroRNA
β2-microglobulin	Neprilysin
Triiodothyronine	Urinary albumin-to-creatinine ratio

ANP, Atrial natriuretic peptide; *BNP*, B-type natriuretic peptide; *CRP*, C-reactive protein; *NT-proBNP*, N-terminal portion of proBNP; *RDW*, red cell distribution width; *TNF-α*, tumor necrosis factor; *TWEAK*, tumor necrosis factor–like weak inducer of apoptosis.
From Ibrahim, N. E., & Januzzi, J. L. (2017). Beyond natriuretic peptides for diagnosis and management of heart failure. *Clin Chem, 63*, 211–222.

TABLE 34.3 Promising Novel Biomarkers for Use in Heart Failure Patients

Biomarker	Diagnosis	Prognosis	Therapy Guidance	Cardiac Production
NT-proBNP, BNP	++++	++++	++	Solely
MR-proANP	+++	++++	?	Solely
$proBNP_{1–108}$	?	+++	?	Solely
sST2	+	++++	?	Not exclusively
GDF-15	–	+++	?	Not exclusively
Gal-3	–	+++	?	Not exclusively
hsTn	+	++++	?	Solely
MPO	–	++	?	Not exclusively
CRP	–	++	?	No
TNF-α	–	++	?	No
IL-6	–	++	?	No
ET-1	–	++	?	Not exclusively
Copeptin	–	++	?	No
MR-proADM	–	++++	?	No
Cystatin-C	–	++++	?	No
NGAL	–	+++	?	No
Neprilysin	?	?	?	No

BNP, B-type natriuretic peptide; *CRP*, C-reactive protein; *NGAL*, neutral gelatinase-associated lipocalin; *NT-proBNP*, N-terminal portion of proBNP. From Ibrahim, N. E., & Januzzi, J. L. (2017). Beyond natriuretic peptides for diagnosis and management of heart failure. *Clin Chem, 63*, 211–222.

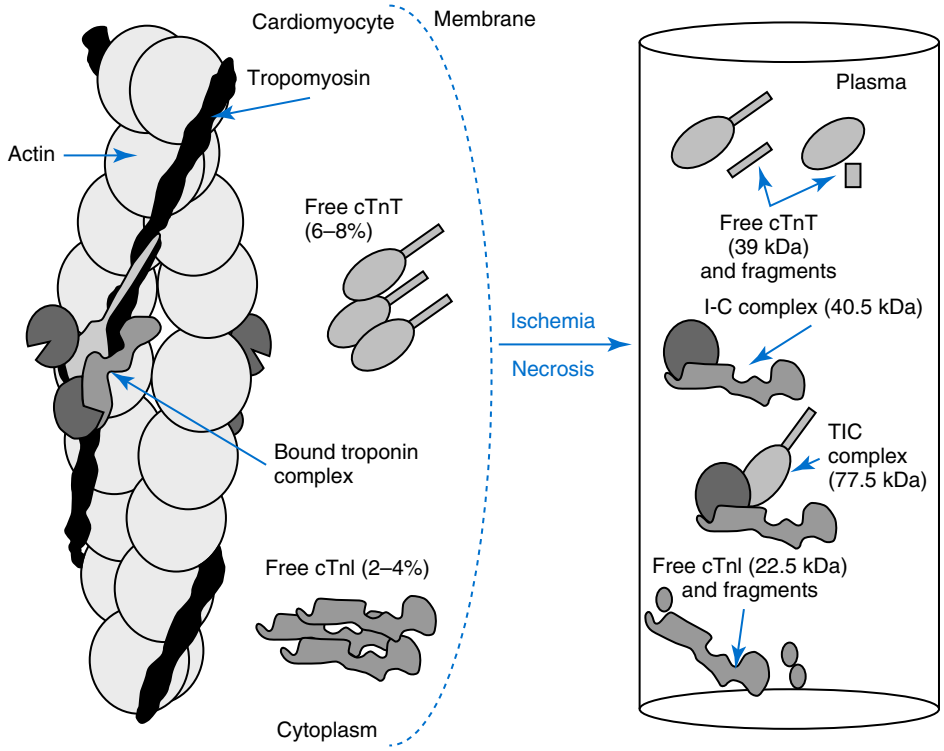


Fig. 34.5 Structure of cardiac troponin (cTn) complex and troponin forms released following myofibril necrosis. (From Gaze, D. C., & Collinson, P. O. [2008]. Multiple molecular forms of circulating cardiac troponin: analytical and clinical significance. *Ann Clin Biochem, 45*:349–359. Figure courtesy of Paul Collinson.)

Cardiac Troponins I and T

cTnI and cTnT are proteins found in cardiac muscle and are measured in the diagnosis of MI. cTnI is 100% cardiac tissue specific, while cTnT has been reported, although infrequently, to be expressed in skeletal muscle in several neuromuscular diseases.

Biochemistry

Three troponin subunits form a complex that regulates the interaction of actin and myosin and thus regulates cardiac contraction (Fig. 34.5). The three troponins are (1) troponin T (the tropomyosin-binding component), (2) troponin I (the inhibitory component), and (3) troponin C (the

calcium-binding component). Troponins are localized primarily in the myofibrils (94% to 97%) with a smaller cytoplasmic fraction (3% to 6%); although some research does challenge that the cytoplasm fraction is experimentally induced. cTnI and cTnT have different amino acid sequences encoded by different genes, and are different from the predominant cTns found in other muscle such as skeletal muscle. Human cTnI has an additional posttranslational 31–amino acid residue on the amino terminal end compared with skeletal muscle TnI, giving it cardiac specificity. Only one isoform of cTnI has been identified. cTnI is not expressed in normal, regenerating, or diseased human or animal skeletal muscle. cTnT is also encoded by a different gene than the one that encodes skeletal muscle isoforms. An 11–amino acid amino terminal residue gives this biomarker cardiac specificity. However, during human fetal development and in some diseased human skeletal muscle, small amounts of cTnT are expressed as one of four identified isoforms in skeletal muscle. In humans, cTnT isoform expression has been reported in skeletal muscle specimens obtained from patients with (1) muscular dystrophy, (2) polymyositis, (3) dermatomyositis, and (4) end-stage renal disease; there is some evidence that diseased skeletal muscle increases circulating cTnT immunoreactivity measured by the current hs-cTnT (US Food and Drug Administration [FDA]-cleared Gen 5) immunoassay. The other troponin, troponin C, is not useful as a cardiac biomarker as the troponin C expressed in the heart is not specific for the heart. Following myocardial injury, multiple forms of troponin appear both in tissue and in blood. These include (1) the complexes of cardiac troponins T, I, and C (T-I-C, or ternary complex); (2) complexes of I and C (I-C binary complex); and (3) free I. Also, multiple modifications of these three forms exist resulting from biochemical processes involving (1) oxidation, (2) reduction, (3) phosphorylation, (4) dephosphorylation, and (5) removal of the amino acids at the ends (C or N) of the molecules.

Analytical Considerations

Many manufacturers have developed assays to measure cTnI, with only one assay for cTnT. These assays are not standardized, and result in different 99th percentile URLs for each assay. A comprehensive listing updated in 2017 by manufacturers can be found on the International Federation of Clinical Chemistry (IFCC) website (http://www.ifcc.org/executive-board-and-council/eb-task-forces/task-force-on-clinical-applications-of-cardiac-bio-markers-tf-cb/).

Assays. Cummins and coworkers were the first to develop a cTn radioimmunoassay (RIA) using (polyclonal) anti-cTnI antisera, followed by an enzyme-linked immunosorbent cTnI assay, which eliminated the need for radioactivity and achieved high specificity by use of monoclonal antibodies.

Numerous manufacturers have since developed monoclonal antibody-based diagnostic immunoassays for the measurement of cTnI in serum, plasma, or whole blood, and most have been cleared by the FDA for patient testing within the United States. Typically, contemporary or point-of-care (POC) assays have limits of detection (LoD) in the range of 0.01 to 0.10 μg/L compared to high-sensitivity (hs)-cTn assays with LoDs of less than 1 ng/L. Analytically insensitive, qualitative assays of cTnI have also been commercialized but are not recommended for clinical care. hs-cTnI and hs-cTnT results are reported in ng/L, whereas contemporary/POC assay results are reported in μg/L. Circumstances that limit the ease of switching from one cTn assay to another in clinical practice or research include: (1) no primary reference cTnI material is currently available for manufacturers to use in standardizing cTnI assays, and (2) cTnI circulates in various forms and the "different antibodies" used in assays recognize different epitopes of cTnI.

For POC immunoassays, the National Academy of Clinical Biochemistry (NACB; now renamed the American Association of Clinical Chemistry [AACC]) has developed Laboratory Medicine Practice Guidelines for use in ACS. These guidelines address (1) administrative issues, (2) cost-effective utilization, and (3) clinical and technical performance of cardiac biomarkers in the emergency medicine department. Proposed elements of the guidelines include:

1. Members of emergency medicine departments, primary care physicians, cardiologists, hospital administrators, and clinical laboratory staff should work collectively to develop an accelerated protocol for the use of biomarkers in the evaluation of patients with possible ACS.
2. Quality assurance measures should be used with monitoring to reduce medical errors and improve patient treatment.
3. The laboratory should perform biomarker testing with a maximum turnaround time (TAT) of 1 hour, optimally 30 minutes. The TAT is defined as the time from blood collection (0 hour) to reporting of results to the provider. Institutions that cannot consistently provide a 1-hour TAT should consider implementing POC assays. Performance specifications and characteristics for central laboratory and POC assays should not differ. POC assays should provide quantitative results.

Specimen requirements. For both central laboratory and POC testing, and for practical considerations, the optimal specimens for rapid processing and testing appear to be anticoagulated whole blood or plasma. This eliminates the extra time needed for clotting and additional sample handling. However, differences have been described among plasma, whole blood, and serum specimens for cTnI concentrations measurement by an individual assay. Some anticoagulants, as well as other factors, are known to interfere with cTnI and cTnT antibody-binding affinity, and produce matrix effects (Fig. 34.6). In addition, it is recommended that different sample types not be used during an individual's workup when serial, timed samples are being drawn to rule in or out a MI. It should be noted that at present, results obtained from all FDA-cleared POC assays are substantially less sensitive than

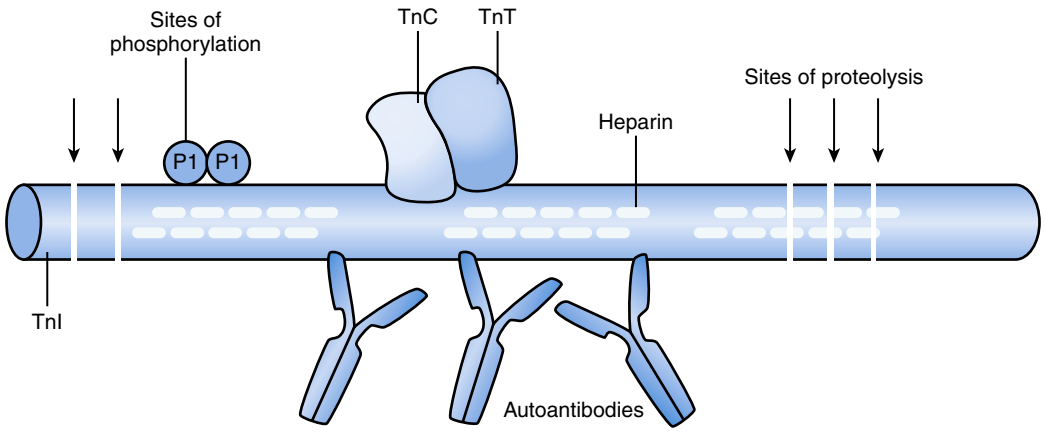


Fig. 34.6 Factors that can influence precise measurement of cTnI circulating in human blood. (With permission from HyTest Ltd.)

central laboratory-based contemporary and hs-assays, and that should be acknowledged by both laboratorians and clinicians when such assays are used. The higher limit of quantification (LoQ; defined at the lowest concentration with a 20% coefficient of variation [CV]) means that small increases of cTn at early time points after AMI will not be detected with these assays. This fact works against the goal of early detection of AMI.

Standardization/harmonization. Standardization of cTn assays remains elusive. The cTnI Standardization Subcommittee of the AACC in collaboration with the National Institute of Standards and Technology (NIST) have produced a cTnI standard reference material (SRM #2921) that is a TnC-cTnI-cTnT complex purified from human heart under non-denaturing conditions. Because this material was found to be commutable with only 50% of current cTnI assays, it is of limited value for assay harmonization of current assays and is not useful as a common calibrator. It does, however, allow for traceability to a common reference material. Currently an IFCC Working Group is working toward developing a secondary serum-based reference material, but it will unlikely provide standardization.

New clinical laboratory practice recommendations for the use of cTn assays in ACS were proposed in 2017 as an expert opinion from the AACC Academy and the IFCC Task Force-Cardiac Biomarkers (TF-CB). This document focuses on the following 10 topics: (1) quality control (QC) utilization, (2) assay limits validating the lower reportable analytical limits, (3) units to be used in reporting measurable concentrations for patients and QC materials, (4) 99th percentile sex-specific URLs to define normality, (5) criteria required to define hs-cTn assays, (6) communication with clinicians and the role laboratories play in educating clinicians on the influence of preanalytical and analytical problems that confound assay results, (7) studies on hs-cTn assays and how authors need to document preanalytical and analytical variables, (8) harmonizing and standardizing assay results and the role of commutable materials, (9) time to reporting of results from sample receipt and sample collection, and (10) changes in hs-cTn concentrations over time and the role of both analytical and biological variabilities in results of serial blood collections. Highlights of this document are shown in Box 34.3.

Defining Healthy Reference Populations for Determining 99th Percentile

Representative 99th percentiles for contemporary, POC, and hs-assays are shown in Table 34.4. There is poor consistency in the composition or numbers of individuals enrolled for determining the 99th percentile. Defining what constitutes a healthy reference individual is a topic of debate. Should individuals be apparently healthy, young individuals, or should they be age-matched patients hospitalized without known cardiovascular disease, similar to the demographics of patients who rule in for an AMI, with ages of 30 to 90 years? How does one determine who is healthy? Should individuals be selected: (1) via personal interview with questions addressing known cardiac medications, such as statins; (2) after obtaining information about known diseases associated with cardiovascular disease, such as renal disease or diabetes; or (3) via definitive physician evaluation of an individual after taking a history and physical examination, including an ECG, echocardiogram or imaging? Both younger (<30 years) and older (>30 years, with median age of 60 to 65 years to be representative of cardiac patients), apparently healthy reference groups should be recruited. Inclusion criteria should be based on data obtained from a history of medications (not on cardiac drugs, i.e., statins), known underlying diseases, and blood surrogate biomarker measurements: NP concentration for ruling out underlying myocardial dysfunction, an estimated glomerular filtration rate (eGFR) for defining underlying renal insufficiency, and HbA1C for underdiagnosed diabetes.

At present, no definitive, multidiscipline, peer-reviewed surrogate biomarker recommendations have been published. Men and women should be equally represented, with a diverse racial and ethnic mix. The number of individuals for determining a 99th percentile has been defined by the IFCC TF-CB to be a minimum of 300 for men and 300 for women

BOX 34.3 Clinical Laboratory Practice Recommendations of the AACC Academy and IFCC Task Force Task on Clinical Applications of Cardiac Bio-Markers for the Use of Cardiac Troponin in Acute Coronary Syndromes

- For hs-cTn assays, laboratories should measure at least three different concentrations of quality control materials at least once per day.
- During initiation of hs-cTn testing, clinical laboratories should validate the limit of blank, limit of detection, or limit of quantification as applicable.
- Report hs-cTn assays in whole numbers using ng/L without decimal points, and report quality control values to one decimal point.
- Use a defined reference population to report 99th percentiles according to sex-specific cutoffs for hs-cTn assays.
- Assays unable to detect cTn at concentrations at or above the limits of detection in at least 50% of healthy men and women, individually, should be labeled as contemporary and not hs-cTn assays.
- For hospitals using two or more cTn assays, differences in the sensitivity of the various cTn assays should be explained to clinicians so they can better understand discrepancies when patients are transferred between facilities.
- Cardiac troponin results should be reported within 60 mins of a sample being received, with the first sample labeled 0 h.
- Determining the degree of serial changes for cTn values assists in differentiating between patients who have acute cardiac injury, including acute myocardial infarction, and those who have more chronic elevations, many of which may be related to structural heart disease.

cTn, Cardiac troponin; *hs*, high sensitivity.
From Apple, F. S., Jaffe, A. S., Collinson, P., et al., on behalf of the IFCC Task Force on Clinical Applications of Cardiac Bio-Markers. (2015). IFCC educational materials on selected analytical and clinical applications of high-sensitivity cardiac troponin assays. *Clin Biochem 48,* 201–203; Wu, A. H. B., Christenson, R. H., Greene, D. N., Jaffe, A. S., Kavsak, P. A., Ordonez-Llanos, J., et al. (2018). Clinical laboratory practice recommendations for the use of cardiac troponin in acute coronary syndrome: expert opinion from the Academy of the American Association for Clinical Chemistry and the Task Force on Clinical Applications of Cardiac Bio-Markers of the International Federation of Clinical Chemistry. *Clin Chem, 64,* 645–655.

(based on 90% confidence intervals). No study has clinically compared all contemporary sensitive assays and/or hs-assays within the same reference or disease population. However, one study has compared multiple assays (n = 19) with the same reference group demonstrating substantial differences for both 99th percentiles and ability to measure concentrations above an LoD, as shown in Fig. 34.7.

Because most laboratories do not have (1) the resources to perform adequately powered 99th percentile reference studies, nor (2) the ability to carry out Clinical and Laboratory Standards Institute (CLSI) protocols to establish total imprecision criteria, these clinical laboratories must rely on the peer-reviewed published literature to assist in establishing both local reference limits and imprecision characteristics. Caution must be taken when comparing the findings reported in the manufacturers' package inserts with the findings reported in the literature because of differences in (1) total sample size, (2) distributions by sex and ethnicity, (3) age ranges, and (4) statistical methods used to calculate the 99th percentile. The IFCC TF-CB website offers the most up-to-date list of cTn assays and 99th percentiles.

Guidelines and acute myocardial infarction. Guidelines have been developed by the (1) Global Task Force for the Fourth Universal Definition of Myocardial Infarction, (2) the joint AACC Academy and IFCC TF-CB document, (3) the ESC/ACC, (4) the American Heart Association (AHA)/ACC), and (5) the Acute Decompensated Heart Failure National Registry (ADHERE) Scientific Advisory Committee. Desirable imprecision of each cTn assay has been defined as less than 10% CV at the 99th percentile reference limit, and is required for an hs-assay. A percentage of CVs up to 20% are acceptable for clinical practice, with assays having greater than 20% CV at the 99th percentile used with caution if no other choice is available.

Clinical utility. Measurement of cTn has become the cornerstone of MI diagnosis (see Box 34.1), with Fig. 34.8 showing the kinetics of the rise and fall of cTn for an hs-cTnI assay for type 1 MI, type 2 MI, chronic myocardial injury, and a nonmyocardial injury patient. High-sensitivity cTn assays also play a substantial role in shortening the time for monitoring cTn to 3 hours, compared to 6 hours with contemporary assays, as shown in Fig. 34.9. Further, in the appropriate low-risk clinical setting, hs-cTn assays now allow for a very early rule-out of MI. Numerous studies have documented that 20% to 50% of patients presenting to emergency departments can be safely discharged predicated on a baseline (single) hs-cTn concentration of less than the LoD, with greater than 99.5% negative predictive value and clinical sensitivity.

B-Type Natriuretic Peptide

In 1981, the Canadian physiologist Adolfo de Bold and his colleagues reported that infusion of atrial tissue extracts produced (1) an increase in the renal excretion of sodium and water, (2) a rapid decrease in BP, and (3) an increase in blood hematocrit. The responsible substance was identified as a peptide comprising 28–amino acid residues and eventually named atrial natriuretic peptide (ANP). Later, brain natriuretic peptide (BNP) and C-type natriuretic peptide (CNP), two structurally related peptides, were isolated and identified in the porcine brain. However, because BNP is mainly expressed in the heart, it was later renamed B-type natriuretic peptide (BNP).

Biochemistry

BNP is a hormone that is mainly released from the myocardial ventricles. Fig. 34.10 illustrates the synthesis of the preprohormone and subsequent secretion of NP proteins from the cardiac myocytes. It is not known whether proBNP is split in the myocyte or later in the plasma. The major circulating forms are NT-proBNP, which has an unknown function;

TABLE 34.4 Analytical Characteristics of Commercial and Research Contemporary, POC, and hs-cTnI and hs-cTnT Assays

Company/Platform/Assay	LoD (μg/L)	99th (μg/L)	% CV at 99th	10% CV (μg/L)	Epitopes/Antibodies	Detection Tag
Contemporary Assays						
Abbott ARCHITECT	<0.01	0.028	15	0.032	C: 87–91, 24–40; D: 41–49	Acridinium
Beckman Access 2	0.01	0.02	14	0.040	C: 41–49; D: 24–40	ALP
Beckman Coulter DxI	0.01	0.03	14	0.040	C:41–49; D:28–80	ALP
Ortho-Clinical Diagnostics Vitros	0.012	0.034	10	0.034	C: 24–40, 41–49; D: 87–91	HRP
Siemens Centaur Ultra	0.006	0.04	10	0.030	C: 41–49, 87–91; D: 27–40	Acridinium
Siemens Dimension RxL	0.04	0.07	20	0.140	C: 27–32; D: 41–56	ALP
Siemens VISTA	0.015	0.045	10	0.040	C: 27–32; D: 41–56	Chemilumi-nescent
Tosoh AIA	0.06	<0.06	8.5	0.090	C: 41–49; D: 87–91	ALP
Roche 4th generation cTnT	0.01	0.01	18	0.03	C: 125–131; D: 136–147	Ruthenium
POC Assays						
Abbott i-STAT	0.02	0.08	16	0.10	C: 41–49, 88–91; D: 28–39,62–78	ALP
Alere Triage	0.05	<0.05	NA	NA	C: NA; D: 27–40	Fluorophore
bioMérieux Vidas	0.01	0.01	27	0.11	C: 41–49, 22–29; D: 87–91, 7B9	ALP
LSI Medience PATHFAST	0.008	0.029	5.1	0014	C: 41–49; D:71–116, 163–209	ALP
Radiometer AQT90 cTnI	0.0095	0.023	17	0.039	C: 41–49, 190–196; D: 137–149	Europium
Radiometer AQT90 cTnT	0.01	0.017	20	0.03	C: 125–131; D:136–147	Europium
Response Biomedical RAMP	0.03	<0.1	18	0.21	C: 85–92; D: 26–38	Fluorophore
Roche Cardiac Reader	<0.05	<0.05	NA	NA	C: 125–131; D:136–147	Gold particles
Siemens Stratus CS	0.03	0.07	10	0.06	C: 27–32; D: 41–56	ALP
Trinity Meritas	0.019	0.036	17	NA	C: 24–40, 41–49; D: 88–90, 137–148,190–196	Fluorophore
ET Health	0.1	0.2	NA	0.42	C: 87–91; D: 27–40	ALP
Nanomix	0.15	NA	NA	0.64	C: 87–91; D: 27–40	ALP

hs-Assays	ng/L	M/F, ng/L	% CV at 99th	10% CV, ng/L	Epitopes/antibodies	Detection tag
Abbott ARCHITECT hs-cTnI	1.2/1.9	34/16	5	3	C: 24–40; D: 41–49	Acridinium
Beckman Coulter Access hs-cTnI	2.5	52/23	<10	8	C: 41–49; D: 24–40	ALP
Ortho-Clinical Diagnostics hs-cTnI	1.0	19/16	<10	3	C: 24–40, 41–49; D: 87–91	HRP
Roche E170 hs-cTnT	5	20/13	8	13	D: 125–131; C: 136–147	Ruthenium
Siemens Vista hs-cTnI	0.5	55/33	5	3	C: 30–35; D: 41–56,171–8	Lumines-cence
Singulex Errena hs-cTnI	0.09	36/30	5	0.9	C: 41–49; D: 27–41	Fluorescence

Assays not designated as cTnT are cTnI.
C, Capture antibody; *D*, detection antibody; *NA*, not available.
From Apple, F. S., Sandoval, Y., Jaffe, A. S., Ordonez-Llanos J, for the IFCC Task Force on Clinical Applications of Cardiac Bio-Markers. (2017). Cardiac troponin assays: guide to understanding analytical characteristics and their impact on clinical care. *Clin Chem, 63*, 73–81.

proBNP, for which the function is also unknown; and BNP, the physiologically active hormone, which is the C-terminal part of pro-BNP. BNP is cleared via degradation by endopeptidases, by receptor-mediated clearance, and perhaps via the kidneys, which also secretes BNP. The NT-proBNP fragment is not cleared via receptor-mediated mechanisms, but is thought to be cleared predominantly by the kidneys. Therefore, it is more sensitive to changes in renal function. The majority of research on NPs in CHF has focused on BNP and NT-proBNP. However, investigations have now described substantial amounts of circulating proBNP, which cross-reacts in BNP and NT-proBNP assays.

Analytical Considerations

A number of immunoassays are available to measure BNP and NT-proBNP that use antibodies directed to different epitopes located on the antigen molecules (Fig. 34.11). Degradation of BNP is known to occur by cleavage at serine and proline residues in vivo and in vitro (Fig. 34.12); cleavage is mediated by proteases, which make the concentration of BNP unstable in collected blood. An improved understanding of potential cross-reactivity with split products of the N-terminal portion of NT-proBNP and proBNP itself is needed. For BNP, ethylenediaminetetraacetic acid (EDTA)-anticoagulated whole blood or plasma appears to be the only

acceptable specimen choices. For NT-proBNP, serum, heparin plasma, and EDTA plasma appear acceptable, although results trend slightly lower with the latter. Plastic blood collection tubes are necessary for BNP, whereas for NT-proBNP, either glass or plastic is acceptable. Recent studies have shown that (1) BNP (Fig. 34.13A) and NT-proBNP (see Fig. 34.13B) assays demonstrated substantial cross-reactivity with proBNP peptides, (2) NT-proBNP assays do not detect glycosylated forms of NT-proBNP or proBNP, and (3) proBNP assays preferentially detect the BNP 1–32 peptide and have minimal cross-reactivity with BNP peptides and glycosylated proBNP. Thus, BNP or NT-proBNP results are not transferable among current immunoassays due to their differences in cross-reactivity and ability to detect

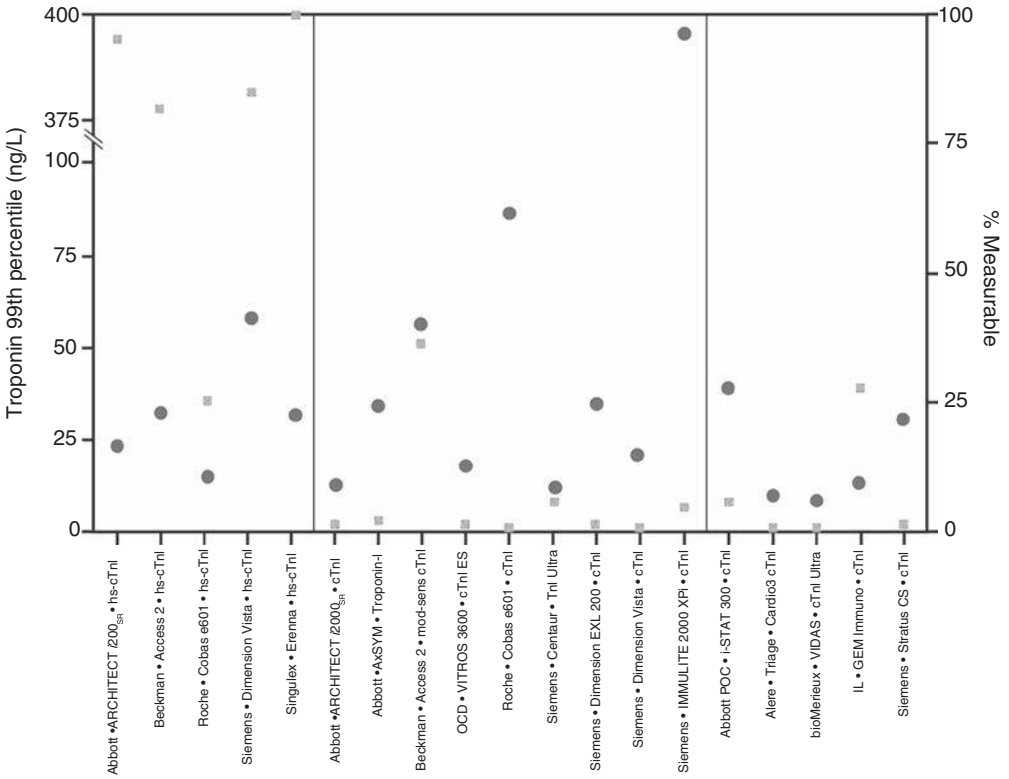


Fig. 34.7 Comparison of both 99th percentile values *(circles)* and percent measurable concentrations *(boxes)* in a presumably healthy population for 19 cardiac troponin assays designated as high-sensitivity (hs) *(left)*, contemporary *(middle)*, and POC *(right)*. (From Apple, F. S., Ler, R., & Murakami, M. M. [2012]. Determination of 19 cardiac troponin I and T assay 99th percentile values from a common presumably healthy population. *Clin Chem, 58*, 1574–1581.)

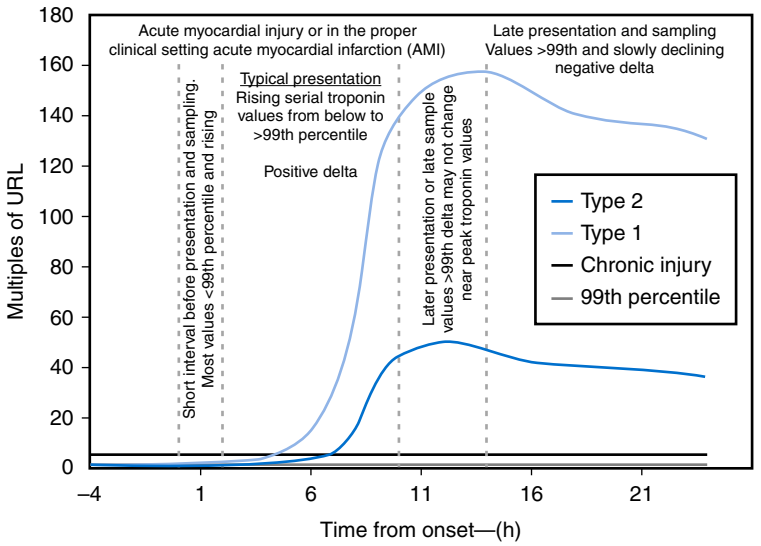


Fig. 34.8 Kinetics for the general rise and fall pattern of cardiac troponin for type 1 myocardial infarction (MI), type 2 MI, and chronic myocardial injury. (With permission from the IFCC Task Force for Clinical Application of Cardiac Bio-Markers [TF-CB]; prepared by Paul Collinson.)

various glycosylated forms of proBNP-derived fragments. It is interesting that, despite a lack of standardized NP assays, clinical cardiovascular and heart failure guidelines recommend uniform cut points for BNP and NT-proBNP assays.

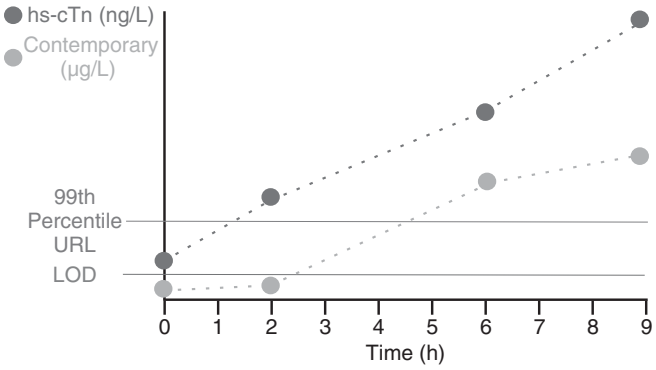


Fig. 34.9 Cardiac troponin kinetics comparing high-sensitivity and contemporary assays in a patient presenting within 30 minutes of acute myocardial infarction. *LOD,* Limit of detection; *URL,* upper reference limit. (From Rifai, N. [2018]. *Tietz Textbook of Clinical Chemistry and Molecular Diagnostics* [6th ed.]. St. Louis: Elsevier.)

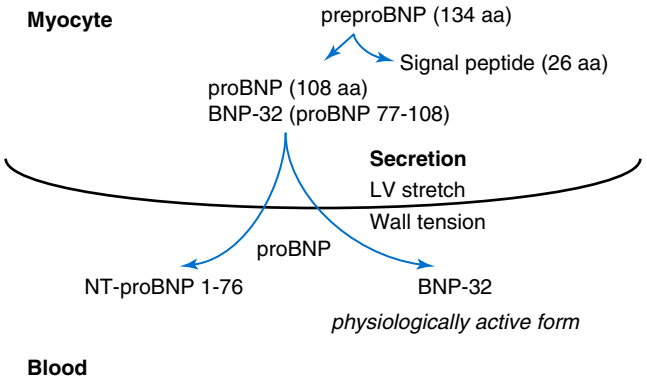


Fig. 34.10 The biotransformation and release of BNP and NT-proBNP from the myocyte into the circulation. *aa,* Amino acid; *LV,* left ventricular.

Opportunities to improve standardization and harmonization efforts for NP assays and ultimately to improve clinical care for heart failure patients, particularly with the advent of new pharmacologic agents, still exist.

Guidelines. In 2005, the IFCC established recommended analytical and preanalytical quality specifications for NP assays. These specifications were intended for use by the manufacturers of commercial assays and by clinical laboratories that use NP assays. Subsequently, the NACB and the IFCC committee developed guidelines pertaining to analytical issues for biomarkers of heart failure.

Recommendations from these guidelines include:

1. Reference limits (95th or 97.5th percentile) should be established independently for both BNP and NT-proBNP based on age (by decade) and by sex. Each commercial assay should be validated separately. The effects of ethnicity and renal function need to be evaluated as possible independent variables.
2. Receiver operating characteristic (ROC) curves (Fig. 34.14) should be established to evaluate clinical effectiveness and to establish optimal medical decision cutoffs for BNP and NT-proBNP assays for diagnostic usefulness.
3. Assays for BNP and NT-proBNP should have total imprecision (%CV) of less than or equal to 15% at their age- and sex-defined URLs, and at the NYHA-defined medical decision concentrations.

Before introduction into clinical practice, BNP and NT-proBNP assays must be characterized with respect to the following preanalytical and analytical issues:

1. Preanalytical:
   a. Effects of storage time and temperature
   b. Influence of different anticoagulants
   c. Influence of gel separator tubes
   d. Need for plastic blood collection tubes for BNP; for NT-proBNP, either glass or plastic is acceptable

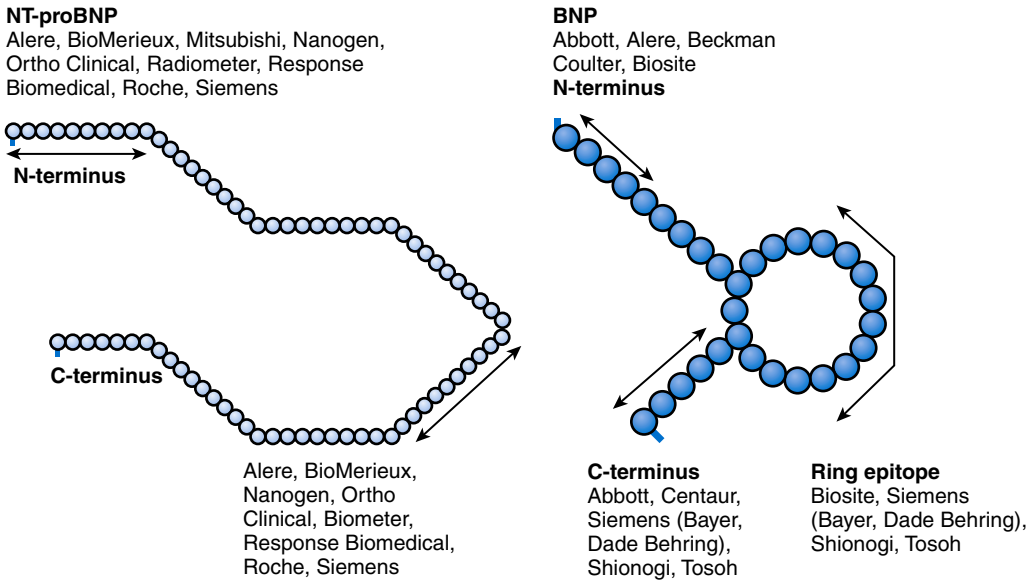


Fig. 34.11 Epitope locations on BNP and NT-proBNP used by commercially available assays. (From Vasile, V. C., & Jaffe, A. S. [2017]. Natriuretic peptides and analytical barriers. *Clin Chem, 63,* 50–58.)

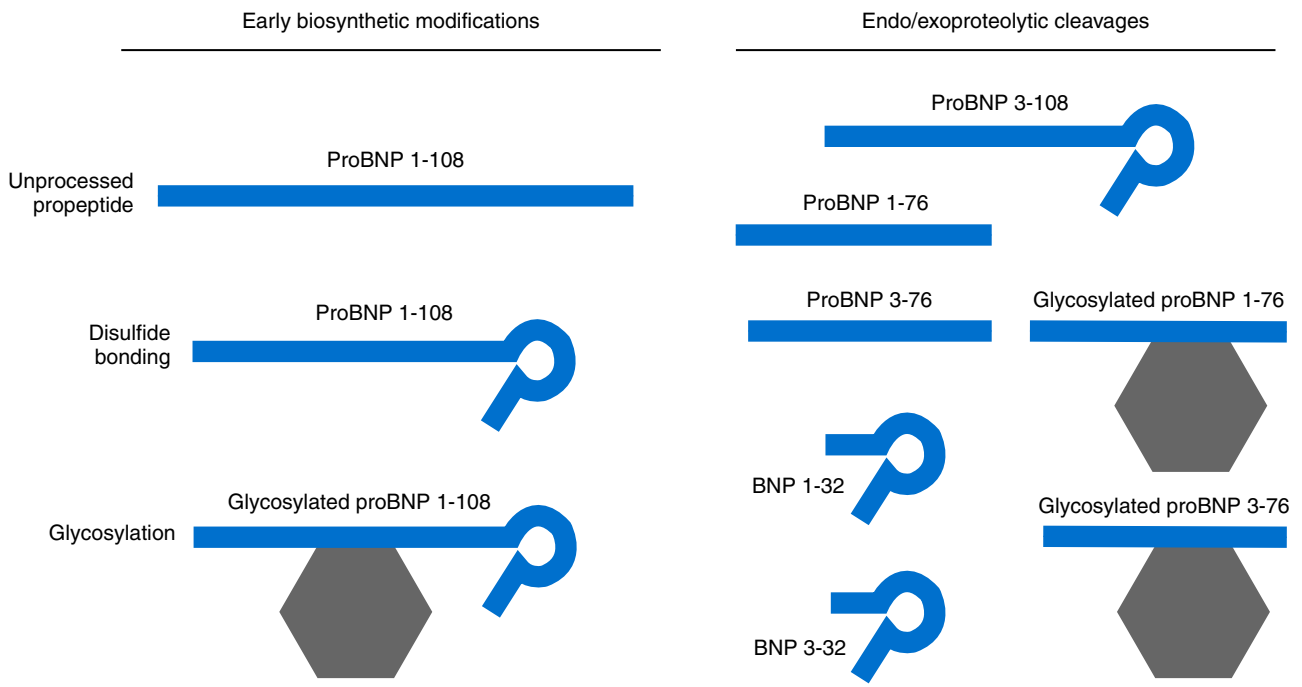


Fig. 34.12 Schematic of possible proBNP-derived peptide products. Note that most peptides are not chemically identified, but are instead suggested by biochemical methods that rely on antibody recognition. Carbohydrate is indicated by the *shaded hexagons*. (From Rifai, N. [2018]. *Tietz Textbook of Clinical Chemistry and Molecular Diagnostics* [6th ed.]. St. Louis: Elsevier.)

2. Analytical:
   a. Identification of the epitopes in the NP that are recognized by the reagent antibodies
   b. Cross-reactivity characteristics with related NPs, including NT-proANP, ANP, CNP, BNP, and glycosylated and nonglycosylated NT-proBNP and proBNP
   c. Identification of interference from heterophile antibodies, rheumatoid factors, and human antimouse antibodies
   d. Description of calibration material used, how the material was defined, and the concentration value assigned
   e. Clarification of dilution response
3. For both BNP and NT-proBNP, measurements should be reported in ng/L, not in pmol/L; and patient specimen comparisons and regression analysis should be performed, in accordance with CLSI guidelines, to establish the degree or lack of harmonization across the dynamic range of each assay. Specifically, harmonization around the current presumed optimal diagnostic medical decision cutoff of 100 ng/L for BNP should be validated.
4. For both BNP and NT-proBNP, biological variability has been determined to be at least 50%. Therefore, caution should be exercised in interpreting concentration changes of less than 50% to 80% as reflective of medical therapy. However, consistent trends should be followed as clinically important.

Reference intervals. At present, multiple companies have FDA-cleared BNP and NT-proBNP assays. A comprehensive listing assays, updated in 2017 by manufacturers, can be found on the IFCC website (http://www.ifcc.org/executive-board-and-council/eb-task-forces/task-force-on-clinical-applications-of-cardiac-bio-markers-tf-cb/). Both laboratory-based assays and POC assays based on the use of whole blood are available. Research assays for proBNP and a novel assay that is capable of distinguishing BNP 1–32 from proBNP have been described. There is also a midregional proatrial natriuretic peptide (MR-proANP) that is equivalent to BNP as a biomarker to rule out and confirm HF. As the number of assays for all four NP biomarkers grows, it is even more essential that appropriate clinical and analytical assay criteria are uniformly adapted.

There are several practical issues regarding the use of serum/plasma/whole blood monitoring of NPs. First, reference intervals vary depending on the assay and the reference population used. Second, a number of biological factors affect the BNP and NT-proBNP concentrations, most importantly (1) age, (2) sex, (3) obesity, and (4) renal function. Significant differences are observed between men and women (higher), and concentrations increase with age, as shown in Fig. 34.15 for BNP (A) and NT-proBNP (B). For both BNP and NT-proBNP, there is an inverse relationship between concentrations and body mass index. For NT-proBNP, review of both the FDA-approved US package insert and the European assay package insert reveals substantial differences in the concentrations that are considered as reference intervals by age and sex.

For BNP, a single cutoff at 100 pg/mL (ng/L) has been designated, likely driven by the FDA clearance at this

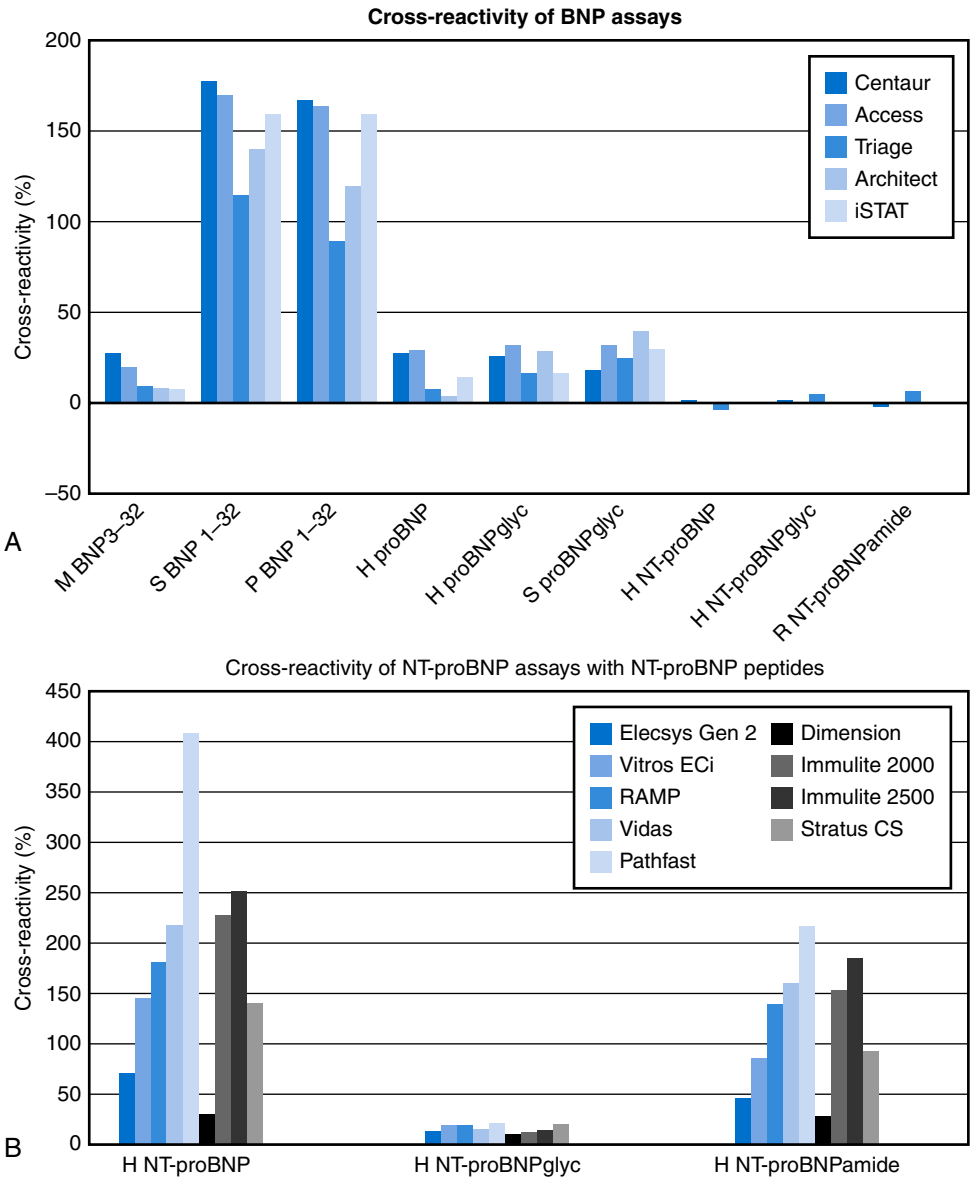
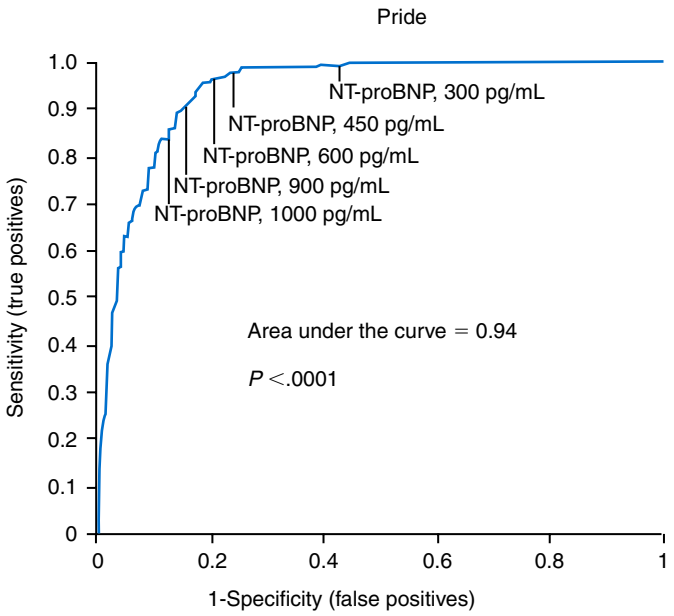
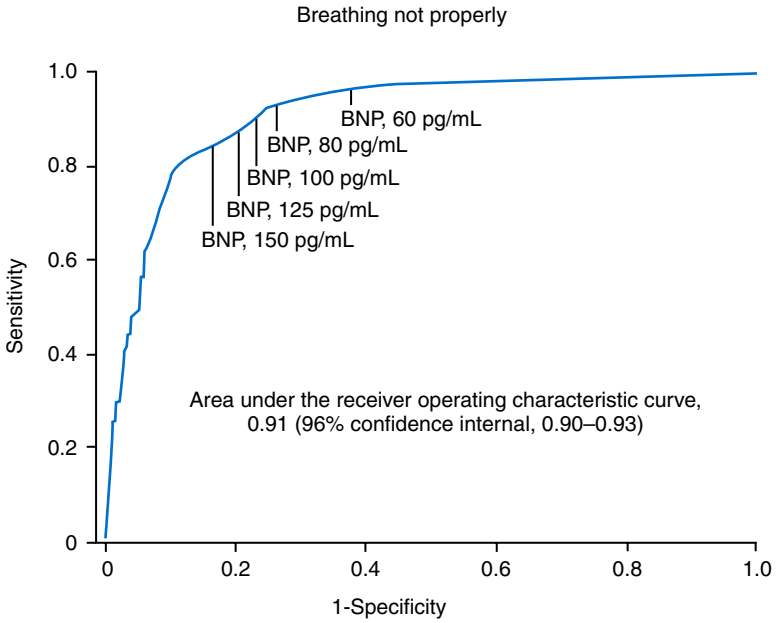


Fig. 34.13 Cross-reactivity profiles for (A) BNP assays and (B) NT-proBNP assays with BNP and NT-proBNP peptides, respectively. (From Saenger, A. K., Rodriguez-Fraga, O., et al. [2017]. Specificity of B-type natriuretic peptide assays: cross-reactivity with different BNP, NT-proBNP, and proBNP peptides. *Clin Chem*, *63*, 351–358.)

value as the ROC curve value optimized for diagnostic accuracy based on the Breathing Not Properly trial. For NT-proBNP, the FDA-cleared assays had cutoffs that are age based: 125 ng/L for patients younger than 75 years and 450 ng/L for patients older than 75 years. These cutoffs appear to misclassify many normal subjects. Clinical studies (proBNP Investigation of Dyspnea in the Emergency [PRIDE]) and the International Collaborative on NT-proBNP [ICON]) have more appropriately defined age-derived and renal function–derived cutoffs as shown in Fig. 34.15B, designated by 50 years of age and eGFR at 60 mL/min per 1.73 m^2, with an optimal negative predictive value of less than 300 ng/L that is age, gender, and eGFR independent. Obesity also has been shown to have an association with BNP and NT-proBNP concentrations, with an inverse relationship between body mass index and BNP in CHF patients. Finally, a lack of understanding of the physiological and biological variability of BNP and NT-proBNP in HF patients may cause clinicians to misinterpret changing (increasing or decreasing) BNP and NT-proBNP concentrations in the context of establishing the success or failure of therapy. It has been shown that both BNP and NT-proBNP exhibit a within-subject biological variability of 35% to 45%. Thus, when considering what is clinically, significantly different between serial BNP and NT-proBNP values, a reference change value of 80% is necessary. This is what has been documented in patients with chronic HF.



Cut point	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	Accuracy
300 pg/mL	99%	68%	62%	99%	79%
450 pg/mL	98%	76%	68%	99%	83%
600 pg/mL	96%	81%	73%	97%	86%
900 pg/mL	90%	85%	76%	94%	87%
1000 pg/mL	87%	86%	78%	91%	87%



BNP (pg/mL)	Sensitivity (%)	Specificity (%)	Positive Predictive Value (%)	Negative Predictive Value (%)	Accuracy (%)
50	97 (98–98)	62 (60–66)	71 (68–74)	96 (94–97)	79
80	98 (91–96)	74 (70–77)	77 (76–80)	92 (89–94)	83
100	90 (88–92)	76 (73–78)	79 (78–81)	92 (87–91)	83
125	87 (86–90)	79 (78–82)	80 (78–83)	87 (84–89)	83
150	86 (82–88)	83 (80–86)	83 (80–86)	85 (83–88)	84

Fig. 34.14 Receiver operating characteristic (ROC) analysis for brain natriuretic peptide (BNP) and N-terminal pro B-type natriuretic peptide (NT-proBNP) for the diagnosis of acute heart failure. (From Rifai, N. [2018]. *Tietz Textbook of Clinical Chemistry and Molecular Diagnostics* [6th ed.]. St. Louis: Elsevier.)

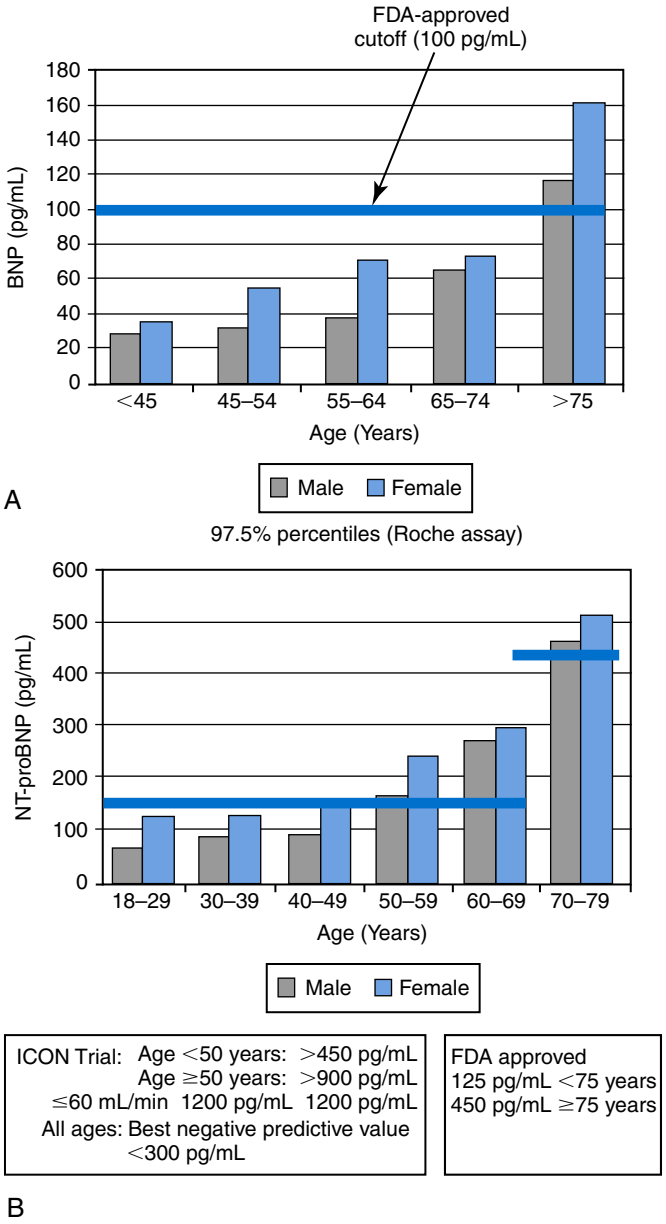


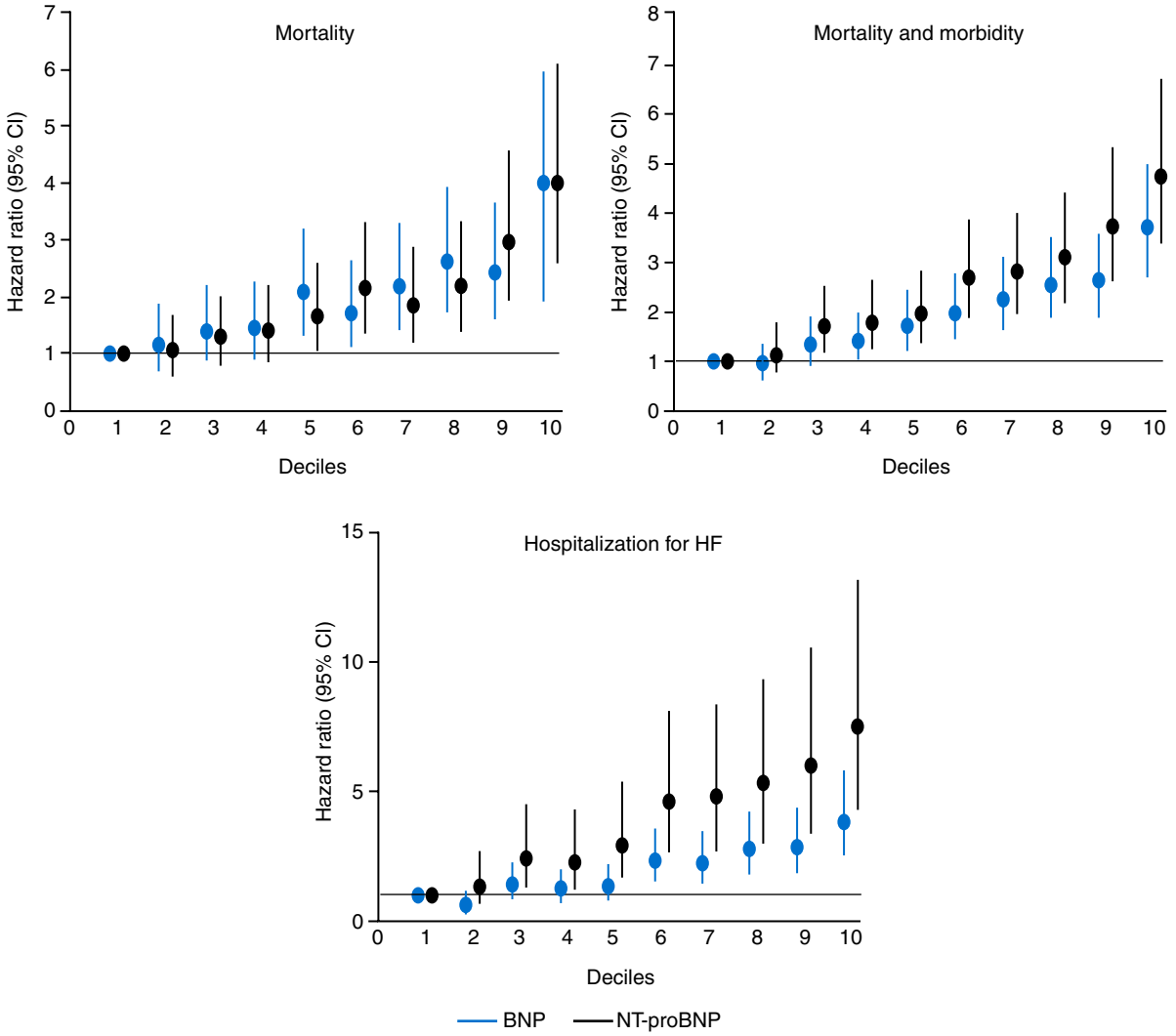
Fig. 34.15 Representative (A) BNP and (B) NT-proBNP concentration distributions in normal males and females by decade (years) with indication of cutoff values by assay. (From Rifai, N. [2018]. *Tietz Textbook of Clinical Chemistry and Molecular Diagnostics* [6th ed.]. St. Louis: Elsevier.)

Clinical Utility

BNP has a multiplicity of cardiac functions and is released as a counterregulatory hormone in response to a variety of cardiac stresses, but most particularly cardiac stretch. It is significantly affected by changes in volume and in cardiac performance, and among its effects are fluid-volume reduction and vasodilation. BNP and NT-proBNP depend on age and sex, are increased in CHF, and are correlated with its severity. NP measurements are useful (1) in identifying patients with moderate to severe CHF and (2) **risk stratification** of CHF patients and of patients who present with ACS. Higher NP values are in general associated with a more adverse prognosis (Fig. 34.16) in both the acute and chronic situations. Perhaps the most interesting use of NPs is in the patient with HF with the hope of finding the ability to titer therapy more efficaciously. BNP and NT-proBNP provide information that is synergistic with the measurement of cTn in these settings and is especially useful for risk stratification when cTn concentrations are within the reference intervals.

Cardiac Biomarkers With Potential and Limited Utilization in Clinical Practice

Several biomarkers that have been used in a limited fashion in clinical practice, but not globally endorsed, are as follows. First, copeptin, a 30–amino acid glycoprotein that constitutes the C-terminal portion of arginine vasopressin,



Covariates for adjustment: age, gender, NYHA class, ischemic etiology, LVEF, LVIDd, serum creatinine and bilirubin, randomized treatment, prescription of β-blockers, digitalis and diuretics, presence of AF or diabetes at study entry.

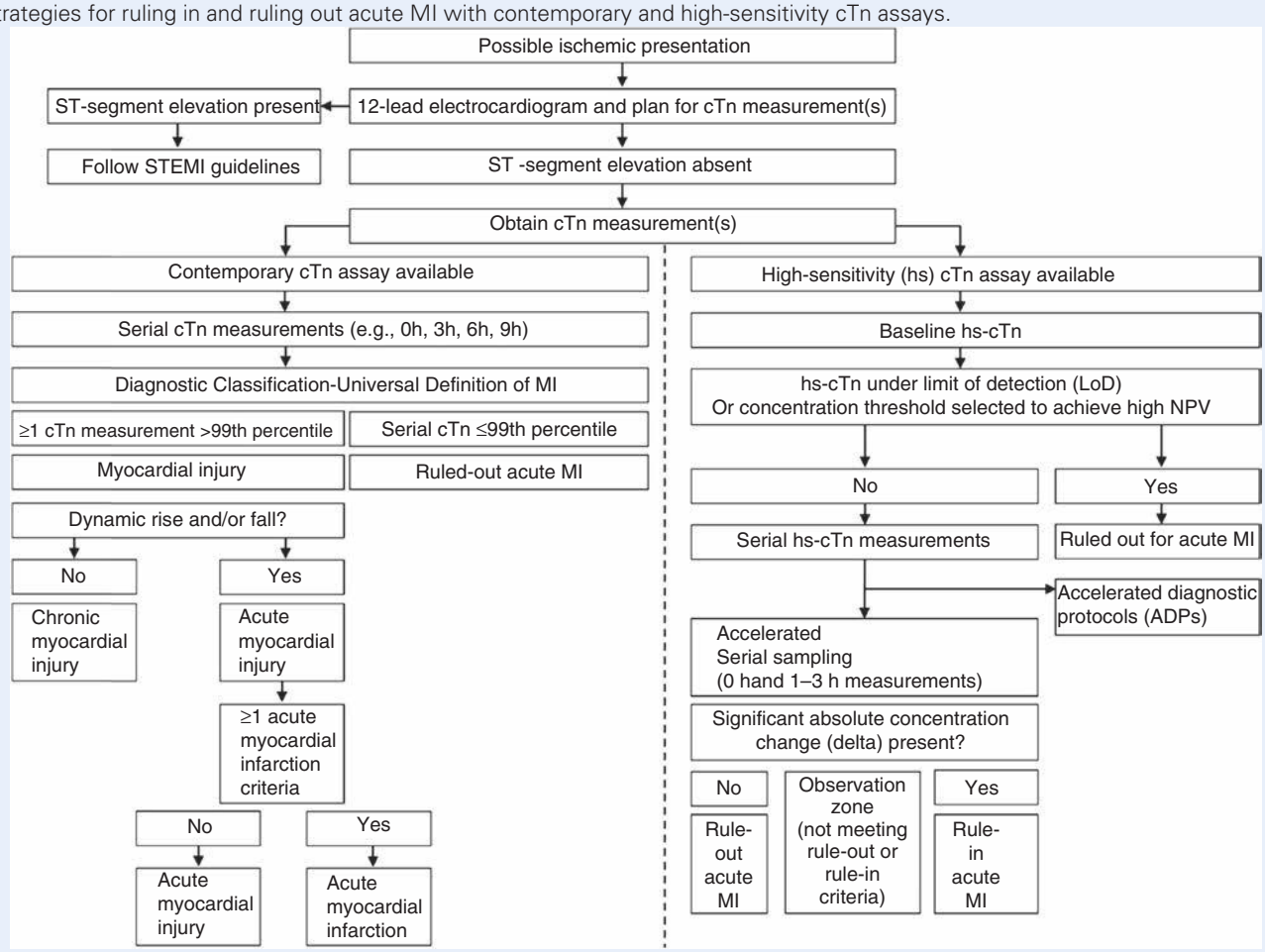
Fig. 34.16 Relationships of BNP and NT-proBNP values and outcomes. (From Rifai, N. [2018]. *Tietz Textbook of Clinical Chemistry and Molecular Diagnostics* [6th ed.]. St. Louis: Elsevier.)

has been described as a rapid and early rule-out biomarker for AMI at presentation in patients with symptoms suggestive of ACS with a normal cTn value. Second, galectin-3—a protein that is a member of the lectin family and that contains a carbohydrate-recognition-binding domain (CRD) of about 130 amino acids that enable the specific binding of β-galactosides—has been shown to be involved in several biological processes. Clinically, galectin-3 has been found to be involved in (1) cancer, (2) inflammation, (3) fibrosis, (4) stroke, and (5) heart disease. It has also been shown that the expression of galectin-3 is implicated in a variety of processes associated with (1) heart failure, (2) myofibroblast proliferation, (3) fibrogenesis, (4) tissue repair, (5) inflammation, and (6) remodeling of ventricles. Third, ST2, a member of the IL-1 receptor protein family and a biomarker of cardiac stress, has predicted the presence and severity of adverse cardiac remodeling and tissue fibrosis, which occurs in response to (1) MI, (2) ACS, or (3) heart failure. Fourth, N- and C-terminal fragments of IGF-binding protein-4 (NT- and CT-IGFBP-4), which are released from vulnerable atherosclerotic plaques, are the most recent biomarker being explored for major adverse cardiac events (MACE) prediction in patients with suspected myocardial ischemia.

Several biomarkers that were once used in clinical practice are no longer recommended for routine patient care in the evaluation of ACS and heart failure, because they lack tissue specificity or provide no added value to cTn or NP for diagnosis or risk assessment. These biomarkers include: CK, CK-MB, the isoforms of CK-MB, C-reactive protein (CRP), and myoglobin. Numerous other investigational biomarkers, which reflect different pathophysiologic mechanisms of ACS and heart failure, are shown in Fig. 34.4.

AT A GLANCE

Strategies for ruling in and ruling out acute MI with contemporary and high-sensitivity cTn assays.



REVIEW QUESTIONS

1. The technique of choice for measuring cTns is a(n):
   a. photometric assay.
   b. immunoassay.
   c. potentiometric assay.
   d. amperometric assay.
2. What are the names of the contractile proteins that are located in the striated muscle fibers of the heart?
   a. Natriuretic peptides
   b. Modified albumins
   c. Cardiac troponins
   d. Actin and myosin
3. Which one of the following cardiac biomarkers is important in detecting moderate to severe CHF?
   a. B-type natriuretic peptide
   b. Cardiac troponin
   c. Nourin
   d. ST2
4. High-sensitivity cardiac troponin assays are characterized by the following except
   a. Ability to measure >50% of normal subjects
   b. Maximum acceptable %CV at the 99th percentile is 20%
   c. Concentrations units express as whole numbers by ng/L
   d. Sex-specific 99th percentiles are used
5. A protein that is involved in a number of disease processes including cancer, inflammation, and heart disease and that is expressed in the process of cardiac ventricular remodeling is:
   a. isoprostane.
   b. phospholipase A_2.
   c. galectin-3.
   d. MPO.
6. The most common cause of ACS is:
   a. myocardial ischemia due to incomplete occlusion of a coronary artery.
   b. atherosclerosis in coronary arteries.
   c. reduced outflow of blood from the left side of the heart.
   d. decreased output of the right side of the heart due to cardiac valve destruction.

7. Which one of the following is not a requirement for an ideal cardiac biomarker?
   a. Specificity
   b. Rapid release from the heart into circulation
   c. Rapid removal from circulation
   d. Sensitivity
8. Loss of oxygen in the form of arterial blood supply to an area of cardiac tissue is referred to as:
   a. ischemia.
   b. infarct.
   c. necrosis.
   d. atherosclerosis.
9. The condition resulting from ineffective pumping of blood to the body's tissues is referred to as:
   a. coronary artery syndrome.
   b. AMI.
   c. stable angina.
   d. CHF.

SUGGESTED READINGS

Apple, F. S., Jaffe, A. S., Collinson, P., et al. (2015). on behalf of the IFCC Task Force on Clinical Applications of Cardiac Bio-Markers. IFCC educational materials on selected analytical and clinical applications of high-sensitivity cardiac troponin assays. *Clinical Biochemistry*, *48*, 201–203.

Apple, F. S., Sandoval, Y., Jaffe, A. S., Ordonez-Llanos, J., for the IFCC Task Force on Clinical Applications of Cardiac Bio-Markers. (2017). Cardiac troponin assays: Guide to understanding analytical characteristics and their impact on clinical care. *Clinical Chemistry*, *63*, 73–81.

de Lemos, J. A., Drazner, M. H., Omland, T., et al. (2010). Association of troponin T detected with a highly sensitive assay and cardiac structure and mortality risk in general population. *JAMA*, *304*, 2503–2512.

Eggers, K. N. M., & Lindhal, B. (2017). Impact of sex on cardiac troponin concentration – a critical appraisal. *Clinical Chemistry*, *63*, 1457–1464.

Jaffe, A. S., Moeckel, M., Giannitsis, E., et al. (2014). In search for the Holy Grail: Suggestions for studies to define delta changes to diagnose or exclude acute myocardial infarction: A position paper from the study group on biomarkers of the Acute Cardiovascular Care Association. *European Heart Journal Acute Cardiovascular Care*, 313–316.

Jaffe, A. S., & White, H. D. (2016). Ruling-in myocardial injury and ruling-out myocardial infarction with the European Society of Cardiology (ESC) 1-hour algorithm. *Circulation*, *134*, 1542–1545.

Januzzi, J. L., van Kimmenade, R., Lainchbury, J., et al. (2006). NT-proBNP testing for diagnosis and short-term prognosis in acute destabilized heart failure: An international pooled analysis of 1256 patients: The International Collaborative of NT-proBNP Study. *European Heart Journal*, *27*, 330–337.

Morrow, D. A., Cannon, C. P., Jesse, R. L., et al. (2007). National Academy of Clinical Biochemistry Laboratory Medicine practice guidelines: Clinical characteristics and utilization of biochemical markers in acute coronary syndromes. *Clinical Chemistry*, *53*, 552–574.

Morrow, D. A. (2017). Clinician's guide to early rule-out strategies with high-sensitivity cardiac troponin. *Circulation*, *13*, 1612–1616.

Pickering, J. W., Than, M. P., Cullen, L., et al. (2017). Rapid rule-out of acute myocardial infarction with a single high-sensitivity cardiac troponin T measurement below the limit of detection. *Annals of Internal Medicine*. https://doi.org/10.7326/M16-2562.

Saenger, A. K., Rodriguez-Fraga, O., et al. (2017). Specificity of B-type natriuretic peptide assays: Cross-reactivity with different BNP, NT-proBNP, and proBNP peptides. *Clinical Chemistry*, *63*, 351–358.

Sandoval, Y., Smith, S. W., & Apple, F. S. (2016). Present and future of high sensitivity cardiac troponin in clinical practice: A paradigm shift to high sensitivity assays. *American Journal of Medicine*, *129*, 354–365.

Sandoval, Y., & Thygesen, K. (2017). Myocardial infarction type 2 and myocardial injury. *Clinical Chemistry*, *63*, 101–107.

Sandoval, Y., Smith, S. W., Thordsen, S. E., et al. (2017). Diagnostic performance of high sensitivity compared to contemporary cardiac troponin I for the diagnosis of acute myocardial infarction. *Clinical Chemistry*, *63*, 1594–1604.

Sandoval, Y., Smith, S. W., Love, S. A., et al. (2017). Single high-sensitivity cardiac troponin I to rule out myocardial infarction. *American Journal of Medicine*, *130*, 1076–1083.

Schellenberger, U., O'Rear, J., Guzzetta, A., et al. (2006). The precursor to B-type natriuretic peptide is an O-linked glycoprotein. *Archives of Biochemistry and Biophysics*, *451*, 160–166.

Shah, A., Griffiths, M., Lee, K. K., et al. (2015). High-sensitivity cardiac troponin and the under diagnosis of myocardial infarction in women. *British Medical Journal*, *350*, 7873.

Thygesen, K., Alpert, J. S., Jaffe, A. S., et al. (2018). Fourth universal definition of myocardial infarction. *Journal of the American College of Cardiology*. https://doi.org/10.1016/jacc2018.08.1038.

Vasile, V. C., & Jaffe, A. S. (2017). Natriuretic peptides and analytical barriers. *Clinical Chemistry*, *63*, 50–58.

Wu, A. H. B., Christenson, R. H., Greene, D. N., et al. (2018). Clinical laboratory practice recommendations for the use of cardiac troponin in acute coronary syndromes: Expert opinion from the Academy of the American Association for Clinical Chemistry and the Task Force on Clinical Applications of Cardiac Bio-Markers of the International Federation of Clinical Chemistry. *Clinical Chemistry*, *64*, 645–655.

35

Kidney Disease

*Michael P. Delaney, Edmund J. Lamb**

OBJECTIVES

1. Define the following:
   - Bence-Jones protein
   - Clearance
   - Cystatin-C
   - Diabetes insipidus (DI)
   - Diabetic nephropathy
   - Dialysis
   - Diuretic
   - Glomerulus
   - Glomerular filtration rate (GFR)
   - Juxtaglomerular apparatus
   - Micturition
   - Nephrolithiasis
   - Nephron
   - Pyelonephritis
   - Renal pelvis
   - Renal replacement therapy (RRT)
   - Ureter
   - Urine
2. Describe the following components of the renal system, including structure, function, contribution to urine formation, and clinical significance:
   - Bladder
   - Bowman capsule
   - Collecting tubules
   - Distal convoluted tubule
   - Glomerulus
   - Juxtaglomerular apparatus
   - Kidneys
   - Loop of Henle
   - Nephron
   - Proximal convoluted tubule
3. State the significance of the kidneys' blood supply in the formation of urine within the structure of the glomerulus; list the arteries that form the glomerular capillary bed.
4. List and describe the three major functions of the kidneys, including processes involved, laboratory analytes affected in each process, and how each function is controlled.
5. Describe the glomerular filtration rate (GFR) and clearance in terms of clinical usefulness and effects of renal disease and age on each process.
6. Describe the renal handling of electrolytes and water, including normal physiology, hormones involved, and effect of disease on each function.
7. Discuss the following conditions including the causes (such as, drugs and toxins), symptoms, acid-base disturbances (if any), laboratory methods used to assess, and pertinent laboratory results obtained:
   - Acute kidney injury (AKI)
   - Acute nephritic syndrome
   - Acute tubular necrosis (ATN)
   - Chronic kidney disease (CKD)
   - Diabetes insipidus (DI)
   - End-stage renal disease (ESRD) (including uremic syndrome)
   - Glomerular disease
   - Interstitial nephritis
   - Nephrotic syndrome
   - Pyelonephritis
   - Renal tubular acidoses (RTAs)
   - Uremic syndrome
   - Urinary tract obstruction
8. Describe two options for renal replacement therapy (RRT), including the clinical need for the therapy and laboratory assessment of each therapy.
9. Evaluate and analyze case studies related to renal disease and laboratory analysis of renal disease.

* We are grateful for the data supplied by the US Renal Data System (USRDS). The interpretation and reporting of these data are the responsibility of the authors and in no way should be seen as an official policy or interpretation of the US government. We are also grateful for data supplied by the UK Renal Registry. The interpretation and reporting of these data are the responsibility of the authors and in no way should be seen as an official policy or interpretation of the UK Renal Registry. We also acknowledge the input of Professor Christopher Price to previous editions of this chapter.

KEY WORDS AND DEFINITIONS

Acidosis Accumulation of acid and hydrogen ions or depletion of the alkaline reserve (bicarbonate content) in the blood and body tissues.

Acute kidney injury (AKI) A rapid decline in kidney function that occurs over hours and days.

Acute nephritic syndrome The sudden onset of hematuria, proteinuria, diminished urine production, azotemia, hypertension, and edema.

Acute tubular necrosis (ATN) Acute renal failure with mild to severe damage or necrosis of tubule cells.

Antidiuretic hormone (ADH; vasopressin) A nonapeptide hormone formed by the neuronal cells of the hypothalamic nuclei and stored in the posterior lobe of the pituitary gland (neurohypophysis). It has both antidiuretic and vasopressor actions.

Azotemia An excess of urea or other nitrogenous compounds in the blood.

Bartter syndrome Hypertrophy and hyperplasia of the juxtaglomerular cells, producing hypokalemic alkalosis and hyperaldosteronism.

Bence-Jones protein An abnormal plasma or urinary protein consisting of monoclonal immunoglobulin light chains, which is excreted in some neoplastic diseases. It is a characteristic protein found in the urine of most patients with multiple myeloma.

Bowman capsule The double-walled globular kidney structure that forms the beginning of a renal tubule and surrounds the glomerulus.

Chronic kidney disease (CKD) Abnormalities of kidney structure or function, present for greater than 3 months, with implications for health.

Clearance The volume of plasma from which a given substance is completely cleared by the kidneys per unit of time.

Diabetes insipidus (DI) A diabetic (defined as the excessive production of urine) disorder due either to insufficient synthesis of antidiuretic hormone (ADH), defective ADH receptors, or end-organ resistance to its action. This results in failure of tubular reabsorption of water in the kidney.

Diabetic nephropathy The nephropathy that commonly accompanies later stages of diabetes mellitus; it begins with hyperfiltration, renal hypertrophy, albuminuria, and hypertension.

Dialysis The removal—for example, by means of a hemodialysis (HD) machine or filter—of certain elements from the blood by virtue of differences in the rates of their diffusion through a semipermeable membrane.

End-stage renal disease (ESRD) A condition where renal function is no longer adequate to support life.

Erythropoietin A glycoprotein hormone secreted chiefly by the kidney in the adult; it increases the production of red blood cells.

Fanconi syndrome A rare recessive disorder with a poor prognosis, characterized by pancytopenia, bone marrow hypoplasia, and patchy brown skin as well as multiple congenital anomalies of the musculoskeletal and genitourinary systems.

Gitelman syndrome A syndrome of hypertrophy of juxtaglomerular cells similar to Bartter syndrome but with hypocalciuria and hypomagnesemia.

Glomerular filtration rate (GFR) The rate in milliliters per minute at which small molecules are filtered through the kidney's glomeruli. It is a measure of the number of functioning nephrons.

Glomerulonephritis Nephritis accompanied by inflammation of the capillary loops of the glomeruli of the kidney. It occurs in acute, subacute, and chronic forms.

Glomerulus A tuft of blood vessels found in each nephron of the kidney that is involved in the filtration of blood.

Hematuria Blood in the urine.

Hemodiafiltration A type of hemofiltration with a dialytic component added. With it, blood flow is accelerated to twice that of conventional dialysis.

Hypertension A medical condition characterized by high arterial blood pressure.

IgA nephropathy A common, chronic form of glomerulonephritis marked by hematuria and proteinuria and by deposits of immunoglobulin A in the mesangial areas of the renal glomeruli.

Interstitial nephritis Primary or secondary disease of the renal interstitial tissue.

Juxtaglomerular apparatus (JGA) A complex in the kidney whose function is the autoregulation of the glomerular filtration rate.

Liddle syndrome A rare autosomal dominant syndrome resulting from mutations affecting the epithelial sodium channel leading to abnormally increased channel function.

Lithotripsy The crushing of a calculus within the urinary system, followed at once by the washing out of the fragments; it is done either surgically or by several different noninvasive methods.

Loop of Henle The U-shaped part of the renal tubule, extending through the medulla from the end of the proximal convoluted tubule to the beginning of the distal convoluted tubule.

Metabolic acidosis Any of the various kinds of acidosis in which the acid-base status of the body shifts toward the acid side because of loss of base or retention of acids other than carbonic acid.

Nephritis Inflammation of the kidney with focal or diffuse proliferation or destructive processes that may involve the glomerulus, tubule, or interstitial renal tissue.

Nephrolithiasis A condition marked by the presence of renal calculi (stones).

Nephron The anatomic and functional unit of the kidney, consisting of the (1) renal corpuscle, (2) proximal convoluted tubule, (3) descending and ascending limbs of loop of Henle, (4) distal convoluted tubule, and (5) collecting tubule.

Nephrotic syndrome General name for a group of diseases involving defective kidney glomeruli, characterized by massive proteinuria and lipiduria with varying degrees of edema, hypoalbuminemia, and hyperlipidemia.

Obstructive uropathy Uropathy resulting from an obstruction in the urinary tract.

Oliguria Diminished urine production and excretion.

Peritoneal dialysis (PD) Diffusion of solutes and convection of fluid through the peritoneal membrane. The dialyzing solution is introduced into and removed from the peritoneal cavity as either a continuous or intermittent procedure.

Polycystic kidney disease The most common renal cystic condition, with concomitant deterioration of renal function.

Polyuria The passage of a large volume of urine in a given period.

Proteinuria Excessive serum proteins in the urine, as in renal disease, after strenuous exercise, and with dehydration.

Pseudohypoaldosteronism type 1 A hereditary disorder of infancy characterized by severe salt wasting, failure to thrive, and other signs of aldosterone deficiency.

Pyelonephritis An inflammation of the kidney and its pelvis as a result of infection.

Rapidly progressive glomerulonephritis (RPGN) Acute glomerulonephritis marked by a rapid progression to end-stage renal disease.

Renal replacement therapy (RRT) Any treatment that replaces kidney function, including dialysis and transplantation.

Renal tubular acidosis (RTA) A variety of metabolic acidosis resulting from the impairment of renal function.

Renin An enzyme of the hydrolase class that catalyzes cleavage of the leucine-leucine bond in angiotensinogen to generate angiotensin I.

Toxic nephropathy Kidney damage caused by the effects of a nephrotoxin.

Uremia An excess in the blood of urea, creatinine, and other nitrogenous end products of protein and amino acid metabolism; also referred to as azotemia.

Uremic syndrome The spectrum of symptoms accompanying uremia.

The kidneys play a central role in homeostasis, and reduced renal function strongly correlates with increasing morbidity and mortality. The basic anatomy and physiology of the kidneys are described as a foundation for understanding the pathophysiology of disease and the rationale for diagnostic and management strategies in kidney disease. Key analytic methods employed during the investigation of kidney disease are dealt with in Chapter 21.

ANATOMY

The kidneys are a paired organ system located in the lumbar region. Their function is to (1) filter the blood and excrete the end products of body metabolism in the form of urine; (2) regulate the concentrations of (a) hydrogen, (b) sodium, (c) potassium, (d) phosphate, and (e) other ions in the extracellular fluid; and (3) produce hormones. In an adult, each kidney is about 12 cm long and weighs about 150 g in men and 135 g in women. Through the characteristically bean-shaped kidney pass the vessels, nerves, and ureter (Fig. 35.1).

Nephron

The functional unit of the kidney is the **nephron**. Each kidney may contain up to 1.5 million nephrons. The nephron consists of a (1) glomerulus, (2) proximal tubule, (3) loop of Henle, (4) distal tubule, and (5) collecting duct (Fig. 35.2). The collecting ducts ultimately combine to develop into the renal calyces, where the urine collects before passing along the ureter and into the bladder. The kidney is divided into

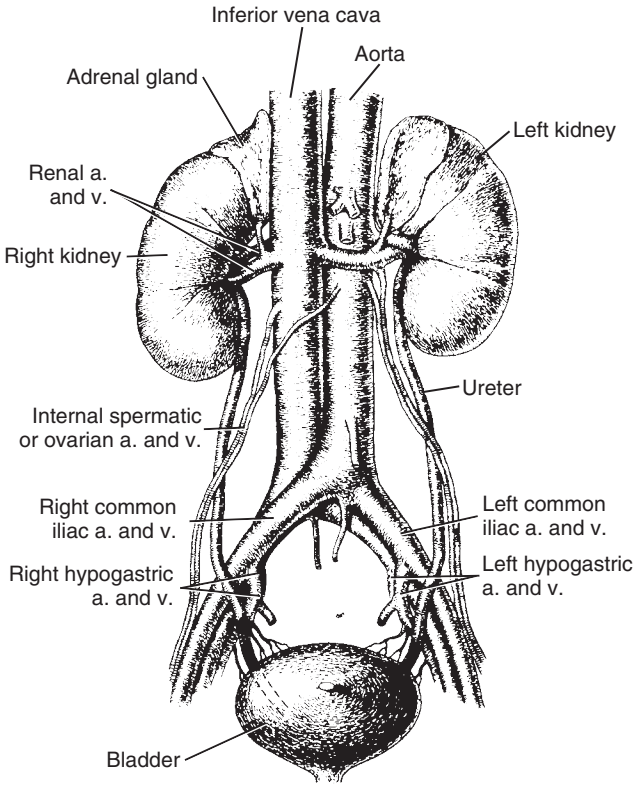


Fig. 35.1 Vascular and anatomic relationships of the kidneys in the human. (From Leaf, A., & Cotran, R. S. [1985]. *Renal pathophysiology* [3rd ed.]. Oxford: Oxford University Press. Reproduced by permission of Oxford University Press.)

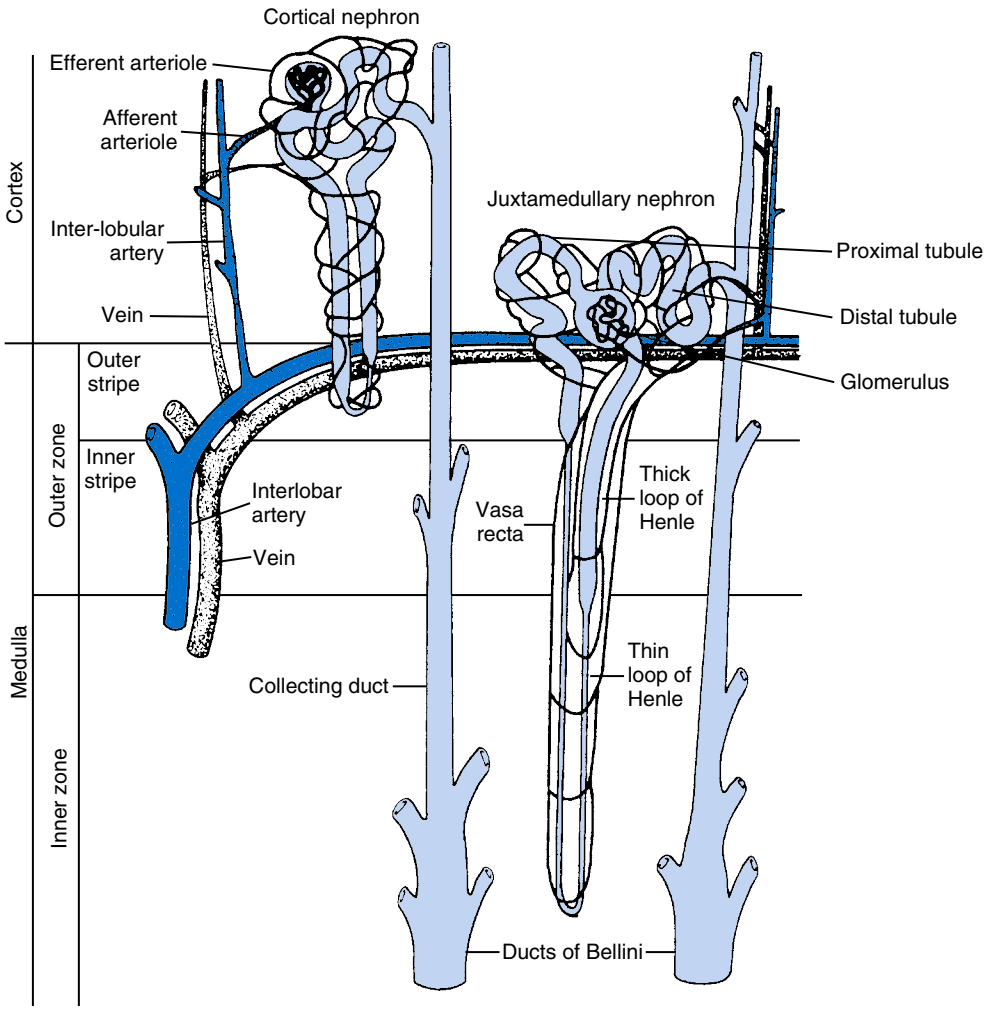


Fig. 35.2 Diagrammatic representation of the nephron, the functional unit of the kidney, illustrating the anatomic and vascular arrangements. (From Pitts, R. F. [1974]. *Physiology of the kidney and body fluids* [3rd ed., p. 8]. Chicago: Year Book Medical Publishers.)

several lobes. The outer, darker region of each lobe—the cortex—consists of the glomeruli and the proximal and distal tubules. The cortex surrounds a paler inner region—the medulla—which is further divided into a number of conical areas known as the renal pyramids, the apex of which extends toward the renal pelvis, forming papillae. Medullary rays are visible striations in the renal pyramids that connect the kidney cortex with the medulla. They are composed of (1) descending (straight proximal) and (2) ascending (straight distal) thick limbs of Henle and (3) collecting ducts and associated blood vessels (the vasa recta). The central hilus is where blood vessels, lymphatics, and the renal pelvis (containing the ureter) join the kidney.

The **glomerulus** is formed from a specialized capillary network. Each capillary develops into approximately 40 glomerular loops around 200 μm in size and consisting of a variety of different cell types supported on a specialized basement membrane. Some endothelial and epithelial cells act in concert with the specialized glomerular basement membrane (GBM) to form the glomerular filtration barrier. The glomerular capillaries are supported by a network of mesangial cells and a mesangial matrix that act as connective tissue for the glomerular apparatus. The basement membrane forms the main size-discriminant barrier to protein passage into the tubular lumen.

The **Bowman capsule** forms the beginning of the tightly coiled proximal convoluted tubule *(pars convoluta)*, which as it progresses toward the renal medulla becomes straightened and is then called the *pars recta.* The proximal tubule is about 15 mm long. It is the most metabolically active part of the nephron, facilitating the reabsorption of 60% to 80% of the glomerular filtrate volume—including 70% of the filtered load of (1) sodium and (2) chloride, most of the (3) potassium, (4) glucose, (5) bicarbonate, (6) phosphate, and (7) sulfate—and secreting 90% of the hydrogen ion excreted by the kidney.

The pars recta drains into the descending thin **loop of Henle**, which, after passing through a hairpin loop, becomes first the ascending thin limb and then the thick ascending loop. At the end of the thick ascending limb is a cluster of cells known as the *macula densa* (Fig. 35.3). The main role of the loop of Henle is to provide the ability to generate a concentrated urine that is hypertonic with respect to plasma.

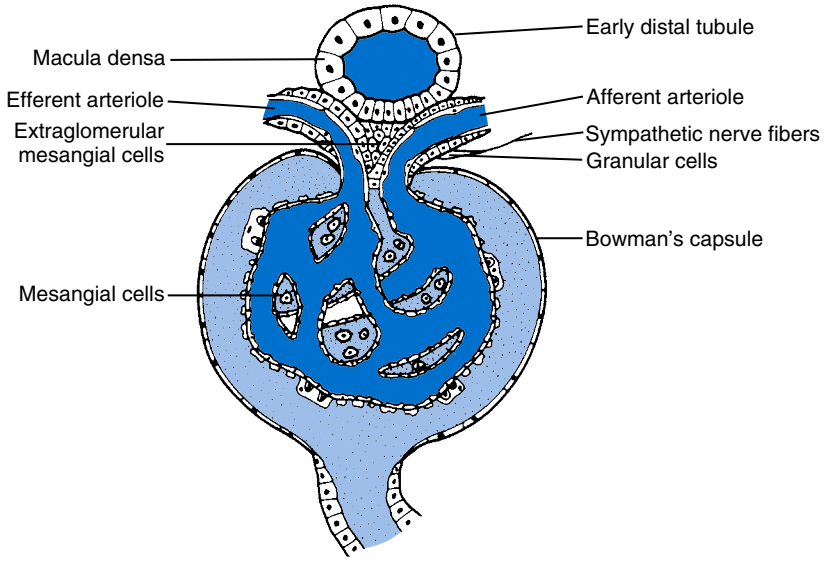


Fig. 35.3 The juxtaglomerular apparatus. The beginning of the distal tubule (i.e., where the loop of Henle reenters the cortex) lies very close to the afferent and efferent arterioles, and the cells of both the afferent arteriole and the tubule show specialization. The cells of the afferent arteriole are thickened, granular (juxtaglomerular) cells innervated by sympathetic nerve fibers. The mesangial cells are irregularly shaped and contain filaments of contractile proteins. Identical cells are found just outside the glomerulus and are termed *extraglomerular mesangial cells* or *Goormaghtigh cells*. (From Lote, C. J. [2000]. *Principles of renal physiology* [4th ed., Chapter 2]. London: Kluwer Academic Publishers.)

The cells forming the distal tubule of the nephron start at the macula densa and extend to the first fusion with other tubules to form the collecting ducts. Sodium chloride reabsorption and some potassium and hydrogen ion excretion occur at this site.

The collecting ducts are formed from approximately six distal tubules. These are successively joined by other tubules to form ducts of Bellini, which ultimately drain into a renal calyx.

Juxtaglomerular Apparatus

Where the thick ascending limb of the loop of Henle passes very close to the glomerulus of its own nephron, the cells of the tubule and the afferent arteriole show regional specialization (see Fig. 35.3). The tubule forms the macula densa; the arteriolar cells are filled with granules (containing renin or its inactive precursor, prorenin) and are innervated with sympathetic nerve fibers. This area is called the **juxtaglomerular apparatus (JGA)**. The JGA plays an important part in maintaining systemic blood pressure through regulation of the circulating intravascular blood volume and sodium concentration. The proteolytic enzyme renin is released primarily in response to decreased afferent arteriolar pressure and decreased intraluminal sodium delivery to the macula densa. Renin release from the macula densa is also influenced by nitric oxide, renal cortical prostaglandins (predominantly PGI_2), and the sympathetic nervous system. The released renin then acts on the plasma protein angiotensinogen to generate angiotensin I. This is converted in the lungs by angiotensin converting enzyme (ACE) to the potent vasoconstrictor and stimulator of aldosterone release, angiotensin II (AII). The vasoconstriction and aldosterone release (with increased distal tubular sodium retention) act in concert with the other action of AII, to increase the release of **antidiuretic hormone (ADH; vasopressin)** and to increase proximal tubular sodium reabsorption, intravascular volume, and pressure. AII also has an inhibitory effect on renin release as part of a negative feedback loop.

Blood Supply

The renal artery divides into posterior and anterior elements, which then divide into the interlobar, arcuate, interlobular, and ultimately afferent arterioles, which expand into the capillary bed that forms the glomerulus (see Fig. 35.2). These capillaries then rejoin to form the efferent arteriole, which then forms the capillary plexuses as well as the elongated vessels (the *vasa recta*) that pass around the remaining parts of the (1) nephron, (2) proximal and distal tubules, (3) loop of Henle, and (4) collecting duct, providing oxygen and nutrients and removing ions, molecules, and water, which are reabsorbed by the nephron. The efferent arteriole then merges with renal venules to form the renal veins, which emerge into the inferior vena cava. The complex architecture of the intrarenal vascular tree is ordered in three dimensions in a characteristic arrangement that probably serves to distribute the blood pressure and flow appropriately to the glomeruli.

In the adult, the kidneys receive approximately 25% of the cardiac output; however, in the newborn infant it is 5%, reaching adult proportions only by the end of the first year of life. About 90% of this blood flow supplies the renal cortex, maintaining the highly active tubular cells. The maintenance of renal blood flow is essential to renal function, and there is a complex array of intrarenal regulatory mechanisms ensuring that it is maintained across a wide range of systemic blood

pressures. The renal glomerular perfusion pressure is independent of the systemic pressure between 90 and 200 mm Hg, being maintained at a constant 45 mm Hg.

KIDNEY FUNCTION

The main biologic functions of the kidneys are (1) excretion, (2) homeostatic regulation, and (3) endocrine function. The kidneys integrate these functions to maintain homeostasis and regulate the internal milieu.

Excretory Function

Urine is (1) excreted by the kidneys, (2) passed through the ureters, (3) stored in the bladder, and (4) discharged through the urethra. In health, it (1) is sterile and clear, (2) is of yellow or amber color, (3) has a slightly acid pH (5.0 to 6.0), and (4) has a characteristic odor. Its specific gravity is about 1.024. In addition to dissolved compounds, it contains a number of cells, proteinaceous casts, and crystals (formed elements). Changes in these formed elements are studied using urine microscopy.

Urination, also termed *micturition,* is the discharge of urine. In normal adults, adequate homeostasis is maintained with a urine output of about 500 mL/day. Alterations in urinary output are described as *anuria* (<100 mL/day), **oliguria** (<400 mL/day), or **polyuria** (>3 L/day or 50 mL/kg body weight per day). The most common disorder of urination is altered frequency, which may be associated with increased urinary volume or partial urinary tract obstruction (e.g., in prostatic hypertrophy).

The first step in urine formation is filtration of plasma water at the glomeruli. A net filtration pressure of about 17 mm Hg in the capillary bed of the tuft drives the filtrate through the glomerular membrane. The filtrate is called an ultrafiltrate because its composition is essentially the same as that of plasma, but with a notable reduction in molecules of molecular weight exceeding 15 kDa. Each nephron produces about 100 μL of ultrafiltrate per day. Overall, approximately 170 to 200 L of ultrafiltrate passes through the glomeruli in 24 hours. In the passage of ultrafiltrate through the tubules, reabsorption of solutes and water in various regions of the tubules reduces the total urine volume, which typically ranges between 0.4 and 2 L/day.

Transport of solutes and water occurs both across and between the epithelial cells that line the renal tubules. Transport is both active (energy-requiring) and passive, but many of the so-called passive transport processes are dependent on or secondary to active transport processes, particularly those involving sodium transport. All known transport processes involve receptor or mediator molecules, many of which have now been identified and characterized using molecular biology techniques. The activity of many of these molecules is regulated by phosphorylation facilitated by protein kinase C or A. Their renal distribution has been shown to correlate with the known regional functional activities. There are inherited disorders of specific tubular transporters and a well-known generalized disorder affecting all of the transport processes, causing **Fanconi syndrome**.

Direct coupling of adenosine triphosphate (ATP) hydrolysis is an example of an active transport process. The most important of these in the nephron is Na^+,K^+-ATPase, which is located on the basolateral membranes of the tubuloepithelial cells. This enzymatic transporter accounts for much of renal oxygen consumption and drives more than 99% of renal sodium reabsorption.

Renal epithelial cell membranes also contain proteins that act as ion channels. For example, there is one for sodium that is closed by amiloride and modulated by hormones such as atrial natriuretic peptide (ANP). Ion channels enable much faster rates of transport than ATPases, but are relatively fewer in number—approximately 100 sodium and chloride channels as against 10^7 Na^+,K^+-ATPase molecules per cell.

In the tubules, the solute composition of the ultrafiltrate is altered by the processes of reabsorption and secretion, so that the urine excreted may have a very different composition from that of the original filtrate. Different regions of the tubule have been shown to specialize in certain functions. In the proximal tubule, 60% to 80% of the ultrafiltrate is reabsorbed in an obligatory fashion, along with (1) sodium, (2) chloride, (3) bicarbonate, (4) calcium, (5) phosphate, (6) sulfate, and (7) other ions. Glucose is virtually completely reabsorbed, predominantly in the proximal tubule by a passive but sodium-dependent process that is saturated at a blood glucose concentration of about 180 mg/dL (10 mmol/L). Uric acid is also reabsorbed in the proximal tubule by a passive sodium-dependent mechanism, but there is also an active secretory mechanism.

In the loops of Henle, chloride and more sodium without water are reabsorbed, generating dilute urine. Water reabsorption in the more distal tubules and collecting ducts is then regulated by ADH. In the distal tubule, secretion is the prominent activity; organic ions, potassium ions, and hydrogen ions are transported from the blood in the efferent arteriole into the tubular fluid. It is also this region that secretes hydrogen ions and reabsorbs sodium and bicarbonate to aid in acid-base regulation. Paracellular (between cell) movement is driven predominantly by concentration, osmotic, or electrical gradients.

Regulatory Function

The *regulatory function* of the kidneys has a major role in homeostasis. The mechanisms of differential reabsorption and secretion, located in the tubule of a nephron, are the effectors of regulation. These mechanisms operate under a complex system of control in which both extrarenal and intrarenal humoral factors participate.

Electrolyte Homeostasis

The proximal convoluted tubule is predominantly concerned with reabsorption (Fig. 35.4). Water reabsorption in the proximal convoluted tubule is termed "obligatory" because its volume is related to the heavy load of solutes being returned to the blood in the efferent arteriole. The amount of bicarbonate reabsorption is related to the **glomerular filtration rate (GFR)** and the hydrogen ion secretory rate. The amount of phosphate reabsorption is controlled in part by plasma

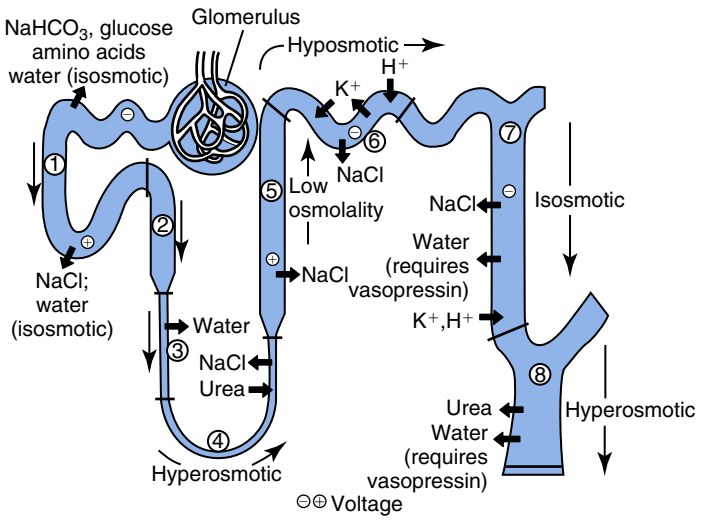


Fig. 35.4 Countercurrent multiplication mechanism: schematic representation of the principal processes of transport in the nephron. In the convoluted portion of the proximal tubule *(1)*, salts and water are reabsorbed at high rates in isotonic proportions. Bulk reabsorption of most of the filtrate (65% to 70%) and virtually complete reabsorption of glucose, amino acids, and bicarbonate take place in this segment. In the pars recta *(2)*, organic acids are secreted and continuous reabsorption of sodium chloride takes place. The loop of Henle comprises three segments: the thin descending *(3)* and ascending *(4)* limbs and the thick ascending limb *(5)*. The fluid becomes hyperosmotic because of water abstraction as it flows toward the bend of the loop, and it becomes hyposmotic because of sodium chloride reabsorption as it flows toward the distal convoluted tubule *(6)*. Active sodium reabsorption occurs in the distal convoluted tubule and in the cortical collecting tubule *(7)*. This last segment is water-impermeable in the absence of vasopressin (antidiuretic hormone), and the reabsorption of sodium in this segment is increased by aldosterone. The collecting duct *(8)* allows equilibration of water with the hyperosmotic interstitium when vasopressin is present. For additional details, see text. (From Burg, M. B. [1978]. The nephron in transport of sodium, amino acids, and glucose. *Hospital Practice, 13*, 100. Modified from a drawing by A. Iselin.)

calcium concentration and in part by the effect of parathyroid hormone (PTH) on the tubular cells. Normally the high-threshold substances—glucose and, to a great extent, amino acids—are reabsorbed here by means of specific intracellular active transport systems. Uric acid may be either reabsorbed or secreted in the proximal convoluted tubule by a two-way carrier-mediated process.

In the ascending loop of Henle, 20% to 25% of filtered sodium is reabsorbed without concomitant reabsorption of water. This process generates dilute urine with an osmolality of 100 to 150 mOsm/kg of water and helps establish the corticomedullary osmotic gradient. The resulting hypertonicity of the interstitium is important in the pathogenesis of renal infections because the hypertonic environment interferes with leukocyte function. Subsequent water reabsorption is regulated by ADH. Although the reabsorption of Na^+ in the loop of Henle is complex and incompletely understood, at least one mechanism consists of an active Cl^- pump with subsequent reabsorption of Na^+ along an electrochemical gradient. This mechanism is inhibited by loop diuretics.

The distal tubule is functionally the most active region of the nephron for the homeostatic regulation of plasma electrolytes and plasma acid-base concentrations. Here a combination of secretion and reabsorption takes place between Na^+, K^+, and H^+. Although excess plasma hydrogen ions are secreted all along the tubule, it is in the distal tubule that exchange of H^+ for Na^+ (which is reabsorbed) fine tunes the balance between H^+ loss and retention (see Chapter 36). Potassium ions are also secreted in the distal tubule. Aldosterone is a potent modulator of Na^+ reabsorption in the distal tubule, particularly when the need arises to conserve Na^+. Production of aldosterone in the adrenal cortex is stimulated by the renin-angiotensin system and by high plasma potassium concentration. Renal secretion of renin is complex but is at least partly regulated by renal perfusion and plasma sodium concentration. Both inadequate perfusion and a low concentration of plasma sodium stimulate renin secretion. Organic anions, such as acetoacetate and β-hydroxybutyrate, also consume H^+ as they are eliminated in part in their nondissociated acid form. When H^+ must be conserved to maintain blood pH, distal tubule cells reduce the secretion of H^+, reduce NH_4^+ generation, reduce Na^+-H^+ exchange, and increase bicarbonate excretion. The net effect is a reduction in plasma bicarbonate and restoration of normal blood pH.

Water Homeostasis

The production of glomerular filtrate normally amounts to about 180 L/day. The unique physiology of the kidney enables approximately 99% of this to be reabsorbed in the production of urine with variable osmolality: between 50 and 1400 mOsmol/kg H_2O at extremes of water intake. Different segments of the nephron show differing permeability to water, enabling the body to both retain water and produce urine of variable concentration. Approximately 70% of the water content of the tubular fluid is reabsorbed in the proximal tubule, 5% in the loop of Henle, 10% in the distal tubule, and the remainder in the collecting ducts.

Water reabsorption occurs both isosmotically, in association with electrolyte reabsorption in the proximal tubule, and differentially, in the loop of Henle, distal tubule, and collecting duct in response to the action of the nonapeptide ADH. Absorption of water depends on the driving force for water reabsorption (predominantly active sodium transport) and the osmotic equilibration of water across the tubular epithelium. The generation of concentrated urine depends upon medullary hyperosmolality, which in turn requires low water permeability in some kidney segments (ascending limb of the loop of Henle), whereas in other kidney segments there is a requirement for high water permeability: differing permeability and the enaction of hormonal control is largely caused by the differential expression along the nephron of a family of proteins known as the aquaporins (AQP), which act as water channels.

Urinary concentration is predominantly achieved by countercurrent multiplication in the loop of Henle (see Fig. 35.4). Although the descending thin limb is very permeable to water, the ascending limb and the collecting duct are not (the collecting ducts are also poorly permeable to urea). The fluid entering the loop of Henle is isotonic to plasma but is hypotonic on leaving it. The ascending limb has active sodium reabsorption driven by Na^+,K^+-ATPase with electroneutralizing transport of chloride, a combined process that can be inhibited by the so-called loop diuretics (e.g., furosemide). In this section of the nephron, sodium reabsorption is not accompanied by water, creating a hypertonic medullary interstitium and facilitating water reabsorption from the anatomically adjacent descending limb. The descending limb cells are permeable to sodium chloride, which is cycled from the descending limb back to the ascending limb. The continuous flow along the loop generates an osmotic gradient at the tip of the loop that can reach 1400 mOsmol/kg per H_2O. Approximately 5% of water is reabsorbed in the loop of Henle.

A further 10% of water reabsorption occurs in the distal tubule, with the remainder (more than 20 L/day) being reabsorbed in the collecting ducts. Entry of water into the collecting duct cells occurs via apical AQP2 channels under the influence of ADH. ADH also increases the permeability of collecting duct cells to urea, which is the major osmotically active component of the luminal fluid in the distal tubule. Fluid of high urea concentration therefore enters the deepest layers of the medullary interstitium, passing down its concentration gradient, contributing to medullary hyperosmolality. The regulation of ADH excretion is of vital importance to fluid homeostasis. The normal plasma osmolality is maintained very tightly between 280 and 290 mOsmol/kg H_2O and is regulated by means of specific osmoreceptors found in the anterior hypothalamus. These receptors modulate the release of ADH and also affect thirst. ADH release may also be stimulated by hypotension, hypovolemia, and vomiting independently of osmoregulation.

Endocrine Function

The *endocrine functions* of the kidneys may be regarded as (1) primary, because the kidneys are endocrine organs producing hormones, or (2) secondary, because the kidneys are a site of action for hormones produced or activated elsewhere. In addition, the kidneys are a site of degradation for hormones such as insulin and aldosterone. In their primary endocrine function, the kidneys produce (1) erythropoietin, (2) prostaglandins and thromboxanes, (3) renin, and (4) 1,25(OH_2) vitamin D_3.

Erythropoietin

Erythropoietin (EPO) is a glycoprotein hormone secreted chiefly by the kidney in the adult and by the liver in the fetus that acts on the bone marrow cells to stimulate erythropoiesis. It is an α-globulin having a molecular weight of 38 kDa. Physiologically the kidneys sense a reduction in O_2 delivery to tissue by blood and release EPO, thereby stimulating the bone marrow to make more red blood cells (RBCs). Conversely, with a surplus of O_2 in blood traversing the kidneys, as in some forms of polycythemia, the release of EPO into blood is diminished. The use of recombinant human EPO (rhEPO, Epoetin) in the management of anemia of kidney disease is discussed further on.

Prostaglandins and Thromboxanes

Prostaglandins and thromboxanes are synthesized from arachidonic acid by the cyclooxygenase (COX) enzyme system (see Chapter 23). This system is present in many parts of the kidneys. The predominant metabolite of its vascular endothelial activity is prostacyclin (PGI_2). Prostaglandin E_2 (PGE_2) appears to be the major metabolite of mesangial and tubular cells. The production and activity of these biologically active compounds have an important role in regulating the physiological action of other hormones on renal vascular tone, mesangial contractility, and tubular processing of salt and water.

Renin

Renin is produced within juxtaglomerular cells after the processing and cleavage of prorenin, which is produced in the liver. Increased production of renin results in formation of AII in the lungs, which is a powerful intrarenal vasoconstrictor and also a key stimulus of aldosterone release from zona glomerulosa cells in the adrenal glands. The net effect is (1) systemic vasoconstriction, (2) intrarenal vasoconstriction, and (3) increased aldosterone release. Aldosterone controls salt and water balance in the kidney. Its effect is predominantly on the distal tubular network, affecting an increase in sodium reabsorption in exchange for potassium.

1,25(OH_2) Vitamin D_3

The kidneys are primarily responsible for producing 1,25(OH_2) vitamin D_3 from 25-hydroxycholecalciferol as a result of the action of the enzyme 25-hydroxycholecalciferol 1α-hydroxylase found in proximal tubular epithelial cells. The regulation of this system is discussed in Chapter 39.

KIDNEY PHYSIOLOGY

The (1) GFR, (2) renal blood flow, and (3) glomerular permeability are important physiologic components of renal function.

Glomerular Filtration Rate

The GFR is considered to be the most reliable measure of the functional capacity of the kidneys and is often thought of as indicative of the number of functioning nephrons. As a physiologic measurement, it has proved to be a useful marker of changes in overall renal function.

The rate of formation of glomerular filtrate depends on the balance between hydrostatic and oncotic forces along the afferent arteriole and across the glomerular filter. The net pressure difference must be sufficient not only to drive filtration across the glomerular filtration barrier but also to drive the ultrafiltrate along the tubules against their inherent resistance to flow. In the absence of sufficient pressure, the lumina of the tubules will collapse.

A decrease in GFR precedes kidney failure in all forms of progressive disease. Different pathologic kidney conditions have been known to progress to end-stage renal disease (ESRD) and dialysis dependency at rates varying from weeks to several decades. The symptoms accompanying progressive kidney disease and their correlation with falling GFR will be influenced by this rate of progression. Measuring GFR in established disease is useful in (1) targeting treatment, (2) monitoring progression, and (3) predicting when **renal replacement therapy (RRT)** will be required. It is also used as a guide to dosage of drugs excreted by the kidneys to prevent potential drug toxicity (see Chapter 30). Measurement of GFR is discussed in Chapter 21.

Glomerular Filtration Rate and Age

Kidney function is not constant throughout life. In utero, urine is produced by the developing fetus from about the ninth week of gestation. The GFR at birth is approximately 30 mL/min per 1.73 m^2. It increases rapidly during the first weeks of life to reach approximately 70 mL/min per 1.73 m^2 by age 16 days, with adult values of body surface area (BSA)–corrected GFR being achieved by 2 years of age. On average, GFR declines with age by approximately 1 mL/min per 1.73 m^2 per year over the age of 40 years, and the rate of decline in GFR accelerates after age 65.

Glomerular Permeability, Filtration, and Protein Loss

The glomerular permeability and filtration capabilities of the kidney control the amount of protein lost in the urine.

Glomerular Permeability and Filtration

The glomerulus acts as a selective filter of the blood passing through its capillaries. The combination of (1) the fenestrated endothelial layer, (2) a basement membrane rich in negatively charged proteoglycans, and (3) a highly specialized terminally differentiated epithelial cell barrier produces a filter that restricts the passage of macromolecules in a (1) size-, (2) charge-, and (3) shape-dependent manner. The epithelial cells (podocytes) have foot processes that are connected to the GBM and form the final barrier to filtration via interdigitating with neighboring cell foot processes connected by a slit diaphragm. Examples of the relationships between size, charge, and mass of the major urinary proteins and their glomerular handling are listed in Table 21.1. In general, proteins of molecular weight greater than that of albumin (66 kDa, diameter 3.5 nm) are retained by the healthy glomerulus and termed high-molecular-weight proteins. However, lower-molecular-weight proteins are also retained to a significant extent.

Urinary Protein Loss

Increased urinary protein loss (**proteinuria**) results from (1) any increase in the filtered load, (2) an increased circulating concentration of low-molecular-weight proteins, or (3) a decrease in reabsorptive capacity. Historically, the pattern of urinary protein loss has been used to identify the cause and classify the proteinuria, of which there are three main types: (1) glomerular, (2) overflow, and (3) tubular proteinuria.

The normal urinary total protein loss is less than 150 mg/day. The proteins lost are made up of mostly albumin (50% to 60%) and some smaller proteins together with proteins secreted by the tubules, of which Tamm-Horsfall glycoprotein (THG, uromodulin) is one. The normal concentrations of proteins found in urine are listed in Table 21.1. Investigation for increased urinary protein loss is mandatory in any patient with suspected kidney disease and is considered in Chapter 21.

Consequences of Proteinuria

Many physicians believe that proteinuria is not just a consequence of the progression of kidney disease but contributes directly to it. The accumulation of proteins in abnormal amounts in the tubular lumen may trigger an inflammatory reaction, which in turn may contribute to interstitial structural damage and expansion and thus to the progression of kidney disease. Evidence gathered from in vitro studies suggests that glomerular filtration of an abnormal amount or types of protein induces mesangial cell injury, leading to glomerulosclerosis, and that these same proteins also have adverse effects on the function of proximal tubular cells. Numerous studies have demonstrated that proteinuria is a potent risk marker for the progression of renal disease in both nondiabetic and diabetic kidney disease. Furthermore, the reduction of protein excretion slows the rate of progression of proteinuric kidney disease. This has been observed in clinical trials in patients treated with ACE inhibitors and angiotensin II receptor blockers (ARBs) either alone or in combination. These drugs reduce protein loss by reducing intraglomerular filtration pressure and possibly by stabilizing the glomerular epithelial cell slit diaphragm proteins. Consequently the reduction of proteinuria is an important therapeutic target.

PATHOPHYSIOLOGY OF KIDNEY DISEASE

Despite the diverse initial causes of injury to the kidney, the progression of kidney disease leading to loss of function and ultimately to kidney failure is a remarkably monotonous process characterized by (1) early inflammation, (2) accumulation and deposition of extracellular matrix, (3) tubulointerstitial fibrosis, (4) tubular atrophy, and (5) glomerulosclerosis (scarring). Proteinuria is thought to be one of

the most important risk factors for the progression of kidney diseases. The renin-angiotensin-aldosterone system (RAAS) plays a pivotal role in many of the pathophysiologic changes that cause kidney injury and is an important therapeutic target.

The kidneys have considerable ability to increase their functional capacity in response to injury. Thus a significant reduction in functioning renal mass (50% to 60%) may occur before the onset of any significant symptoms or even before any major biochemical alterations appear. The most sensitive and specific measure of functional change—the GFR—can be reduced to less than 60 mL/min per 1.73 m^2 before signs and symptoms of kidney failure will be observed. An increase in workload per nephron is thought to be an important cause of progressive renal injury.

Overview of Kidney Disease and Its Clinical Manifestations

Most often kidney disease is detected opportunistically by measurement of blood pressure and urine and blood testing in asymptomatic individuals. Such testing can occur in the primary care setting or for health clearance purposes for insurance. Typical findings include isolated **hematuria** and isolated proteinuria. Kidney disease may also present with macroscopic hematuria, swollen ankles, headaches, and visual disturbances due to severe **hypertension**, or as a manifestation of systemic disease. Symptoms suggestive of advanced kidney disease include fatigue, nausea, vomiting, poor appetite, shortness of breath, fluid retention, poor memory, loss of libido, and itching. Unfortunately many individuals present very late in their disease and may require urgent **dialysis** without having had any previous experience with the specialist nephrology service. These patients have a poor prognosis compared with patients who have been cared for in a multidisciplinary specialist environment for at least 1 year. Therefore early recognition of kidney disease is of paramount importance to outcome.

Effective management of the patient with kidney disease is dependent on establishing a definitive diagnosis. Initial management includes a (1) detailed clinical history, (2) clinical examination, and (3) urinalysis and assessment of the urinary sediment.

Terminology of Kidney Disease

The commonly used term *acute renal failure* (ARF) has been replaced by **acute kidney injury (AKI)**. Kidney *failure* is defined as a GFR of less than 15 mL/min per 1.73 m^2. Not all patients with kidney failure require RRT (dialysis or transplantation) to sustain life. **End-stage renal disease (ESRD)** is a US government–defined term that indicates the need for long-term chronic RRT. Each patient with ESRD is registered through the Medical Evidence Form (2728) that all dialysis and transplant providers must submit. This now includes both Medicare and non-Medicare populations.

AT A GLANCE

Who Should be Tested for Chronic Kidney Disease (CKD)?[a]

Offer testing for CKD using estimated GFR and a urinary albumin-to-creatinine ratio (ACR) to people with any of the following risk factors:

- Diabetes mellitus
- Hypertension
- Acute kidney injury
- Cardiovascular disease (ischemic heart disease, chronic heart failure, peripheral vascular disease, or cerebral vascular disease)
- Structural renal tract disease, recurrent renal calculi, or prostatic hypertrophy
- Multisystem diseases with potential kidney involvement—e.g., systemic lupus erythematosus
- Family history of kidney failure (GFR category G5) or hereditary kidney disease
- Opportunistic detection of hematuria

Do not use age, gender, or ethnicity as risk markers to test people for CKD. In the absence of metabolic syndrome, diabetes mellitus, or hypertension, do not use obesity alone as a risk marker to test people for CKD.

Monitor GFR at least annually in people taking drugs known to be nephrotoxic, such as calcineurin inhibitors (e.g., cyclosporin or tacrolimus), lithium, and nonsteroidal antiinflammatory drugs.

[a]Modified from National Institute for Health and Care Excellence. (2014). *Chronic kidney disease. Early identification and management of chronic kidney disease in adults in primary and secondary care.* http://www.nice.org.uk/nicemedia/live/13712/66658/.pdf.

Classification of Chronic Kidney Disease

Chronic kidney disease is an important public health problem, emphasizing the need for earlier identification and treatment. Historically, data obtained from epidemiologic surveys were compromised by lack of consistent surrogate markers of kidney function to identify established disease. For example, serum creatinine, calculated creatinine **clearance** (C_{Cr}), and measured C_{Cr} were variously used. Landmark guidelines developed by the National Kidney Foundation–Kidney Disease Outcomes Quality Initiative (NKF-K/DOQI) were published in 2002 and were based on categories of GFR. They have subsequently been revised and updated by Kidney Disease: Improving Global Outcomes (KDIGO) and broadly adopted by other national organizations (Table 35.1). The KDIGO 2012 guideline added a second dimension to the classification system with three identified levels of albuminuria, acknowledging the powerful additional prognostic information imparted by the presence of proteinuria (see Table 35.1).

A high prevalence of CKD is identified with the classification system, with an estimated 1 in 7 adults affected within the United States. Most individuals with CKD do not progress to kidney failure, with prevalence rates of patients in GFR category G3 10 to 20 times higher than the prevalence rates of GFR categories G4 and G5. There is some concern that many individuals being identified with G3 CKD are not at increased risk; use of cystatin C to delineate risk in this population is being evaluated.

TABLE 35.1 Classification of Chronic Kidney Disease Indicating Prognosis and Secondary/Tertiary Care Referral Decision Making by Glomerular Filtration Rate and Albuminuria Categories

				PERSISTENT ALBUMINURIA CATEGORIES: DESCRIPTION AND RANGE		
				A1	A2	A3
				Normal to mildly increased <30 mg/g (<3 mg/mmol)	Moderately increased 30–300 mg/g (3–300 mg/mmol)	Severely increased >300 mg/g (>30 mg/mmol)
Glomerular filtration rate categories (mL/min/1.73 m²): description and range	G1	Normal or high	>90	55.6	1.9 Monitor	0.4 Refer[a]
	G2	Mildly decreased	60–89	32.9	2.2 Monitor	0.3 Refer[a]
	G3a	Mildly to moderately decreased	45–59	3.6 Monitor	0.8 Monitor	0.2 Refer
	G3b	Moderately to severely decreased	30–44	1.0 Monitor	0.4 Monitor	0.2 Refer
	G4	Severely decreased	15–29	0.2 Refer[a]	0.1 Refer[a]	0.1 Refer
	G5	Kidney failure	<15	<0.1 Refer	<0.1 Refer	0.1 Refer

DARK BLUE, very high risk; *MIDDLE BLUE,* high risk; *BLUE,* moderately increased risk; *LIGHT BLUE,* low risk (if no other markers of kidney disease, no CKD).

[a]Referring clinicians may wish to discuss with their local nephrology service, depending on local arrangements regarding referral and monitoring.

NOTE: Numbers in the cells show the proportion (in %) of the adult population in the United States.

Modified from Kidney Disease Improving Global Outcomes. (2013). Clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney International, 3,* 1–150.

The main causes of CKD leading to kidney failure in the United States are indicated in Fig. 35.5. As shown, diabetes mellitus is the largest single cause of advanced CKD and accounts for almost 50% of new dialysis patients in the United States. Hypertension is the underlying diagnosis in around 28% of new dialysis patients and is particularly prevalent among African Americans. In addition, the rate of new ESRD in African Americans is 3.5 times higher than that among whites. The myriad of kidney diseases include (1) **glomerulonephritis**, (2) infection, (3) hereditary conditions, (4) systemic conditions, (5) interstitial conditions, (6) obstructive conditions, and (7) those of unknown origin.

AT A GLANCE

Defining and Classifying CKD[a]

CKD is defined as abnormalities of kidney structure or function present for at least 3 months with implications for health.

Markers of reduced kidney function or damage include:

- Reduced glomerular filtration rate (GFR <60 mL/min per 1.73 m²)
- Albuminuria (≥30 mg/24 h; ACR ≥30 mg/g [≥3 mg/mmol])
- Urine sediment abnormalities
- Electrolyte and other abnormalities due to tubular disorders
- Abnormalities detected by histology
- Structural abnormalities detected by imaging
- History of kidney transplantation

Assign GFR and albuminuria categories as described in Table 35.1.

[a]Modified from Kidney Disease Improving Global Outcomes. (2013). *Clinical practice guideline for the evaluation and management of chronic kidney disease. Kidney International, 3,* 1–150.

Management of Chronic Kidney Disease

Irrespective of the underlying cause, the rate of progression of CKD (i.e., the rate at which a patient approaches kidney failure) is dependent on both nonmodifiable factors, such as age, gender, race, and level of kidney function at diagnosis, and modifiable characteristics, including proteinuria, blood pressure control, and smoking.

Lowering blood pressure and reduction of proteinuria have been shown to decrease the progression of CKD. The Modification of Diet in Renal Disease (MDRD) study compared the rates of decline in GFR in patients with various causes of CKD assigned to either a "usual" or "low" blood pressure goal. Outcome data suggest that the low blood pressure goal had some beneficial effect in those patients with higher concentrations of proteins in their urine. In practice, medications that inhibit parts of the RAAS are typically preferred because these drugs have been demonstrated to reduce proteinuria and rate of progression of CKD.

Protein intake is restricted spontaneously to approximately 0.6 to 0.8 g/kg per day in uremic patients not receiving dietary advice. To prevent malnutrition, patients receive professional dietary advice with diets containing an increased proportion of first-class protein and increased calorie content of up to 35 kcal/kg per day. General health measures, including cessation of cigarette smoking, are encouraged. Complications of CKD that develop before the need for RRT are numerous and include (1) cardiovascular disease, (2) bone disease, and (3) anemia. Albuminuria and proteinuria have been shown to be associated with increased risk of cardiovascular disease, cardiovascular mortality, and all-cause mortality.

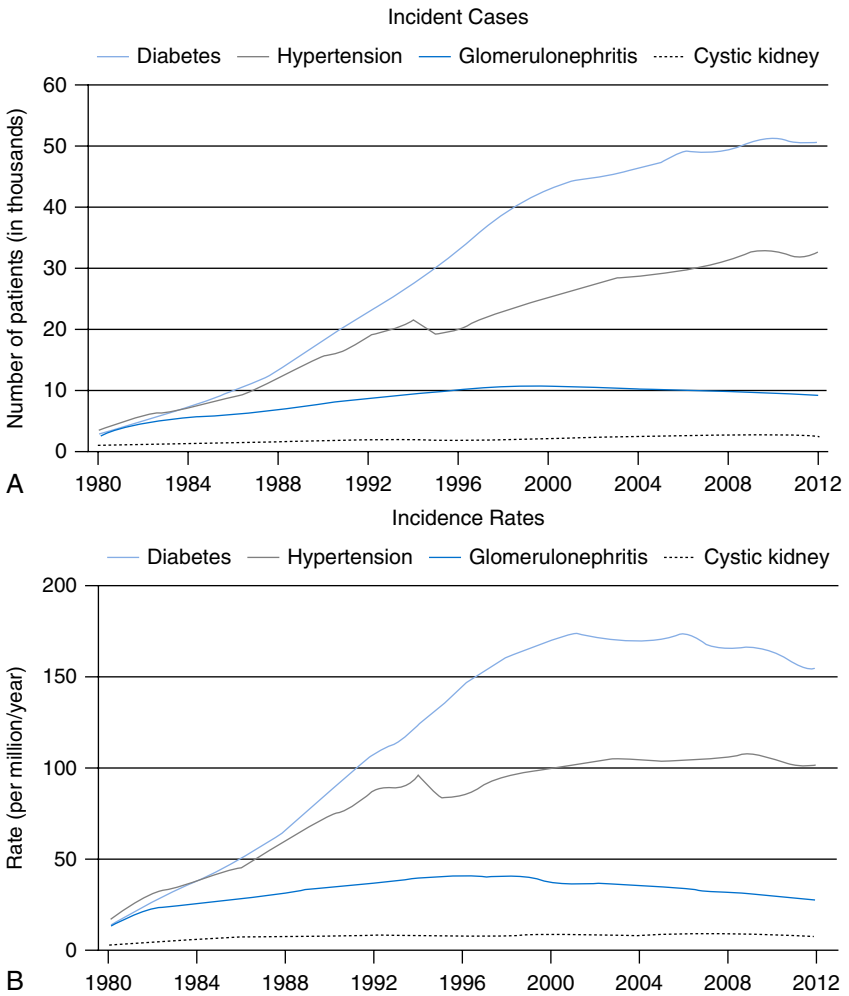


Fig. 35.5 Trends in (A) ESRD incident cases, in thousands, and (B) adjusted* ESRD incidence rate, per million/year, by *primary cause of ESRD*, in the US population, 1980–2012. (From US Renal Data System, 2014 USRDS Annual Data Report. [2014]. *An overview of the epidemiology of kidney disease in the United States*. Bethesda, MD: National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases.)

Cardiovascular Complications of Chronic Kidney Disease

The incidence of cardiovascular disease is seven- to ten-fold greater in patients with CKD than in non-CKD age- and sex-matched controls; by the time patients develop the need for RRT, the risk becomes approximately 17 times greater.

The spectrum of cardiovascular disease in CKD includes (1) angina, (2) congestive heart failure, (3) myocardial infarction, (4) peripheral vascular disease, (5) stroke, (6) sudden cardiac death, and (7) transient ischemic attack. Structural heart disease, such as left ventricular hypertrophy (LVH) and valvular heart disease, is very common. Up to 75% of patients commencing dialysis have echocardiographic evidence of LVH. Vascular calcification is highly prevalent in dialysis patients and is associated with reduced survival.

Dyslipidemia in Chronic Kidney Disease

The pattern of dyslipidemia (abnormal concentration of lipids in the blood) in CKD differs from that seen in non-CKD. Although total cholesterol concentration may be normal, there is often a highly abnormal lipid subfraction profile with a predominance of atherogenic small, dense, low-density lipoprotein (LDL) particles. In addition, triglyceride concentrations are high. Low cholesterol is associated with increased mortality in dialysis patients, likely reflecting coincident inflammation and malnutrition ("reverse causality"). Although there are no data to suggest that statins are of benefit in patients receiving dialysis, the Study of Heart and Renal Protection (SHARP) evaluated the use of cholesterol lowering with simvastatin and ezetimibe in 9000 patients (including 3000 dialysis patients) and demonstrated significant reductions in stroke and arterial revascularization in the treated patients followed for a median of 4.9 years.

Disturbances in calcium and phosphorus metabolism and bone disease in chronic kidney disease. Bone disease as a consequence of CKD has long been recognized. As GFR declines, renal activation of vitamin D decreases and plasma phosphate concentration rises, resulting in reduced ionized calcium. Consequently the parathyroid glands increase the

production of PTH. This increased secretion of PTH stimulates resorption of calcium and phosphate from the bone, the body's major calcium reservoir. Problems develop early, typically once GFR has fallen to approximately 45 mL/min per 1.73 m^2. Secondary hyperparathyroidism classically causes (1) bone changes, (2) so-called high-turnover bone disease, and (3) increased risk of fracture.

"Adynamic," low-turnover bone disease is also highly prevalent in patients with CKD and is characterized by poor bone formation. It is more common in the elderly and those with diabetes mellitus and malnutrition. Adynamic bone is associated with (1) a low PTH concentration, (2) abnormal calcium balance, (3) hyperphosphatemia, (4) **acidosis**, and (5) the use of high doses of vitamin D analogues (e.g., alfacalcidol [1α-hydroxyvitamin D_3]). Vascular calcification is commonly associated with adynamic bone disease. The diagnosis of adynamic bone is ultimately made on bone biopsy, but patients are reluctant to undertake this procedure. Unfortunately PTH alone does not correlate with bone biopsy findings. Serum bone-specific alkaline phosphatase (ALP) measured using an immunoassay technique has good predictive value in separating high from low bone turnover, particularly in combination with PTH measurements. For example, low bone formation has been associated with bone-specific ALP concentrations below 20 μg/L and PTH below 200 ng/L.

High concentrations of plasma phosphate are associated with increased mortality in hemodialysis (HD) patients. Strategies to reduce phosphate concentrations are employed routinely in the treatment of patients on dialysis. Phosphate is present in many foods and is linearly associated with protein ingestion. The recommended allowance of phosphate is reduced for patients on dialysis to around 800 mg/day. Treatment with vitamin D analogues increases gut absorption of phosphate from approximately 65% to almost 85%. The use of phosphate binders, taken with meals, is almost universal in dialysis patients, reducing phosphate absorption to 30% to 40%. It has been possible to normalize phosphate balance in patients treated by daily HD.

Anemia

The World Health Organization (WHO) defines anemia as a hemoglobin concentration of less than 13 g/dL in men and less than 12 g/dL in women. In the absence of treatment, anemia is inevitable as CKD progresses. Therapies are available to correct anemia; therefore it is mandatory to assess a patient with CKD for this condition. The NKF KDOQI guideline recommends that an estimated GFR of less than 60 mL/min per 1.73 m^2 should be the cutoff value for determining the presence or absence of anemia. Detection is important, since treatment may alleviate many of the symptoms of CKD and hopefully reduce risk of LVH. The pathology of anemia in CKD is multifactorial, but the predominant cause is loss of peritubular fibroblasts (specialized cells that produce collagen and other materials) within the renal cortex that synthesize EPO. Failure to produce EPO leads to decreased numbers of red cells and concomitant decreased concentrations of hemoglobin. Other causes of anemia include (1) absolute or functional iron deficiency, (2) folic acid and vitamin B_{12} deficiencies, and (3) chronic inflammation. Red cell survival may also be reduced. Treatment with recombinant erythropoiesis-stimulating agents is recommended to correct anemia. The gene for human EPO was cloned in 1985 and rhEPO-, epoetin-, or EPO-stimulating agents (ESAs) were introduced into clinical practice shortly thereafter. The most common side effect is hypertension therefore blood pressure should be well controlled before treatment is introduced. The majority of patients respond to treatment. However, failure to respond requires thorough investigation for many potential causes, such as (1) occult blood loss, (2) hyperparathyroidism, (3) iron deficiency, (4) vitamin B_{12} deficiency, (5) folate deficiency, and (6) inadequate dialysis. Many clinical benefits are derived from correcting anemia with ESAs, including (1) improved exercise capacity, (2) improved cognitive function, (3) better quality of life, and (4) increased libido. Later studies have, however, highlighted the risks of high doses of ESAs in nondialysis CKD patients with associated increases in mortality and cardiovascular morbidity. These studies have been instructive in setting the current hemoglobin target of 10.0 to 12.0 g/dL and in recommending ESA dose adjustments when hemoglobin is less than 10.5 or greater than 11.5 g/dL to balance benefit versus safety to patients.

Treatment of anemia in CKD requires adequate iron stores. For example, in patients with CKD, a plasma ferritin less than 100 μg/L is considered to suggest iron deficiency, and a plasma ferritin of 100 to 200 μg/L in association with a transferrin saturation (TSAT) of less than 20% represents "functional" iron deficiency. Parenteral iron is the treatment of choice for absolute and functional iron deficiency, since oral iron has low efficacy in CKD.

Uremic Syndrome

Uremic syndrome comprises (1) symptoms, (2) physical signs, and (3) abnormal findings on diagnostic studies that result from the failure of the kidneys to maintain adequate (1) excretory, (2) regulatory, and (3) endocrine function. It is considered the terminal clinical manifestation of kidney failure. At least 90 organic compounds have been shown to be retained in uremia.

The classic signs of **uremia** (**azotemia**) include (1) progressive weakness and easy fatigue, (2) loss of appetite followed by (3) nausea and vomiting, (4) muscle wasting, (5) tremors, (6) abnormal mental function, (7) frequent but shallow respirations, and (8) **metabolic acidosis**. The syndrome evolves to produce (1) stupor, (2) coma, and (3) ultimately death unless support is provided by dialysis or successful kidney transplantation. The composition of plasma is abnormally labile in response to such factors as (1) diet, (2) state of hydration, (3) gastrointestinal bleeding, (4) vomiting, (5) diarrhea, and (6) intake of therapeutic drugs. Patients with kidney failure (GFR of ≤15 mL/min per m^2) will generally exhibit signs and symptoms of uremia, or a need for RRT.

The most characteristic laboratory findings are increased concentrations of nitrogenous compounds in plasma, such as urea and creatinine, as a result of reduced GFR and decreased

tubular function. Retention of these compounds and of metabolic acids is followed by progressive (1) hyperphosphatemia, (2) hypocalcemia, and (3) potentially dangerous hyperkalemia. Although most patients eventually exhibit acidemia, respiratory compensation by the elimination of carbon dioxide is extremely important. In addition, reduced endocrine function is manifested by inadequate synthesis of EPO and calcitriol, with resulting anemia and osteomalacia. Disordered regulation of blood pressure generally leads to hypertension. Biochemical characteristics of the uremic syndrome are summarized in the following Points to Remember box.

In addition to the consequences of reduced (1) excretory, (2) regulatory, and (3) endocrine function of the kidneys, the uremic syndrome has several systemic manifestations, among which are (1) pericarditis, (2) pleuritis, (3) disordered platelet and granulocyte function, and (4) encephalopathy.

Many retained metabolites have been implicated in the systemic toxicity of the uremic syndrome. Although urea was the first of these metabolites to be identified as being increased in uremia, it does not appear to be responsible for the systemic manifestations of uremia. Urea is a 60-Da water-soluble compound (see Chapter 21) that has the highest concentration of presently known uremic retention solutes in uremic plasma. Although its removal by dialysis is directly related to patient survival, the effects of urea on biologic systems are not clear. Urea removal by dialysis is not necessarily representative of other molecules retained in the uremic syndrome, particularly protein-bound solutes or lower-molecular-weight proteins, such as PTH and cystatin C.

POINTS TO REMEMBER

Biochemical Characteristics of the Uremic Syndrome

Retained Nitrogenous Metabolites

Urea
Cyanate
Creatinine
Guanidine compounds
Uric acid

Fluid, Acid-Base, and Electrolyte Disturbances

Fixed urine osmolality
Metabolic acidosis (decreased blood pH, bicarbonate)
Hyponatremia or hypernatremia
Hypokalemia or hyperkalemia
Hyperchloremia
Hypocalcemia
Hyperphosphatemia
Hypermagnesemia

Carbohydrate Intolerance

Insulin resistance (hypoglycemia may also occur)
Plasma insulin normal or increased
Delayed response to carbohydrate loading
Hyperglucagonemia

Abnormal Lipid Metabolism

Hypertriglyceridemia
Decreased high-density lipoprotein cholesterol
Hyperlipoproteinemia

Altered Endocrine Function

Secondary hyperparathyroidism
Osteomalacia (secondary to abnormal vitamin D metabolism)
Hyperreninemia and hyperaldosteronism
Hyporeninemia
Hypoaldosteronism
Decreased erythropoietin production
Altered thyroxine metabolism
Gonadal dysfunction (increased prolactin and luteinizing hormone, decreased testosterone)

Acute Kidney Injury

The definition of AKI, endorsed by KDIGO during 2012, is the occurrence of any one of the following:

1. Increase of plasma creatinine by equal to or greater than 0.3 mg/dL (≥26 μmol/L) within 48 hours
2. Increase in plasma creatinine to equal to or greater than 1.5 times baseline, which is known or presumed to have occurred within the prior 7 days
3. Reduction in urine output (documented oliguria <0.5 mL/kg per hour for >6 hours)

In addition, AKI is staged from 1 to 3, depending on severity. In the United Kingdom, the National Confidential Enquiry into Patient Outcome and Death: Adding Insult to Injury (http://www.ncepod.org.uk/2009report1/Downloads/AKI_summary.pdf, accessed April 08, 2013) reported that failure (1) to identify intravascular volume depletion, (2) to withhold nephrotoxic drugs, and (3) of early diagnosis of causative conditions (such as, sepsis) directly contributed to in-hospital mortality associated with AKI. Approximately 20% of postadmission AKI episodes are predictable and avoidable. Therefore prompt administration of intravenous crystalloid solutions (such as, 0.9% sodium chloride) may prevent further deterioration of AKI in many cases. Patients at risk for AKI include (1) older persons, (2) those with preexisting CKD, (3) those with sepsis, (4) those with diabetes mellitus, (5) patients with heart disease, and (6) those receiving nephrotoxic drugs, particularly in the setting of hypovolemia. Clinical assessment of AKI should consider whether the precipitant is (1) prerenal, (2) intrarenal (intrinsic), or (3) postrenal. As intrinsic AKI is caused by primary vascular, glomerular, or interstitial disorders, it is important that all patients presenting with AKI undergo urinalysis to test for (1) infection, (2) hematuria, and (3) proteinuria. In most cases, the kidney lesion seen on histology is referred to as **acute tubular necrosis (ATN)**. ATN is caused by ischemic or nephrotoxic injury to the kidney. In 50% of cases of hospital-acquired AKI, the cause is multifactorial. Urinary electrolyte measurements are seldom required or useful in the investigation of AKI. Laboratory testing of blood is crucial in the management of AKI. Blood tests also assist in establishing the underlying diagnosis, and specific investigations are requested if kidney function has not improved following volume correction. An acute renal screen should clearly focus on the most likely diagnoses and include the tests shown in Table 35.2.

TABLE 35.2 Investigation of Acute Kidney Injury

Test	Indication/Comments
Urine Testing	
Urine reagent strip ("dipstick")	Hematuria and proteinuria may indicate glomerular origin
Red cell casts on microscopy	(Not available universally; may need bedside microscope)
Urine microscopy and culture	Identify urinary tract infection
Urine protein electrophoresis and immunofixation	
Blood Tests	
Baseline Studies	
Urea, electrolytes, and creatinine	Check previous laboratory reports: AKI or AKI with preexisting CKD
Calcium, phosphate, albumin	
Liver function tests	Suspected multiorgan involvement or abnormal coagulation
Acid-base studies	Arterial blood gas or venous plasma bicarbonate concentration
Full blood count	Anemia, hemolysis, thrombocytopenia
Coagulation studies	Evidence of intravascular coagulation; need to normalize if considering kidney biopsy and central line insertion
Selected Additional Investigations	
Blood culture	Any infection, but especially endocarditis, severe pneumonia, or urinary tract sepsis
Creatine kinase	Very high in cases of muscle inflammation and necrosis (rhabdomyolysis)
Lactate dehydrogenase	If high, suspect renal infarction and consider hemolysis
Antineutrophil cytoplasmic antibodies	Vasculitides
Antiglomerular basement membrane antibody	Antiglomerular basement membrane disease
Antinuclear antibodies	SLE
Anti-dsDNA antibodies, extractable nuclear antigens	SLE
Low C3 complement	SLE, MPGN, C3 glomerulopathies
Low C4 complement	Systemic lupus erythematosus, atheroembolism, cryoglobulinemia, MPGN (immune complex positive)
Cryoglobulin	Cryoglobulinemia
Urate	Urate nephropathy
Serum protein electrophoresis	Myeloma
Virology studies	Hepatitis serology, human immunodeficiency virus
Other serology	Antistreptolysin O titer
Imaging	
Chest x-ray	Pulmonary edema, pneumonia, effusions, malignancy, and granulomas
Abdominal x-ray (kidney, ureter, and bladder)	Renal stones
Renal tract ultrasound scan	Identify size and symmetry of kidneys Evidence of an obstructed system Small, shrunken kidneys in advanced CKD
Computed tomography scan	Anatomy and perfusion
Magnetic resonance imaging	Angiography to identify renovascular lesions
Formal angiography	Critical renal artery stenosis
Kidney Biopsy	
	Reserved for patients with unexplained AKI in whom acute tubular necrosis is not suspected. It is anticipated that additional therapy as with steroids, cytotoxic drugs, and plasma exchange may be required.

AKI, Acute kidney injury; *CKD,* chronic kidney disease; *dsDNA,* double-stranded DNA; *MPGN,* membranoproliferative glomerulonephritis; *SLE, systemic lupus erythematosus*

In addition to biochemical testing, there is a role for ultrasound and radiologic imaging in kidney disease, in particular for the exclusion of obstruction. Kidney biopsy is generally reserved for cases of AKI wherein an ultrasound scan has excluded obstructed kidneys, kidney sizes are maintained, the cause of AKI is otherwise unexplained, and an intrinsic pathology is suspected.

Metabolic acidosis is the most common acid-base disorder in patients with AKI. Reduced renal excretion of potassium and the effects of acidosis on the generation of extracellular potassium may lead to a very high concentration of potassium in the plasma. Severe hyperkalemia (a plasma potassium concentration >6.5 mmol/L) is associated with life-threatening cardiac arrhythmias. Emergency

treatment of hyperkalemia should be instituted as necessary. Infusion of high concentrations of glucose stimulates insulin secretion from the pancreas and the uptake of potassium into the intracellular space. A fall in potassium concentration should be expected within 60 minutes. In addition, low-dose rapid-acting insulin is sometimes administered with the glucose bolus. Serum potassium concentration should be monitored hourly until the risk of a life-threatening cardiac event has passed and no evidence of potassium rebound is found. Blood glucose concentration should be monitored because hypoglycemia may occur following exogenous insulin. When hyperkalemia persists despite appropriate medical measures, dialysis should be considered. Recovery from AKI usually occurs within days or weeks following removal of the initiating event. However, uncomplicated AKI has a mortality rate of 5% to 10%, although AKI complicating the failure of a nonrenal organ system in the intensive care unit setting is associated with mortality rates approaching 50% to 70%.

POINTS TO REMEMBER

Causes of AKI

- Prerenal AKI:
  - Hemorrhage
  - Diarrhea
  - Postoperative fluid and blood losses
  - Sepsis
  - Acute cardiac failure
- Renal (intrinsic renal disease) AKI:
  - Tubular
    - Any of the "prerenal" causes that are severe or that are not corrected promptly leading to ATN. Other causes of ATN include:
    - Drug nephrotoxicity
      - NSAIDs, ACE inhibitors
      - Aminoglycoside antibiotics
      - Amphotericin
    - Contrast nephropathy
    - Poisoning
    - TIN
      - Allergic TIN associated with antibiotics and NSAIDs
      - Sarcoidosis
      - Pyelonephritis
    - Glomerular
      - RPGN (ANCA-associated vasculitides, Goodpasture disease, SLE, other crescentic glomerulonephritides)
      - Thrombotic microangiopathies
      - Cryoglobulinemia
      - Atheroembolism
    - Vascular
      - Aortic dissection
      - Renal vein thrombosis
    - Miscellaneous
      - Rhabdomyolysis
      - Urate nephropathy
      - Hepatorenal syndrome
    - Postrenal AKI
      - Bladder outflow obstruction
      - Benign and malignant prostate disease
      - Invasive bladder carcinoma
      - Bilateral renal calculi or calculi within a single kidney
      - Retroperitoneal fibrosis

ACE, Angiotensin converting enzyme; *AKI,* acute kidney injury; *ANCA,* antineutrophil cytoplasmic antibody; *ATN,* acute tubular necrosis; *NSAIDs,* nonsteroidal antiinflammatory drugs; *RPGN,* rapidly progressive glomerulonephritis; *SLE,* systemic lupus erythematosus; *TIN,* tubulointerstitial nephritis.

DISEASES OF THE KIDNEY

Diabetic Nephropathy

Diabetic nephropathy is a clinical diagnosis based on the finding of proteinuria in a patient with diabetes mellitus. Overt nephropathy is characterized by protein loss greater than 0.5 g/day, approximately equivalent to albumin loss of greater than 300 mg/day or category A3 albuminuria. For a variety of reasons, it is preferable to assess proteinuria as albuminuria (see Chapter 21), and albumin has long been uniformly adopted as the "criterion standard" in evaluating diabetes-related kidney damage. Patients with albuminuria exceeding 30 mg/day have pathologically increased albuminuria. Diabetic nephropathy is the most common cause of ESRD in the United States, accounting for approximately 40% of incident patients in RRT programs. The prevalence of diabetes mellitus as a cause of ESRD is variable among ethnic groups within the United States, with whites having rates approximately 25% of those observed among African Americans. Among patients who require dialysis, those with diabetes mellitus have a 22% higher mortality at 1 year and a 15% higher mortality at 5 years than patients without diabetes. The cornerstones of treatment for diabetic nephropathy consist of (1) glycemic control; (2) blood pressure control, particularly with drugs that block the RAAS; and (3) management of cardiovascular risk.

Hypertension

Hypertension is second only to diabetes mellitus as a primary diagnosis of ESRD for incident patients commencing dialysis in the United States (see Fig. 35.5). The incidence is higher in older people and especially among the black population in the United States. Hypertension often develops as a consequence of CKD because of alterations in salt and water metabolism and activation of the sympathetic nervous and renin-angiotensin systems. Hypertension can act as an accelerating force in the development of ESRD. Treatment of hypertension to predefined target blood pressure values is critical in preventing progression to ESRD.

Glomerular Diseases

Glomerular disease is suggested clinically by the finding of blood and protein in the urine on urine reagent strip testing. Proteinuria of greater than 1 g/day (0.88 g/g creatinine;

100 mg/mmol creatinine) in the absence of an overflow-type proteinuria, such as myoglobinuria or light chain–related disease, is invariably glomerular in origin. Although a detailed discussion of each glomerular disease is beyond the scope of this book, the most important diseases are discussed to illustrate the spectrum of disease. For further information regarding clinical guidelines, the reader is directed to the KDIGO *Clinical Practice Guideline for Glomerulonephritis*. In the United Kingdom, glomerulonephritis accounts for 14% of new cases of established renal failure.

Primary Glomerular Disease

Glomerulonephritis can be primary (affecting only the kidneys) or secondary (in which the kidneys are involved as part of a systemic process). Histopathologic classification of glomerulonephritis may appear slightly cumbersome, but it is readily simplified by consideration of the glomerular structures and cells that may be involved and the presence or absence of immune complexes. In essence, only three cell types are involved—endothelial, epithelial, and mesangial—plus the acellular GBM. The glomerular cells and the GBM have a limited range of response to injury—namely, proliferation, scarring (sclerosis), and GBM thickening. The term *focal* is used if fewer than half of the glomeruli are involved in the disease process as seen on light microscopy, whereas *diffuse glomerulonephritis* refers to cases in which all glomeruli are involved. Immune deposits identified following immunofluorescence or immunoperoxidase staining do not define whether a disease is focal or diffuse.

IgA nephropathy. **IgA nephropathy** is an example of a focal glomerulonephritis with focal mesangial cell proliferation demonstrated by light microscopy. However, diffuse and global deposition of the immunoglobulin IgA can be demonstrated following immunostaining. It is the most common type of glomerulonephritis worldwide and has a particularly high prevalence around the Pacific rim. The disease tends to be slowly progressive (in terms of loss of kidney function), depending, as with most kidney diseases, on the degree of proteinuria, kidney function at the time of diagnosis, and degree of interstitial fibrosis on kidney biopsy. Up to 50% of patients exhibit increased concentrations of serum IgA, although diagnosis depends on kidney biopsy findings. Clinical presentation varies considerably from asymptomatic microscopic hematuria to macroscopic hematuria, proteinuria including nephrotic syndrome, and crescentic glomerulonephritis with kidney failure. Episodic macroscopic hematuria is seen in some patients at the same time as an upper respiratory tract infection.

Treatment options range from tonsillectomy in patients with macroscopic hematuria associated with respiratory infection to no treatment in those with isolated microscopic hematuria or those with proteinuria of less than 1 g/day. In progressive disease, all patients are treated in a similar generic fashion as for most kidney diseases, including targeting blood pressure to less than 125/75 mm Hg and using comprehensive RAAS blockade to minimize proteinuria. In addition, selected patients may receive corticosteroids in the form of prednisone.

Nephrotic syndrome. **Acute nephrotic syndrome** is defined as heavy proteinuria (>3 g/day), reduced serum albumin concentration, and edema. In comparison with nephritic syndromes, nephrotic patients may exhibit an otherwise bland urinary sediment with little hematuria. Nephrotic syndrome can occur in patients of any age, from neonates to the elderly. Although the underlying kidney disease tends to vary with age, in all cases the lesion is within the glomerulus and is associated with damage to the specialized visceral epithelial cells, the podocytes. Proteinuria is a consequence of a reduction in the charge-selective properties of the filtration barrier, particularly the GBM, and of alterations in the slit diaphragms of interdigitating foot processes of adjacent podocytes.

The most common causes of nephrotic syndrome are minimal change disease (MCD), focal segmental glomerulosclerosis (FSGS), and membranous nephropathy. Detection of autoantibodies directed to M-type phospholipase A_2 receptor PLA_2R (anti-PLA_2R antibody) in patients with membranous nephropathy indicates primary disease rather than a secondary process and guides therapy.

Secondary causes include diabetic nephropathy, amyloidosis, and SLE. A kidney biopsy is generally undertaken in all adult patients who present with nephrotic syndrome. Nephrotic syndrome is associated with increased cardiovascular disease as a result of marked hyperlipidemia and increased risk of infection and thromboembolic disease. In addition, AKI may supervene in cases of nephrotic syndrome, and prolonged proteinuria with a poor response to treatment may lead to kidney failure.

The management of nephrotic syndrome depends on the underlying glomerular lesion, although general principles apply in all cases (Box 35.1). In addition to general measures, specific treatment targeted at inducing remission from proteinuria usually requires a combination of immunosuppressive drugs, including corticosteroids and cytotoxic drugs.

BOX 35.1 Management of Nephrotic Syndrome

- Low-sodium diet
- Protein intake of 1.0 g/kg per day
- Fluid management: usually includes loop diuretics
- Thromboembolism prophylaxis: may include formal anticoagulation in high-risk patients (serum albumin <20 g/L or in nephrotic syndrome due to membranous nephropathy or mesangioproliferative glomerulonephritis)
- Vigilance for infections
- Treatment of hyperlipidemia with 3-hydroxy-3-methylglutaryl-Co-enzyme A (HMG)-CoA-reductase inhibitors (statins)
- Treatment of hypertension, primarily with RAAS blockade
- Supportive treatment of AKI
- Education and psychological support for patients and relatives

AKI, Acute kidney injury; *RAAS,* renin-angiotensin-aldosterone system.

Rapidly Progressive Glomerulonephritis

Rapidly progressive glomerulonephritis (RPGN) is a heterogeneous group of disorders characterized by a fulminant clinical course that leads to kidney failure in only weeks or months. The clinical picture of RPGN is often preceded by a systemic illness for several months.

These syndromes are often characterized by focal glomerulonephritis with glomerular ischemia, infarction, and tissue death (necrosis). Following release of inflammatory cytokines and chemokines from the necrotic capillaries, there is proliferation of the epithelial cells of the Bowman capsule. The proliferated cells lie on top of adjacent cells and form a partial circle around the inner rim of the Bowman capsule, which is referred to as a *crescent*. Proliferating epithelial cells and macrophages eventually compress the glomeruli and obstruct the proximal convoluted tubules, thus severely compromising nephron function.

Anti-GBM antibodies may be present along the GBM in anti-GBM disease. Most commonly, however, there is no or little Ig deposition within the glomerulus (so-called pauci-immune). Approximately 80% of patients with active pauci-immune necrotizing and crescentic glomerulonephritis have been shown to possess antineutrophil cytoplasmic antibodies (ANCAs) irrespective of the presence or absence of a concomitant systemic vasculitis. This strong association has allowed serologic discrimination of this type of glomerulonephritis from other types of RPGN. Two subtypes of ANCA have been described—cytoplasmic (C-ANCA) and perinuclear (P-ANCA)—reflecting the patterns observed by indirect immunofluorescence microscopy using alcohol-fixed neutrophils as a substrate. C-ANCAs are directed toward a plasma proteinase (PR3) in neutrophil primary granules, whereas the P-ANCAs target antigen is usually myeloperoxidase (MPO). Immunoassays have been used to measure anti-PR3 and anti-MPO antibody titers. ANCAs appear in the plasma of almost all patients with active and generalized disease and are useful in diagnosis of the disease. However, false positives and false negatives may occur, necessitating histologic diagnosis where possible.

In addition to measurement of ANCA, C-reactive protein (CRP) is extremely helpful in assessment of the acute-phase reaction in active disease processes and is critical in helping to define disease remission. In patients in whom the CRP has returned to normal, disease remission is indicated; this prompts the clinician to reduce immunosuppressant treatment doses.

Systemic Lupus Erythematosus

Systemic lupus erythematosus (SLE) is a chronic inflammatory disease of unknown cause that can affect the skin, joints, kidneys, lungs, nervous system, serous membranes, and/or other organs of the body. Renal involvement, termed lupus **nephritis**, occurs in up to 60% of adults with SLE. Lupus nephritis is especially common in black and Hispanic patients in the United States. It may present variably from incidental hematuria and proteinuria, nephrotic syndrome, or a fulminating RPGN. Most (75%) patients with SLE develop abnormal urinalysis or impaired kidney function during the course of the disease. Detection of lupus nephritis involves urine testing for blood and protein and tests of kidney function. In addition, serologic testing for autoantibodies to nuclear antigens and measurement of complement components C3 and C4 are undertaken. Significant hypocomplementemia and increased anti–double stranded DNA (anti-dsDNA) titers suggest active disease. Combined use of corticosteroids and intravenous or oral cyclophosphamide has been the conventional treatment for diffuse proliferative lupus nephritis since the 1970s. Recent data suggest that mycophenolate mofetil and corticosteroids can be as effective as but not superior to intravenous cyclophosphamide for induction treatment for lupus nephritis.

Interstitial Nephritis

A variety of chemical, bacterial, and immunologic injuries to the kidney may cause generalized or localized changes that primarily affect the tubulointerstitium rather than the glomeruli. This group of disorders is characterized by alterations in tubular function that, in advanced cases, may cause secondary vascular and glomerular damage. **Interstitial nephritis** is associated with proteinuria that is less severe than in glomerular disease. In addition to infection-related **pyelonephritis**, interstitial nephritis may present in acute and chronic forms and has many causes. Acute allergic interstitial nephritis presents with AKI and marked inflammation of the interstitium. A drug hypersensitivity reaction is the most common form of acute interstitial nephritis. Urinary findings may be normal, or low-level proteinuria and eosinophils may be seen on light microscopy. More than 100 different drugs have been implicated, but NSAIDs and β-lactam antibiotics are the drugs most commonly identified. Treatment is directed at removing any causative agent. Steroids are used to promote early resolution of the clinical course, although patients can develop chronic interstitial fibrosis.

Monoclonal Light Chains and Kidney Disease

Ig molecules are formed in secretory B cells (plasma cells) and consist of heavy chains, which denominate the antibody isotype, and either kappa (κ) or lambda (λ) light chains. The proportion of Ig containing κ versus λ is 3:2 in humans. The molecular weight of light chains is approximately 23 kDa. Excess production of light over heavy chains appears to be required for efficient Ig synthesis, resulting in the release of free light chains (FLCs) into the circulation. In normal individuals, the small quantity of circulating polyclonal light chains is filtered by the glomerulus, and 90% is reabsorbed in the proximal tubule and degraded by proteases.

Myeloma is a neoplastic proliferation of a clone of plasma cells that produce excessive amounts of a monoclonal (M) protein and FLCs. In multiple myeloma, complete monoclonal Igs (usually IgG or IgA) are accompanied in the plasma by variable concentrations of FLCs that appear in the urine as **Bence Jones proteins** (named after Henry Bence Jones, who first described these in 1848). M-proteins and FLCs can be identified in the blood and/or the urine in 98% of patients

with myeloma using protein electrophoresis and immunofixation. Immunoparesis, with reduction in non–M-protein Ig, is characteristic of myeloma. The FLC immunoassay is now also used to detect monoclonal Ig. The kidneys are often affected in myeloma, with diverse clinical and pathologic presentations. The three most common forms of monoclonal Ig-mediated kidney disease are cast nephropathy, monoclonal Ig deposition disease (MIDD), and AL amyloidosis. Impairment of kidney function at presentation occurs in almost 50% of patients with myeloma. Although most recover following treatment for other factors contributing to renal impairment (e.g., dehydration, hypercalcemia, infection, nephrotoxic drugs), about 10% have severe renal involvement caused by the effects of monoclonal FLCs on the kidney. Severe kidney failure may occur in myeloma following deposition of light chains within tubules—so-called cast nephropathy (or myeloma kidney).

Polycystic Kidney Disease

Autosomal dominant **polycystic kidney disease** (ADPKD) is the second most common inherited monogenic disease (after familial hypercholesterolemia), with an estimated incidence of 1:1000. It is by far the most common inherited kidney disease; 12.5 million people worldwide are affected. Approximately 50% of ADPKD patients develop kidney failure by age 55 years. It is therefore important to make the diagnosis in affected families and to monitor kidney function regularly. The intervals between estimations of GFR will depend on the stage of CKD, as with other progressive kidney diseases. An important clinical observation is the highly variable phenotype within families. The disease causes the development of multiple kidney cysts and extrarenal cysts in the liver and pancreas. About 10% of ADPKD families have a strong family history of intracranial arterial aneurysm rupture. Hypertension is an early and frequent manifestation, and gross hematuria is a common presenting symptom. On the basis of effectiveness, cost, and safety, ultrasound is the imaging modality most commonly used to make the diagnosis.

ADPKD is caused by mutations in the genes (*PKD1* and *PKD2*) that encode polycystin 1 and 2, which are located in primary cilia. Mutations affecting *PKD1* are more prevalent than those of *PKD2* and tend to have a worse prognosis. Genetic testing is not used routinely as a screening tool because current techniques identify only 70% of the hundreds of different *PKD1* and *PKD2* mutations. Generic treatment should include treatment of hypertension with ACE inhibitors and/or ARBs and maintaining a fluid intake of 2 to 3 L/day to reduce the risk of kidney stone disease. Specific therapies are targeted at reducing cyst development and enlargement. Estimates of cyst volume can be determined using magnetic resonance imaging techniques and changes documented over a relatively short period of time. This has allowed the performance of clinical trials of novel drugs such as the vasopressin V_2-receptor antagonists, tolvaptan, and antiproliferative drugs such as sirolimus. Tolvaptan has been shown in the 3-year clinical study (TEMPO3:4) to slow the increase in size of the cysts, reduce pain associated with cysts, and slow the rate of change in kidney function. Adverse events, primarily thirst and polyuria, and abnormalities in liver function tests are reported, and the drug is subject to a US Food and Drug Administration review.

Obstructive Uropathy

Benign prostatic hyperplasia (BPH) is one of the most common types of **obstructive uropathy** and is an almost universal finding in aging men. Among the most common symptoms are disorders of micturition, in particular increased frequency; in many cases this can progress to bladder outflow obstruction. Between 10% and 40% of men with bladder outflow obstruction caused by BPH present in acute retention. Approximately 5% of the men this group have high-pressure chronic retention of urine, which can result in upper urinary tract obstruction and consequently CKD as a result of glomerular and tubular damage.

Tubular Disease

Renal tubular acidoses and inherited tubulopathies are types of renal tubular disease.

Renal Tubular Acidoses

The **renal tubular acidoses (RTAs)** comprise a diverse group of both inherited and acquired disorders affecting either the proximal or distal tubule. They are characterized by a hyperchloremic, normal anion gap metabolic acidosis and urinary bicarbonate or hydrogen ion excretion inappropriate for the plasma pH. They are the result of either failure to retain bicarbonate or inability of the renal tubules to secrete hydrogen ion. Typically the GFR in RTA is normal or slightly reduced, and there is no retention of anions, such as phosphate and sulfate (as opposed to the acidosis of renal failure).

The three categories of RTA are distal (dRTA, type I); proximal (pRTA, type II); and type IV, which is secondary to aldosterone deficiency or resistance. The term "type III RTA" (mixed proximal/distal defect) has been abandoned because this is no longer considered a separate entity.

The finding of a hyperchloremic metabolic acidosis in a patient without evidence of gastrointestinal bicarbonate losses and with no obvious pharmacologic cause should prompt suspicion of an RTA. In addition to plasma electrolyte (including potassium) measurement, preliminary investigation should include measurement of urinary pH in a fresh early-morning urine sample. The finding of a urine pH greater than 5.5 in the presence of a systemic acidosis supports the diagnosis of dRTA.

Inherited Tubulopathies

The inherited tubulopathies comprise a heterogeneous set of rare disorders, including (1) **Bartter syndrome**, (2) **Gitelman syndrome**, (3) **Liddle syndrome**, (4) **pseudohypoaldosteronism type I**, (5) Dent disease, and (6) X-linked dominant hypophosphatemic rickets (previously known as vitamin D–resistant rickets). Most are characterized by electrolyte disturbances. In addition to these, general reasons to suspect a tubulopathy include (1) a familial disease pattern, (2) renal

impairment, (3) nephrocalcinosis, and (4) stone formation, especially if these should present at an early age. In cases in which a diuretic-sensitive channel is affected, these disorders will clearly mimic the effects of diuretic use (see discussion later in this chapter), and exclusion of covert use of diuretics is important. Although they are individually uncommon or rare, an awareness of these disorders is critical for the clinical biochemist considering the potential differential diagnoses in patients having electrolyte imbalances.

Diuretics

Diuretics are predominantly prescribed to treat either hypertension and/or disorders associated with fluid overload. All diuretics act by interfering with tubular reabsorption of sodium and/or chloride and therefore have accompanying effects on water retention. Different classes of diuretics act at different sites along the nephron. Classes include (1) loop, (2) thiazide, and (3) "potassium-sparing" diuretics. Many diuretics will cause hypokalemia to some degree, depending on the (1) potency, (2) dose, (3) duration of treatment, and (4) patient's underlying potassium balance.

Diabetes Insipidus

Diabetes insipidus (DI) is a disorder in which there is an abnormal increase in urine output, fluid intake, and often thirst. DI is due to the absence of an ADH effect, either because of impaired or failed secretion (cranial or central DI) or lack of end-organ response to ADH (nephrogenic DI). A further disorder, psychogenic polydipsia, or compulsive water drinking, has been known to present as DI. The investigation of DI involves first demonstrating, by undertaking a water deprivation test, that the patient is unable to form a concentrated urine. If the patient fails to concentrate urine during the test, then synthetic ADH is given to establish whether the defect in concentrating ability is at the renal or pituitary level.

Renal Calculi

Nephrolithiasis is a condition marked by the presence of renal calculi. Such calculi ("kidney stones") occur in the (1) renal pelvis, (2) ureter, and (3) bladder. Kidney stone formation is often considered to be a nutritional or environmental disease, linked to affluence, but genetic or anatomic abnormalities are also significant. Approximately 5% to 10% of the population of the western world are thought to have formed at least one kidney stone by the age of 70 years, and the prevalence of kidney stones may be increasing. In both males and females, the average age of first stone formation is decreasing. For most stone types, there is a male preponderance. The passage of a stone is associated with severe pain, called renal colic, which may last for 15 minutes to several hours and is commonly associated with nausea and vomiting.

The majority of kidney stones found in the western world are composed of one or more of the following substances: (1) calcium oxalate with or without phosphate (frequency 67%), (2) magnesium ammonium phosphate (12%), (3) calcium phosphate (8%), (4) urate (8%), (5) cystine (1% to 2%), and (6) complex mixtures of the above (2% to 3%). These poorly soluble substances crystallize within an organic matrix, the nature of which is not well understood. Most kidney stones are treated by **lithotripsy**, which entails crushing a calculus within the urinary system, followed at once by the washing out of the fragments.

Prostaglandins and Nonsteroidal Antiinflammatory Drugs in Kidney Disease

The prostaglandins are a series of C_{20} unsaturated fatty acid derivatives of cyclooxygenase (COX) on cell membrane arachidonic acid (see Chapter 23). The major renal vasodilatory prostaglandin is PGE_2, which is synthesized predominantly in the medulla of the kidney. The major vasoconstrictor prostaglandin is thromboxane A_2, which is produced primarily within the renal cortex. PGE_2 (1) increases renal blood flow rate, (2) inhibits sodium reabsorption in the distal nephron and collecting duct, and (3) stimulates renin release. These actions promote natriuresis and diuresis. In patients with CKD, renal PGE_2 excretion rates are three to five times higher than those in healthy subjects; therefore PGE_2 production represents a compensatory response to loss of nephron mass. Vasodilatory prostaglandins are synthesized following stimulation with renal sympathetic adrenergic and AII-dependent mechanisms to offset or modulate vasoconstriction. In the tubule, prostaglandins act as autocoids, exerting their effects locally near the site of synthesis.

NSAIDs have analgesic, antipyretic, and antiinflammatory effects. They also block the synthesis of COX products of arachidonic acid, which have a critical role in (1) renal hemodynamics, (2) control of tubular function, and (3) renin release. Analgesic nephropathy is a common cause of ESRD in a number of countries, reaching 10% in Switzerland and Australia, but it is essentially a preventable condition for which biochemical monitoring has proved useful. Older people demonstrate significant reductions of GFR within 1 week of ingestion of NSAIDs. Acute interstitial nephritis and nephrotic syndrome have been reported to occur with NSAIDs.

Toxic Nephropathy

A wide variety of nephrotoxins exist in the environment, in some cases associated with particular occupations (e.g., heavy metals, such as cadmium and lead). Both glomerular and tubulointerstitial damage may result from exposure to toxins; detection of both requires biochemical monitoring of GFR/serum creatinine concentration and tubular and glomerular proteinuria. Anatomic physiologic and biochemical features make the kidney susceptible to insult from a variety of medicinal and environmental agents. Factors contributing to the sensitivity of the kidney include its large blood flow, the concentration of filtered solutes during urine production, and the presence of a variety of xenobiotic transporters and metabolizing enzymes. **Toxic nephropathy** commonly occurs as a result of decreased renal perfusion because of precipitation within the tubule or due to direct toxic effects at the proximal tubule level. In some cases the conjugation of environmental

chemicals (e.g., mercury, cadmium) to glutathione and/or cysteine targets these chemicals to the kidney, where inhibition of renal function occurs through a variety of mechanisms that are not completely understood. Although some drugs can cause kidney damage in the presence of normal renal function, a far greater variety of drugs can cause problems in patients with kidney disease, predominantly because of accumulation resulting from decreased renal elimination. A list of drugs and environmental toxins commonly known to cause kidney damage is given in Table 35.3.

RENAL REPLACEMENT THERAPY

RRT includes dialysis procedures such as hemodialysis (HD), **peritoneal dialysis (PD)**, continuous hemofiltration (HF), and continuous **hemodiafiltration** (HDF). These techniques are used to temporarily or permanently remove toxic substances from the blood when the kidneys cannot satisfactorily remove them from the circulation. In addition, kidney transplantation has become an effective form of RRT. Extensive laboratory support is required by an RRT program (Table 35.4).

Dialysis

Dialysis is the process of separating macromolecules from ions and low-molecular-weight compounds in solution based on the difference in their rates of diffusion through a semipermeable membrane through which crystalloids can pass readily but colloids pass very slowly or not at all. Two distinct physical processes are involved: diffusion and ultrafiltration.

TABLE 35.3 Toxic Nephropathy: Causes, Patterns, and Markers

Compound Category	Drug/Toxin	Type of Renal Injury/Pathology	Biomarkers/Notes
Antibacterial agents	Aminoglycosides (e.g., neomycin, gentamicin, tobramycin, amikacin)	Acute tubular necrosis and interstitial nephritis. Nonoliguric AKI.	Plasma: ↓K, ↓Mg, ↓Ca Urine: ↑LMWP, ↑Glycosuria Nephrotoxicity major and common side effect
	Amphotericin	Initially distal tubular injury followed by medullary injury	Plasma: ↓K, ↑creatinine dRTA
Antiviral/antiprotozoal agents	Acyclovir	Nonoliguric AKI due to tubular obstruction and interstitial inflammation	Crystalluria and hematuria
	Pentamidine	Tubular toxicity	Plasma: ↓Mg, ↓Ca Urine: ↑Mg, ↑Ca
	Indinivar	Nephrolithiasis, irreversible kidney failure in some patients	Crystalluria and hematuria
Radiocontrast agents	Iothalamate, iodixanol	Oliguric or nonoliguric AKI, generally reversible. Proximal tubular damage.	↑Plasma creatinine after contrast administration
Antitumor drugs	Cisplatin	Irreversible dose-related and cumulative kidney failure. TIN with heavy proteinuria. Often AKI.	Urine: ↑Mg, ↑PO_4, tubular casts and ↑LMWP in early stages.
	Methotrexate	Nonoliguric AKI. Tubular atrophy and interstitial fibrosis.	Only seen in association with high-dose therapy
	Interleukin-2	Reversible AKI due to ↓RBF	Observed in up to 90% of cases of high-dose therapy
Other drugs	ACE inhibitors	Dramatic ↓GFR due to ↓ efferent arteriolar tone	Especially in the setting of bilateral renal artery stenosis. ↑Plasma creatinine and K.
	5-Aminosalicylic acid (e.g., mesalazine, olsalazine)	Occasional ATN and irreversible kidney damage	Tubular proteinuria
	Cyclooxygenase (COX)-2 inhibitors	Probably a similar pattern of renal injury to NSAIDs (see below)	
	Lithium	Distal tubular damage with nephrogenic diabetes insipidus ± dRTA	
	Penicillamine	Membranous glomerulopathy with NS, occasionally AKI	Proteinuria
	NSAIDs	Several forms of nephropathy identified, including (1) hemodynamically mediated AKI, (2) TIN ± NS, (3) salt and/or water retention, (4) hyperkalemia, and (5) CKD/ESRD ("analgesic nephropathy")	Depends on type of effect

Continued

TABLE 35.3 Toxic Nephropathy: Causes, Patterns, and Markers—cont'd

Compound Category	Drug/Toxin	Type of Renal Injury/Pathology	Biomarkers/Notes
Heavy metals	Cadmium	Subtle but irreversible TIN	Fanconi syndrome with RTA. ↑Urinary metallothionein.
	Gold	Membranous glomerulopathy but normal GFR maintained	Proteinuria <3.5 g/24 h
	Lead	Proximal tubular atrophy with interstitial fibrosis	Reversible Fanconi syndrome in children with acute poisoning. In lead workers, urinary proteinuria <2 g/24 h in association with ↑plasma urate, hypertension ± gouty arthritis.
	Mercury	Proximal tubular damage with ATN	Urine: ↑LMWP
Other environmental agents	Hydrocarbons (e.g., paints, dry-cleaning solvents)	ATN, chronic TIN, glomerulonephritis caused by renal cytochrome P450 metabolism of chloroform to toxic metabolites	Tubular proteinuria ± ↑plasma creatinine
	Paraquat, diquat	ATN secondary to ↓RBF due to shock and direct toxic effects of paraquat; intrinsic kidney damage due to production of reactive oxygen species	↑Plasma creatinine

ACE, Angiotensin converting enzyme; *AKI,* acute kidney injury; *ATN,* acute tubular necrosis; *CKD,* chronic kidney disease; *dRTA,* distal renal tubular acidosis; *ESRD,* end-stage renal disease; *GFR,* glomerular filtration rate; *LMWP,* low-molecular-weight proteinuria; *NS,* nephrotic syndrome; *NSAIDs,* nonsteroidal antiinflammatory drugs; *RBF,* renal blood flow; *TIN,* tubulointerstitial nephropathy.

The list of agents shown is not exclusive but illustrates the range of compounds that may affect the kidney. For further information, readers should consult: Palmer, B. F., & Henrich, W. L. (2004). Toxic nephropathy. In: B. M. Brenner (Ed.), *Brenner and Rector's the Kidney* (7th ed., Chapter 34) Philadelphia: Saunders.

TABLE 35.4 Laboratory Support for Dialysis Programs

Clinical Condition	Laboratory Tests
Acute Dialysis	
Dialysis disequilibrium	Urea and electrolytes, bicarbonate, calcium
Pyrexia	C-reactive protein, white cell count, blood cultures
Bleeding	Clotting screen, platelets
Chronic Dialysis	
Anemia	Ferritin, transferrin saturation, B12, folate Blood film, PTH, C-reactive protein
Sepsis	C-reactive protein, blood, urine specimens for microscopy, culture, and sensitivity
Nutrition	Albumin, phosphate
Cardiovascular disease risk	Lipid profile
Dialysis-related amyloid	β_2-microglobulin (not routinely measured)
CKD-MBD	Predialysis plasma calcium, phosphate (monthly in hemodialysis patients; every 3 months in peritoneal dialysis patients) Alkaline phosphatase PTH (at least every 3 months) Aluminum in patients receiving aluminum-based phosphate binders (every 3 months)
Adequacy of hemodialysis as assessed by urea clearance	Predialysis and postdialysis urea
Sepsis, abdominal pain in peritoneal dialysis	Microscopy and culture of peritoneal dialysate
Adequacy of peritoneal dialysis as assessed by weekly small solute clearance	Dialysate creatinine, urea
Peritoneal membrane characteristics assessed by peritoneal equilibration test (PET)	Plasma and dialysate glucose and creatinine

CKD, Chronic kidney disease; *MBD,* mineral; *PTH,* parathyroid hormone.

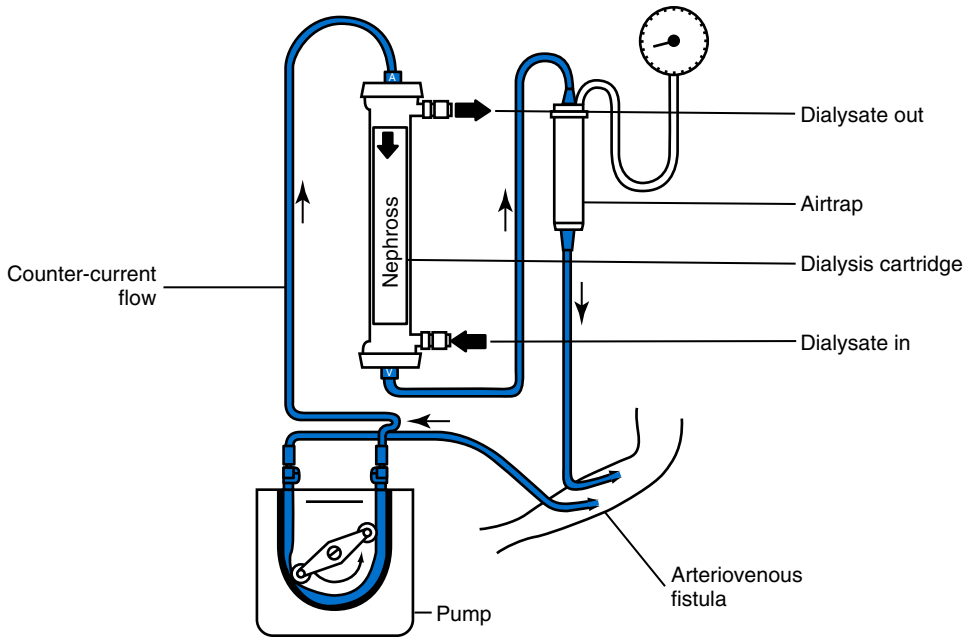


Fig. 35.6 A hemodialyzer setup.

No absolute recommendation of commencement of dialysis based on GFR alone can be made. KDIGO suggest that dialysis be initiated when one or more of the following are present: symptoms or signs attributable to kidney failure (serositis, acid-base or electrolyte abnormalities, pruritus), inability to control volume status or blood pressure, a progressive deterioration in nutritional status refractory to dietary intervention, or cognitive impairment. This often but not invariably occurs in the GFR range between 5 and 10 mL/min per 1.73 m^2. Not all individuals will be suitable for RRT, and in this setting it is important that the multidisciplinary team facilitate care for people on the "conservative management" pathway.

Dialysis procedures include (1) HD and (2) PD.

Hemodialysis

HD is the method most commonly used to treat advanced and permanent kidney failure. Clinically it is considered the default therapy that is used in patients unsuitable for the alternate modalities of PD and kidney transplantation. Operationally it involves connecting the patient to a circuit into which his or her blood flows to and from a semipermeable large-surface-area membrane, the hemodialyzer (Fig. 35.6). After filtration to remove wastes and extra fluid, the cleansed blood is returned to the patient. This is a complicated and inconvenient therapy requiring a coordinated effort from a health care team that includes the patient, nephrologist, dialysis nurse, dialysis technician, dietitian, and others. Patients are dialyzed in home- or hospital-based units, with dialysis usually performed three times a week for sessions lasting between 3 and 5 hours. This dialysis schedule is largely empirical insofar as it reconciles adequate treatment with breaks between treatments to provide the patient with a reasonable quality of life.

HD relies on good vascular access, preferably via a surgically created upper limb arteriovenous fistula (AVF), to enable blood to be pumped around the extracorporeal circuit at a rate in excess of 300 mL/min.

Fluid management on HD is crucial for patient well-being and survival. Because conventional dialysis is based on a thrice weekly schedule, fluid is accumulated by the patient between dialysis sessions. Many patients are anuric or at least oliguric; therefore unrestricted fluid intake would result in fluid overload and complications of pulmonary edema and hypertension. Patients receiving HD are advised to restrict fluid intake to 1 L/day or so. This allowance is recommended to the individual patient by the dialysis nursing staff and the dietitian to ensure that adequate nutrition is maintained. Nevertheless many patients find the fluid restriction very difficult to maintain; therefore large weight gains between dialysis sessions are a common occurrence.

Assessment of adequacy of hemodialysis. Assessment of adequacy of dialysis treatment for individual patients in the clinical setting includes consideration of the patient's well-being, cardiovascular risk, nutritional status, and degree of achievable ultrafiltration. It also includes estimates of a number of laboratory parameters such as hemoglobin, phosphate, and albumin and clearance of the small solutes urea and creatinine. Although a full description of adequacy is beyond the scope of this text, in practice a simple calculation may be performed to obtain an estimate of dialysis adequacy: the urea reduction ratio (URR). The URR is the percentage fall in plasma urea attained during a dialysis session and is measured as follows:

$$[(\text{Predialysis \{urea\}} - \text{Postdialysis \{urea\}}) / (\text{Predialysis \{urea\}})] \times 100\,\%$$

Observational studies in populations of dialysis patients have shown that variations in URR are associated with major differences in mortality.

Peritoneal Dialysis

Peritoneal dialysis (PD) is a type of dialysis in which dialysate is passed into the patient's peritoneal cavity, with the peritoneum then employed as the dialysis membrane (Fig. 35.7). Use of PD varies among countries, depending on access to

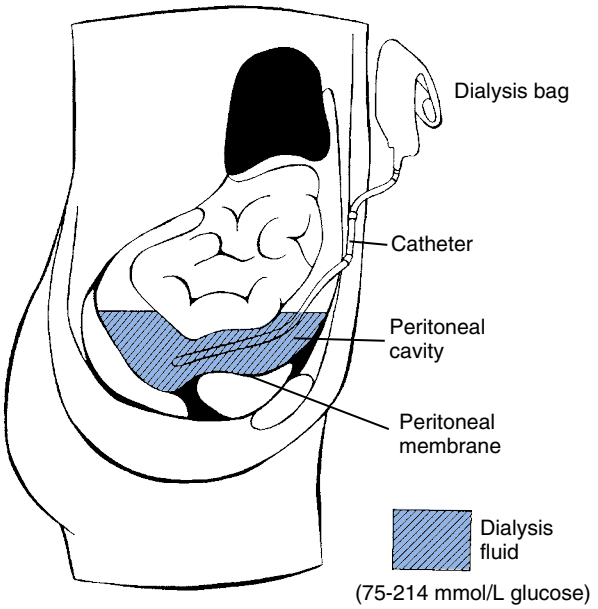


Fig. 35.7 Diagrammatic sketch of peritoneal dialysis. To convert glucose concentration in mmol/L to mg/dL, multiply by 18. (Redrawn from Nolph, K. D. [1989]. Peritoneal anatomy and transport physiology. In: J. F. Maher [Ed.], *Replacement of renal function by dialysis* [3rd ed., Chapter 23]. Dordrecht, The Netherlands: Kluwer Academic Publishers/Springer.)

HD. For example, in 2013 in the United Kingdom, 3 months after entering an RRT program, 20% of incident patients were receiving PD (compared with 66% receiving HD and 10% with a functioning transplant); whereas in Mexico, 90% of patients received PD.

Operationally, PD uses the patient's own peritoneal membrane (surface area approximately 2 m^2), across which fluids and solutes are exchanged between the peritoneal capillary blood and the dialysis solution placed in the peritoneal cavity. Fluid removal (ultrafiltration) is achieved by using dialysis fluids containing high concentrations of dextrose acting as an osmotic agent; as dextrose passes across the peritoneal membrane, the concentration gradient diminishes and the rate of fluid removal decreases. Conventional therapies use four daily exchanges of approximately 2 L of fluid, with approximately 10 L of spent dialysate generated (including ultrafiltration). RRF is critical to the success of PD because only a few milliliters per minute can contribute substantially to urea clearance and C_{Cr}, with each additional milliliter resulting in an extra 10 L of clearance per week. Practical reasons for opting for PD include preservation of RRF and vascular access sites, a home treatment facilitating increased patient autonomy, flexibility as to where the treatment can be administered, and ease of self-treatment, with lower capital costs involved. Blood pressure control and extremes of fluid shifts are not as problematic as those that occur on conventional HD.

Automated PD is now widely available. It requires a programmable machine to regulate flow, dwell time, and drainage, and may be performed at night. Solute clearance can be increased by leaving fluid in the peritoneum during the day and by performing an additional daytime exchange.

Peritonitis, a particular complication of PD, typically presents with a cloudy dialysate effluent and abdominal pain. Additional features such as vomiting and a high temperature suggest serious infection. Blood and dialysate samples should be taken for urgent microbiologic analysis and antibiotics administered via the dialysis catheter directly into the peritoneum. If antibiotic treatment fails, then the catheter is removed and the patient is converted to HD. In the majority of cases, the episode of peritonitis responds to treatment and PD can continue, although it is likely that repeated episodes will cause scarring and fibrosis of the peritoneal membrane, with permanent loss of ultrafiltration. Long-term serious complications may occur, such as sclerosing encapsulating peritonitis caused by adhesions and peritoneal thickening encasing the peritoneal contents and causing bowel obstruction. This unusual condition is associated with increased frequency of peritonitis episodes and longer duration of PD.

Malnutrition in Dialysis Patients

Dialysis patients with ESRD tend to have a poor appetite. Protein metabolism is altered in the setting of chronic acidosis and low-grade inflammation. These factors in combination place patients at risk of protein and energy malnourishment. Nutritional screening is recommended in dialysis patients. Such screening may involve measurement of weight, a recent history of edema-free weight loss, the body mass index, and subjective global assessment. Serum albumin is often used as a marker of malnutrition, even though it is a relatively poor nutritional marker. However, good evidence indicates that the lower the albumin concentration, the worse the long-term prognosis.

Kidney Transplantation

Kidney transplantation is the most effective form of RRT in terms of long-term survival and quality of life. Approximately, 100,000 patients are on the waiting list for a kidney transplant, and almost 17,000 kidney transplants are performed annually in the United States. Median waiting time for a listed patient depends on his or her age, with children given priority listing. Although patients with kidney failure should have equitable access to kidney transplantation, currently only 23% of adult patients on dialysis in the United Kingdom are on the active transplant waiting list. Waiting time spent on dialysis has been shown to be an important factor in determining mortality. The very best outcomes are achieved following preemptive (i.e., before dialysis has become necessary) live donor transplants. Typically, graft and patient survival at 1-year are in excess of 90%. Long-term graft survival remains a major problem, with half of transplants failing within 14 years, usually as a result of chronic allograft injury or death with a functioning graft. Transplantation medicine provides a constant challenge to balance the immunologic risk of damage to the allograft (rejection) versus the well-being of the recipient while avoiding excess immunosuppression, as the latter increases the likelihood of opportunistic infection and malignancy. In addition, many of the powerful immunosuppressive drugs have idiosyncratic side-effect profiles.

Preoperative Assessment

Some patients are considered unsuitable candidates for transplantation (Box 35.2). Laboratory assessment includes indicators of general operative health (e.g., electrolytes, acid-base status, clotting profile, full blood cell count, cross-matching). In addition, a full screen for infectious diseases—particularly cytomegalovirus (CMV), Epstein-Barr virus, hepatitis B and C, varicella-zoster virus, and human immunodeficiency virus status—is undertaken; these infections can be activated by immunosuppressive therapy.

The Operation

The donor kidney is usually placed extraperitoneally in the right or left iliac fossa. Anastomoses are constructed joining the transplant renal artery and vein to the recipient's respective iliac vessels. The ureter is joined to the bladder. The recipient's native kidneys are left in situ in almost all cases. Living donor kidneys can be retrieved through open surgery or with the aid of laparoscopic techniques. The workup for a live donor transplant is beyond the scope of this chapter.

Postoperative Assessment

During the initial postoperative phase of 1 to 2 weeks, careful monitoring of serum creatinine and urine output is required to monitor graft function. Most grafts produce measurable amounts of urine within a matter of hours, and this is a clear sign of a functioning graft; however, in a certain proportion, perhaps 5% to 10% of cases, primary nonfunction is apparent. In this subgroup, continuing dialysis support is necessary. In some patients the condition resolves without treatment, but in others a percutaneous kidney biopsy may be necessary to establish whether the graft is still viable and what form of therapy should be initiated. In otherwise uncomplicated cases, the serum creatinine concentration falls rapidly postoperatively (Fig. 35.8). Early allograft rejection episodes are suspected if the creatinine does not fall to the expected level or if there is a creatinine concentration indicating allograft dysfunction.

Immunosuppression and therapeutic drug monitoring. The currently used immunosuppressive drugs and regimens have potentially numerous and serious side effects; these are summarized in Table 35.5.

Cyclosporin is insoluble and is presented for clinical use as a microemulsion. It has a narrow therapeutic window; therefore in clinical transplantation it is important to monitor the blood concentration frequently. The most widely accepted practice is to monitor the "trough" concentration (C-0) just before the next dose. Accepted trough concentrations range from 100 to 300 μg/L. The highest concentrations are targeted during the induction phase of treatment for 2 to 3 months; subsequently lower maintenance concentrations are desirable. The trough concentration within the blood may not provide a truly accurate guide to total drug exposure because wide variation in absorption is seen over the first 2 to 4 hours following dosing. This is important in that most of the pharmacodynamic effects of cyclosporin occur within 2 hours.

BOX 35.2 Exclusion Criteria for Consideration for a Kidney Transplant

- Serious concomitant illness (particularly if likely to shorten life expectancy or to be exacerbated by immunosuppressive treatment)
- Active malignancy
- Inoperable ischemic heart disease
- Severe chronic lung disease
- Active systemic infection (e.g., tuberculosis)
- Active immunologic disease
- Severe irreversible hepatic disease
- Severe peripheral vascular disease
- Severe obesity (body mass index >40 kg/m^2)
- Lower urinary tract dysfunction not amenable to surgical repair
- Substance abuse
- Significant psychiatric disturbance

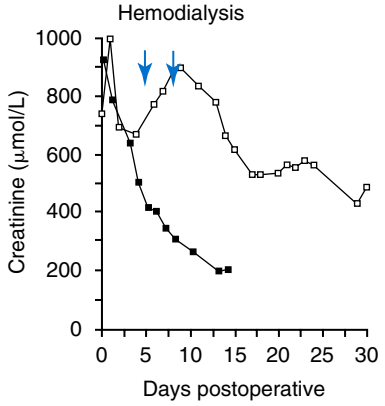


Fig. 35.8 Posttransplantation biochemical profile. *Open squares* represent the course of a patient who experienced an early rejection episode (confirmed by biopsy, ↓) and requires initial hemodialysis support. *Solid squares* represent the typical profile of an uncomplicated transplant recipient. To convert creatinine concentration in μmol/L to mg/dL, multiply by 0.011.

Studies from Canada suggest that trough concentrations do not reflect clinical outcomes in terms of acute rejection rates, although high trough concentrations were associated with increased nephrotoxicity. Among kidney transplant patients, the trough level of tacrolimus is correlated with acute rejection episodes and nephrotoxicity. Trough concentrations also guide sirolimus therapy (see Chapter 30).

MMF is morpholinoethyl ester of mycophenolic acid (MPA), a potent and reversible inhibitor of inosine monophosphate dehydrogenase isoform 2 (IMPDH); it has become the single most used immunosuppressant in solid organ transplantation. Excellent results have been obtained with a fixed-dose regimen. IMPDH is a target for immunosuppression because lymphocytes depend on the de novo guanosine nucleotide synthesis pathway for DNA synthesis and cell division. MMF, because it is a prodrug of MPA, is rapidly absorbed following an oral dose and is deesterified to MPA, which is highly protein-bound. Free MPA concentrations

TABLE 35.5 Noninfectious Complications of Immunosuppressant Drugs

Drug	Drug Dose	Target Therapeutic Range[a]	Toxicity Profile
Corticosteroids—e.g., prednisone	Dose depends on weight of patient and time since transplant. Typically 40 mg daily during first week and tapering to 5 mg at 6 months and withdrawal at 12 months.	Not appropriate.	Increased risk of developing diabetes mellitus Deterioration in diabetes control Osteopenia Osteoporosis Psychosis Fat redistribution Hypertension Dyslipidemia Cataracts Weight gain
Calcineurin Inhibitors			
Cyclosporin	Variable; depends on weight, time since transplant, and achieved drug concentration. Given in two divided doses and predose trough concentration measured in morning blood sample.	200–300 μg/L for first 3–12 months. Thereafter, aim for 100 μg/L.	Nephrotoxicity Hypertension Neurotoxicity Hemolytic-uremic syndrome Tubular electrolyte abnormalities (hypophosphatemia, hypomagnesemia, hyperkalemia) Hirsutism Gingival hyperplasia Bone pains Dyslipidemia
Tacrolimus		Level depends on time since transplant. Typical early (less than 3 months) targets are 8–12 μg/L and thereafter 4–8 μg/L.	As for cyclosporin, except no hirsutism or gingival hyperplasia Increased risk of diabetes mellitus Cardiomyopathy (children) Alopecia
Mycophenolate mofetil	Initially 2 g daily in divided doses.	Not routinely measured.	Abdominal pain Diarrhea Myelosuppression
Sirolimus	Dose depends on weight and achieved drug concentration. The drug is administered once daily.	Level depends on time since transplant. Typical early (less than 3 months) targets are 8–12 μg/L and thereafter 4–8 μg/L.	Lymphocele (a fluid-filled collection near to transplanted kidney) Thrombocytopenia Hyperlipidemia
Azathioprine	Usual starting dose of 2 mg/kg body weight in a single daily dose.	Levels not measured. Because the enzyme thiopurine methyltransferase (TPMT) metabolizes azathioprine, the risk of myelosuppression is increased in patients with low activity of the enzyme. Enzyme activity may be determined prior to commencing treatment, and full blood counts are performed for several weeks following commencement of drug.	Myelosuppression Severe interaction if used with allopurinol (treatment for gout)
Selected Biological Agents			
AntiCD25 monoclonal antibodies: basiliximab and daclizumab	Given at time of transplant and once thereafter.		Very well tolerated
Polyclonal antithymocyte globulin (ATG) and antilymphocyte globulin (ALG), and monoclonal OKT3	Given in response to refractory rejection episodes in selected patients.		Increased risk of malignancy, posttransplant lymphoproliferative disease Hypersensitivity reactions

[a]These are not recommendations but are illustrative and will vary among centers.

determine the level of immunosuppressive activity, and this can be affected by hypoalbuminemia and renal insufficiency. Therapeutic drug monitoring of MPA is possible using high-performance liquid chromatography (HPLC) and immunoassay techniques. However, it has not been universally accepted in kidney transplantation programs because prospective studies have given conflicting information of its value.

In summary, long-term graft failure is a major problem, and graft loss accounts for the return of increasing numbers of patients to dialysis. The most common cause of graft loss is death with a functioning graft. Kidney failure carries a considerable burden of cardiovascular morbidity. Although some risk factors, such as volume overload and anemia, are improved following transplantation, others, including dyslipidemia and hypertension, persist. The drugs used to prevent rejection can exacerbate these. Challenges to the nephrology community are complex and include improving access to transplantation, reducing side effects of the powerful drugs used to prevent rejection, and reducing cardiovascular risk profiles for individual patients.

Simultaneous Pancreas-Kidney Transplantation

Patients with kidney failure and diabetes mellitus, predominantly type 1, but increasingly certain patients with type 2, and limited secondary complications of diabetes may be considered for simultaneous pancreas and kidney (SPK) transplantation. Patients tend to be younger than kidney-only recipients (e.g., aged between 20 and 40 years). Patient survival now reaches over 95% at 1 year posttransplant and over 83% after 5 years.

The surgical technique for SPK involves whole-organ pancreas transplantation with the duodenal segment draining either into the urinary bladder through a duodenocystostomy or, more commonly today, enterically via an anastomosis between the graft duodenal segment and the recipient's small bowel. The kidney is attached as usual to the iliac vessels, and the donor ureter is inserted into the bladder separately. Postoperatively, blood glucose concentrations are monitored closely, and intravenous insulin is given as necessary. Exocrine pancreatic secretion can be measured in the urine for bladder-drained pancreas allografts. The major fear is rejection, and a number of parameters are monitored, including plasma glucose, amylase, lipase, and 12- or 24-hour urinary amylase (again for bladder-drained allografts). For patients with bladder drainage, enteric conversion may be required for refractory problems such as dehydration, metabolic acidosis, chronic urethritis, urinary tract infection, and recurrent reflux pancreatitis. Because of high fluid, bicarbonate, and electrolyte losses into the urine in these patients, the need for supplementation is increased in SPK recipients. There is a long-term need for high-dose oral sodium bicarbonate supplementation in bladder-drained pancreatic transplantation because of exocrine secretory losses. Hyperamylasemia is common postoperatively and may or may not signify allograft rejection.

REVIEW QUESTIONS

1. It is difficult to measure the GFR of a kidney directly; therefore which one of the following is assessed to determine GFR?
   a. Renal blood flow
   b. Renal threshold
   c. Serum creatinine
   d. Urine albumin
2. Which one of the following tests evaluates renal tubular (including loop of Henle) function?
   a. Urine osmolality
   b. Inulin clearance
   c. Urine albumin
   d. Urine protein
3. The structural and functional unit of the kidney is the nephron. What structures make up a nephron?
   a. Only the structures located in the kidney cortex
   b. The glomeruli, tubules, and associated blood vessels
   c. The ureters, bladder, and urethra
   d. Only the structures located in the kidney medulla
4. Pyelonephritis is
   a. caused by a lack of intrinsic factor.
   b. the destruction of kidney glomeruli by immune complexes.
   c. a tumor of the stomach.
   d. inflammation of both the lining of the renal pelvis and the parenchyma of the kidney, especially due to bacterial infection.
5. Which one of the following hormones synthesized in cells of the JGA is involved in control of blood pressure through its action on angiotensinogen?
   a. Erythropoietin
   b. Antidiuretic hormone
   c. Renin
   d. Aldosterone
6. Regarding laboratory findings, uremia/azotemia specifically refers to
   a. reduced renal function.
   b. elevated nitrogenous compounds in blood.
   c. elevated serum proteins in blood.
   d. decreased urine albumin.
7. The most common glomerular disease caused by damage to the glomerular membrane from the deposition of immune complexes is
   a. IgA nephropathy.
   b. chronic glomerulonephritis.
   c. uremic syndrome.
   d. pyelonephritis.
8. A disorder in which there is an abnormal increase in urine output, fluid intake, and often thirst and that is caused by the absence of antidiuretic hormone is
   a. diabetes mellitus.
   b. IgA nephropathy.
   c. diabetes insipidus.
   d. nephrolithiasis.

9. Which one of the following hormones affects water reabsorption in the proximal tubule, the loop of Henle, the distal tubule, and the collecting duct of the kidney?
   a. Aldosterone
   b. Renin
   c. 1,25(OH_2) vitamin D_3
   d. Antidiuretic hormone

10. A drug prescribed to an individual to treat hypertension and/or disorders associated with fluid overload is referred to as
   a. an ACE inhibitor.
   b. a diuretic.
   c. cystatin C.
   d. an exogenous marker.

SUGGESTED READINGS

Chapman, A. B., Devuyst, O., Eckardt, K. U., et al. (2015). Autosomal-dominant polycystic kidney disease (ADPKD): executive summary from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. *Kidney International*, *88*, 17–27.

Coresh, J., Selvin, E., Stevens, L. A., et al. (2007). Prevalence of chronic kidney disease in the United States. *The Journal of the American Medical Association*, *298*, 2038–2047.

De Vriese, A. S., Glassock, R. J., Nath, K. A., Sethi, S., & Fervenza, F. C. (2017). A proposal for a serology-based approach to membranous nephropathy. *Journal of the American Society of Nephrology*, *28*, 421–430.

Devoe, D. J., Wong, B., James, M. T., et al. (2016). Patient education and peritoneal dialysis modality selection: a systematic review and meta-analysis. *American Journal of Kidney Diseases*, *68*, 422–433.

Diabetes Control and Complications Trial Research Group. (1993). The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. The Diabetes Control and Complications Trial Research Group. *The New England Journal of Medicine*, *329*, 977–986.

Haraldsson, B., Nystrom, J., & Deen, W. M. (2008). Properties of the glomerular barrier and mechanisms of proteinuria. *Physiological Reviews*, *88*, 451–487.

Kanagasundaram, N. S. (2015). Pathophysiology of ishaemic acute kidney injury. *Annals of Clinical Biochemistry*, *52*, 193–205.

Kidney Disease Improving Global Outcomes. (2012). KDIGO clinical practice guideline for anemia in chronic kidney disease. *Kidney International*, *2*, 279–335.

Kidney Disease Improving Global Outcomes. (2013). Clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney International*, *3*, 1–150.

Kidney Disease Improving Global Outcomes (KDIGO). (2012). KDIGO clinical practice guideline for glomerulonephritis. *Kidney International*, *2*(Suppl.), 139–274.

Kidney Disease Improving Global Outcomes. (2017). KDIGO clinical practice guideline update for the diagnosis, evaluation, prevention, and treatment of chronic kidney disease-mineral and bone disorder (CKD-MBD). *Kidney International. Supplement*, *7*, 1–59.

Lamb, E. J., Stevens, P. E., & Deeks, J. J. (2015). What is the best glomerular filtration marker to identify people with chronic kidney disease most likely to have poor outcomes? *BMJ*, *350*, g7667.

Loudon, K. W., & Fry, A. C. (2014). The renal channelopathies. *Annals of Clinical Biochemistry*, *51*, 441–458.

Matsushita, K., van der Velde, M., Astor, B. C., et al. (2010). Association of estimated glomerular filtration rate and albuminuria with all-cause and cardiovascular mortality in general population cohorts: a collaborative meta-analysis. *Lancet*, *375*, 2073–2081.

National Institute for Health and Care Excellence. (2013). *Acute kidney injury: prevention, detection and management of acute kidney injury up to the point of renal replacement therapy (CG169)*. https://www.nice.org.uk/.

National Kidney Foundation. (2006). Clinical practice guidelines for hemodialysis adequacy, update 2006. *American Journal of Kidney Diseases*, *48*(Suppl. 1), S2–S90.

Parikh, C. R., & Mansour, S. G. (2017). Perspective on clinical application of biomarkers in AKI. *Journal of the American Society of Nephrology*, *28*, 1677–1685.

United Kingdom Prospective Diabetes Study (UKPDS) Group. (1998). Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). *Lancet*, *352*, 837–853.

Wanner, C., & Tonelli, M. (2014). KDIGO clinical practice guideline for lipid management in CKD: summary of recommendation statements and clinical approach to the patient. *Kidney International*, *85*, 1303–1309.

Webster, A. C., Nagler, E. V., Morton, R. L., & Masson, P. (2017). Chronic kidney disease. *Lancet*, *389*, 1238–1252.

Physiology and Disorders of Water, Electrolyte, and Acid-Base Metabolism

Mark A. Cervinski, Christopher McCudden, Marc Berg, Mitchell G. Scott

OBJECTIVES

1. Define the following:
   - Acid-base balance
   - Aldosterone
   - Anion gap
   - Antidiuretic hormone
   - Compensation
   - Depletional hyponatremia
   - Dilutional hyponatremia
   - Hyper- and hypovolemia
   - Interstitial fluid
   - Solute diuresis
2. State the significance of and describe the total body water (TBW) distribution between compartments, including approximate volume and composition of each compartment, electrolyte distribution, and active/passive transport for ion exchange; list the hormonal regulators of water and sodium.
3. For each of the following electrolytes, state and describe regulatory mechanisms, disorders, and causes/effects of these disorders:
   - Chloride
   - Potassium
   - Sodium
4. State the Henderson-Hasselbalch equation, and explain each of its components; explain its relation to compensatory mechanisms in acid-base disturbances.
5. Categorize the physiological buffer systems relative to their specific roles in the regulation of blood pH.
6. Explain the specific respiratory and renal mechanisms important in the regulation of acid-base balance.
7. For each of the acid-base imbalances listed below, list the causes and state the primary deficit, compensatory mechanisms, and laboratory values obtained with each:
   - Metabolic acidosis
   - Metabolic alkalosis
   - Respiratory acidosis
   - Respiratory alkalosis
8. State the formula for and calculate the anion gap; discuss the clinical usefulness of anion gap, including eight causes of an increased anion gap and four causes of normal anion gap acidosis.
9. Compare chloride responsive, chloride resistant, and exogenous base metabolic alkalosis, including causes and laboratory values.
10. Assess and solve case studies related to electrolyte disturbances and acid-base balance disorders.

KEY WORDS AND DEFINITIONS

Acid-base balance The homeostatic maintenance of acids and bases within the body to achieve a physiological pH (approximately 7.40).

Acidemia An arterial blood pH less than 7.35.

Alkalemia An arterial blood pH greater than 7.45.

Anion gap (AG) The difference between the serum sodium concentration and the sum of the serum chloride and bicarbonate concentrations; the anion gap is high in some forms of metabolic acidosis.

Chronic obstructive pulmonary disease (COPD) Any disorder characterized by persistent or recurring obstruction of bronchial air flow, such as chronic bronchitis, asthma, or pulmonary emphysema.

Compensated acidosis A state of acidosis in which the pH of the blood has been returned toward normal by compensatory mechanisms.

Depletional hyponatremia A condition characterized by low plasma concentration of sodium associated with low total body sodium and normal blood volume; also called euvolemic hyponatremia.

Diabetes insipidus A condition associated with increased thirst, polyuria, and large volume of dilute urine, exceeding 3 L/day, and causing dehydration.

Dilutional hyponatremia A condition characterized by low plasma concentration of sodium resulting from loss of sodium from the body with nonosmotic retention of water.

Extracellular fluid (ECF) A general term for all the body fluids outside the cells, including the interstitial fluid, plasma, lymph, and cerebrospinal fluid (CSF).

Henderson-Hasselbalch equation An equation that defines the relationship between pH, bicarbonate, and the partial pressure of dissolved carbon dioxide gas.

Hyperkalemia A concentration of serum potassium above 5.0 mmol/L.
Hypernatremia A concentration of serum sodium above 150 mmol/L.
Hypervolemia Abnormal increase in the volume of circulating fluid (plasma) in the body.
Hypokalemia A concentration of serum potassium below 3.5 mmol/L.
Hyponatremia A concentration of serum sodium below 136 mmol/L.
Hypovolemia Abnormally decreased volume of circulating fluid (plasma) in the body.
Intracellular fluid (ICF) The portion of the total body water (TBW) with its dissolved solutes that is within the cell membranes.
Ketoacidosis A condition characterized as acidosis accompanied by the accumulation of ketone bodies (ketosis) in the body tissues and fluids.
Metabolic acidosis A pathological process that leads to the accumulation of acid that lowers the bicarbonate concentration and decreases pH.
Metabolic alkalosis A pathological process that leads to the accumulation of base that raises the bicarbonate concentration and increases pH.
Respiratory acidosis A pathological process that leads to the accumulation of carbon dioxide that raises the PCO_2 and decreases pH; usually caused by emphysema or hypoventilation.
Respiratory alkalosis A pathological process that leads to the excessive elimination of carbon dioxide, which lowers the PCO_2 and increases the pH; caused by hyperventilation.
Sodium–hydrogen exchanger (Na^+–H^+ exchanger) A membrane protein that is primarily responsible for maintaining the balance of sodium; also called the sodium–hydrogen antiporter.
Total body water (TBW) Any of various estimates of the water content of the human body, taking into consideration the person's height, weight, and age.

The human body is in a dynamic state of flux as fluids and electrolytes are constantly being gained and lost. Homeostasis within a narrow window is necessary for life, and the body must defend against excessive gains or losses. These homeostatic systems require the interaction of multiple organ systems and a variety of chemical buffers and highly specialized mechanisms of the lungs and kidneys work together to regulate water, electrolytes, and pH between and within intracellular and extracellular compartments. Perturbations in the dynamic equilibria of water, electrolytes, and pH may arise from external (e.g., trauma, changes in altitude, ingestion of toxic substances) or internal (e.g., normal metabolism, disease state) sources.

Water accounts for 60% of body weight in adolescents and adult males and ≈55% for adult females. As depicted in Fig. 36.1, approximately two-thirds of **total body water (TBW)** is distributed into the **intracellular fluid (ICF)** compartment, and one-third exists in the **extracellular fluid (ECF)** compartment. The ECF may be further subdivided by the capillary endothelium into interstitial (≈75% of ECF) and intravascular (≈25% of ECF) compartments. The average adult has ≈5 L of blood volume (intravascular compartment) and a plasma volume of ≈3.0 L when the hematocrit is 40%. Although fluid from other clinically relevant ECF compartments (e.g., cerebrospinal fluid [CSF], urine) are analyzed in the clinical laboratory, most laboratory tests used to determine hydration status and electrolyte and acid-base status are performed on samples from the *intravascular* compartment.

The minimum daily requirement for water can be estimated from renal (1200 to 1500 mL in urine) and "insensible" losses (≈400 to 700 mL due to evaporation from the skin and respiratory tract). Activity, environmental conditions, and disease all have dramatic effects on daily water (and electrolyte) requirements. On average, an adult must take in ≈1.5 to 2.0 L of water daily to maintain fluid balance. Table 36.1 lists common causes and clinical manifestations of expansion and contraction of the ECF compartment.

WATER AND ELECTROLYTES—COMPOSITION OF BODY FLUIDS

The primary cationic (positively charged) electrolytes are sodium (Na^+), potassium (K^+), calcium (Ca^{2+}), and magnesium (Mg^{2+}), whereas the anions (negatively charged) include chloride (Cl^-), bicarbonate (HCO_3^-), phosphate (HPO_4^{2-}, $H_2PO_4^-$), sulfate (SO_4^{2-}), organic ions such as lactate, and negatively charged proteins. Electrolyte concentrations of the body fluid compartments are shown in Table 36.2. Plasma or serum Na^+, K^+, Cl^-, and HCO_3^- concentrations, commonly analyzed in an *electrolyte profile*, provide the most relevant information about the osmotic, hydration, and acid-base status of the body. Although hydrogen (H^+) is a cation, its concentration is approximately 1 million-fold lower in plasma than the major electrolytes listed in Table 36.2 (10^{-9} mol/L vs. 10^{-3} mol/L) and is negligible in terms of osmotic activity.

In order to maintain electroneutrality, any increase in the concentration of one anion is accompanied by a corresponding decrease in other anions, by an increase in one or more cations or a combination of both. The converse is true for any decrease in anion concentration. In the case of polyvalent ions (e.g., Ca^{2+}, Mg^{2+}), it is important to distinguish between the substance concentration of the ion itself and the concentration of the ion charge. Thus, although the concentration of total calcium ions in normal plasma is ≈2.5 mmol/L, the concentration of the total calcium ion *charge* is 5.0 mmol/L (also called 5 milliequivalents per liter [mEq/L]).

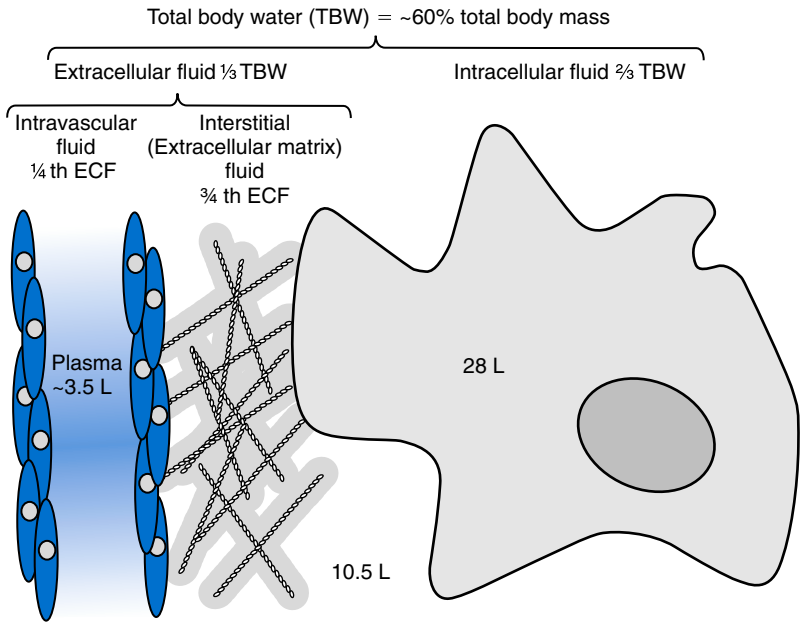


Fig. 36.1 Volume and distribution of total body water (TBW). The intracellular and extracellular fluid compartments (ICF and ECF, respectively) are separated by cellular plasma membranes, and within the ECF, interstitial and intravascular fluids are separated by the capillary endothelium (blue cells). The volumes indicated represent water and not total volume. Endothelial cells, *blue*; interstitial cell, *gray*; collagen matrix fibers, *black cables*.

TABLE 36.1 Causes and Clinical Manifestations of Changes in Extracellular Fluid Volume

	Clinical Manifestations	Causes
ECF loss	Thirst, anorexia, nausea, light-headedness, orthostatic hypotension, syncope, tachycardia, oliguria, decreased skin turgor and "sunken eyes," shock, coma, death	Trauma (and other causes of acute blood loss), "third-spacing" of fluid (e.g., burns, pancreatitis, peritonitis), vomiting, diarrhea, diuretics, renal or adrenal (i.e., sodium wasting) disease
ECF gain	Weight gain, edema, dyspnea (secondary to pulmonary edema), tachycardia, jugular venous distention, portal hypertension (ascites), esophageal varices	Heart failure, cirrhosis, nephrotic syndrome, iatrogenic (intravenous fluid overload)

ECF, Extracellular fluid.

TABLE 36.2 Electrolyte and Water Composition of Body Fluid Compartments

Component	Plasma	Interstitial Fluid	Intracellular Fluid[a]
Volume, H_2O (TBW = 42 L)	3.5 L	10.5 L	28 L
Na^+	140	145	12
K^+	4	4	156
Ca^{2+}	2.4	2–3	0.3
Mg^{2+}	1	0.5–1	13
Trace elements	1	—	—
Cl^-	103	114	4
HCO_3^-	27	31	12
Protein$^-$	16	—	55
Organic acids$^-$	5	—	—
HPO_4^{2-}	1	—	—
SO_4^{2-}	0.5	—	—

[a]These values are derived from skeletal muscle.
All electrolyte values are expressed in millimoles per liter of *fluid*. Because the H_2O content of plasma is ≈93% by volume, the corresponding electrolyte concentrations in plasma water are ≈10% higher.
TBW, Total body water.

EXTRACELLULAR AND INTRACELLULAR COMPARTMENTS

Plasma

Plasma generally has a volume of 1300 to 1800 mL/m^2 of body surface and constitutes approximately 5% of the body volume (≈3.5 L for a 66 kg subject). Total body volume is derived from body mass by using an estimated body density of 1.06 kg/L. Table 36.2 describes the electrolyte composition of plasma. The mass concentration of water in normal plasma is approximately 0.933 kg/L, depending on the protein and lipid content (see "Electrolyte Exclusion Effect" in Chapter 24).

Interstitial Fluid

Interstitial fluid is essentially an ultrafiltrate of blood plasma. When all extracellular spaces except plasma are included, the volume accounts for about 26% (10.5 L) of the total body

volume (see Fig. 36.1). Plasma is separated from the interstitial fluid by capillary endothelial lining, which acts as a semipermeable membrane and allows the passage of water and diffusible solutes but not compounds of high molecular mass such as proteins. The "impermeability" to proteins is not absolute, and in some pathologic conditions causing "shock," such as bacterial sepsis, the permeability of the vascular endothelium increases dramatically, resulting in leakage of albumin, a reduction in the effective circulating volume, and hypotension.

Intracellular Fluid

The exact composition of ICF is difficult to measure and varies by tissue type. Data for ICF (see Table 36.2), therefore, are considered only approximations. The ICF constitutes ≈66% of the total body volume (see Fig. 36.1).

Distribution of Ions by Active and Passive Transport

The electrolyte compositions of blood plasma and interstitial fluid are similar (both are ECFs) and differ markedly from that of ICF (see Table 36.2). The major ECF ions are Na^+, Cl^-, and HCO_3^-, but in ICF, the main ions are K^+, Mg^{2+}, organic phosphates, and protein. These unequal distributions of ions are the consequence of active and passive transport of ions and larger molecules. The difference in Na^+ concentration, for instance, is due to active transport of Na^+ from inside to outside the cell against its electrochemical gradient. This gradient is maintained predominantly by an active sodium pump deriving its energy from glycolysis generated adenosine triphosphate (ATP). A Na^+/K^+-ATPase frequently couples the active transport of Na^+ out of the ICF with transport of K^+ into the cell against the K^+ concentration gradient.

ELECTROLYTES

Disorders of Na^+, K^+, Cl^-, and HCO_3^- will now be separately considered, even though disorders of electrolyte and water homeostasis need a systematic evaluation rather than an individual review of each ion.

SODIUM

Disorders of Na^+ homeostasis can occur because of excessive loss, gain, or retention of Na^+, or that of H_2O. It is difficult to separate disorders of Na^+ and H_2O balance because of their close relationship in establishing normal osmolality in all body water compartments. As described in detail in Chapter 35, the primary organ for regulating body water and extracellular Na^+ is the kidney.

The kidney is responsible for clearing uremic toxins from the circulation as well as for maintaining electrolyte homeostasis across a wide range of gains and losses. In the proximal tubules, 70% to 80% of filtered Na^+ is actively reabsorbed, with H_2O and Cl^- following passively to maintain electrical neutrality and osmotic equivalence. In the descending loop of Henle, H_2O, but not electrolytes, is passively reabsorbed because of the high osmotic strength of interstitial fluid in the renal medulla. In the ascending loop of Henle, Cl^- is reabsorbed actively, with Na^+ following. At the distal tubule, the first of the two primary Na^+/H_2O regulating processes occurs. Here, aldosterone stimulates the cortical collecting ducts to reabsorb Na^+ (with water following passively) and secrete K^+ (and to a lesser extent, H^+) to maintain electrical neutrality. Aldosterone is produced by the adrenal cortex in response to angiotensin II derived via the action of renin. The secretion of renin by renal juxtaglomerular cells is stimulated by low Cl^-, by β-adrenergic activity, and by low arteriolar pressure. Thus, when the kidneys are hypoperfused (as occurs when blood volume decreases, or when the renal arteries are obstructed), the distal tubules, under the influence of aldosterone, reclaim Na^+.

Further water regulation in the kidney occurs from the distal tubule through the collecting duct, where tubular permeability to H_2O is under the influence of antidiuretic hormone (ADH), also known as vasopressin (see Chapters 35 and 40). ADH is released by the posterior pituitary under the influence of baroreceptors in the aortic arch and of hypothalamic chemoreceptors that are responsive to circulating osmolality, which is primarily a reflection of Na^+ concentration. When ECF volume is decreased, or when plasma osmolality is increased, ADH is secreted, tubular permeability to H_2O increases via aquaporins, and H_2O is reabsorbed in an attempt to restore blood volume or to decrease osmolality. In contrast, when ECF volume is increased or osmolality decreased, ADH secretion is inhibited, and more H_2O is excreted in urine (diuresis).

The body's only nonrenal mechanism for restoring Na^+/H_2O homeostasis is ingestion of H_2O. Thirst is stimulated by decreased blood volume or hyperosmolality. It is important to remember that baroreceptors that influence renal handling of Na^+ and H_2O, and thirst, sense changes only in the intravascular blood volume and not the total ECF, while osmoreceptors in the brain sense the osmolality of the ECF surrounding the cells. Laboratory assessment of water and electrolyte disorders is made primarily from the blood volume (plasma); the clinician must assess the status of TBW and blood volume before interpreting laboratory values (see Table 36.1).

Hyponatremia

Hyponatremia is defined as a decreased plasma Na^+ concentration (defined as <130 to 135 mmol/L). Hyponatremia manifests clinically as nausea, generalized weakness, and mental confusion at values less than 120 mmol/L, and severe mental confusion plus convulsions at less than105 mmol/L. The Na^+ concentration at which symptoms develop depends on how quickly hyponatremia develops, and symptoms may manifest at higher Na^+ concentrations (≈125 mmol/L) when hyponatremia develops rapidly. Central nervous system (CNS) symptoms are due to the movement of H_2O into cells to maintain osmotic balance and subsequent swelling of CNS cells. These symptoms can occur more rapidly in children, so there is a need to be particularly vigilant in the pediatric population.

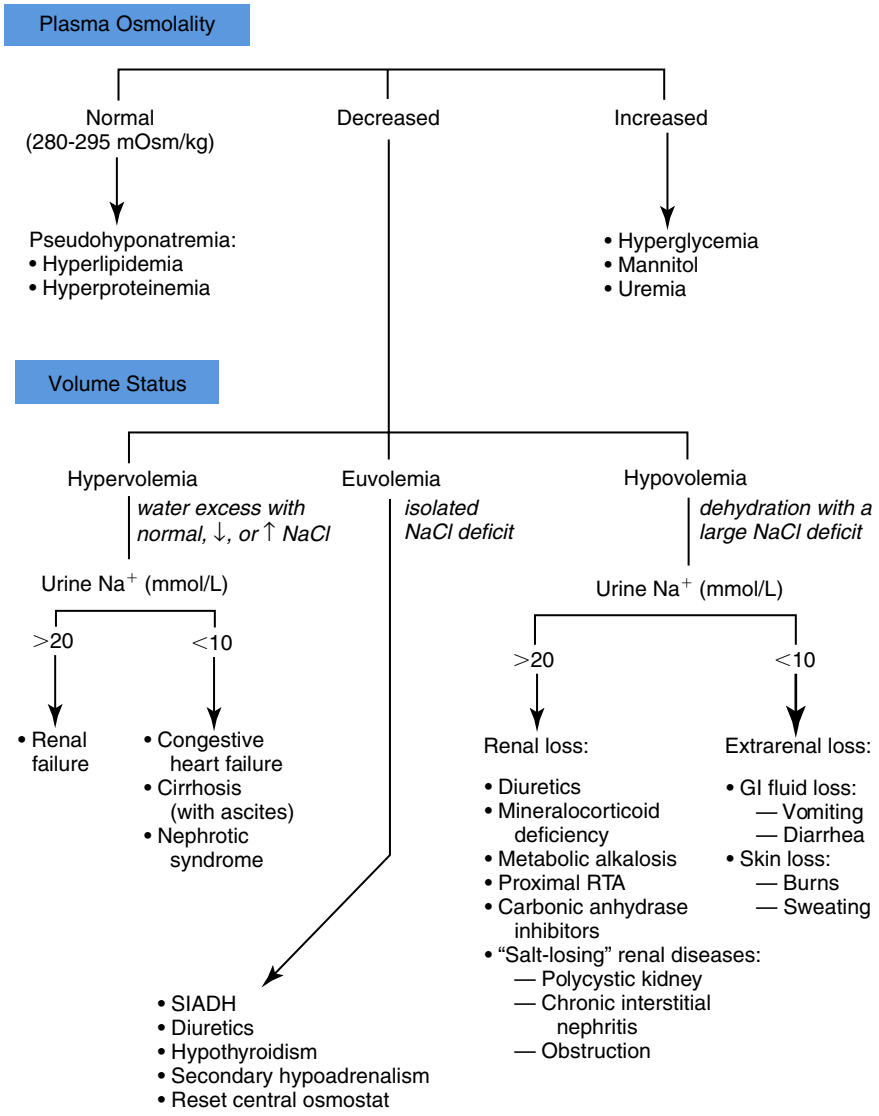


Fig. 36.2 Algorithm for the differential diagnosis of hyponatremia. *GI,* Gastrointestinal; *RTA,* renal tubular acidosis; *SIADH,* syndrome of inappropriate secretion of antidiuretic hormone. (Modified from Kirkpatrick, W., & Kreisberg, R. [1986]. Acid-base and electrolyte disorders. In P. Liu [Ed.], *Blue book of diagnostic tests* [pp. 239–254]. Philadelphia: WB Saunders.)

Hyponatremia can be hypo-osmotic, hyperosmotic, or iso-osmotic. Thus measurement of plasma osmolality is an important initial step in the assessment of hyponatremia. Of these, the most common form is hypo-osmotic hyponatremia due to the large role that Na^+ plays in maintaining osmolality. Fig. 36.2 describes an algorithm for laboratory measurements and physical examination findings in the differential diagnosis of plasma Na^+ less than 135 mmol/L.

Hypoosmotic Hyponatremia

Typically when plasma Na^+ concentration is low, calculated or measured osmolality will also be low. This type of hyponatremia can be due to excess loss of Na^+ (**depletional hyponatremia**) or increased ECF volume (**dilutional hyponatremia**). Differentiating these initially requires clinical assessment of TBW and ECF volume by history and physical examination.

Depletional hyponatremia results from a loss of Na^+ from the ECF space that exceeds the concomitant loss of water. The net loss of Na^+ from the ECF space stimulates thirst and production of ADH, both of which contribute to the maintenance of hyponatremia. **Hypovolemia** is apparent in the physical examination (orthostatic hypotension, tachycardia, decreased skin turgor). If urine Na^+ is low (<10 mmol/L), the kidneys are properly retaining filtered Na^+ and the loss is extrarenal, most commonly from the gastrointestinal (GI) tract or skin (see Fig. 36.2). Preventing ongoing loss and restoring ECF volume with isotonic fluid is sufficient to correct hyponatremia in these situations. Alternatively, if urine Na^+ is increased in this setting (generally >20 mmol/L), renal loss of Na^+ is likely. Renal loss of Na^+ occurs with (1) osmotic diuresis, (2) use of diuretics (which inhibit reabsorption of Cl^- and Na^+ in the ascending loop), (3) adrenal insufficiency (lack of aldosterone or cortisol prevents distal tubule reabsorption of Na^+), or

(4) salt wasting nephropathies, as can occur with interstitial nephritis and tubular recovery after acute tubular necrosis or obstructive nephropathy. Renal loss of Na^+ in excess of H_2O can also occur in metabolic alkalosis from prolonged vomiting, because increased renal HCO_3^- excretion is accompanied by Na^+ ions. In this case, urine sodium is increased (>20 mmol/L), but urine chloride remains low. In proximal renal tubular acidosis (RTA) type 2, bicarbonate is lost because of a defect in HCO_3^- reabsorption and is accompanied by Na^+ to maintain electrical neutrality. As with extrarenal Na^+ loss, management of hyponatremia attributable to renal Na^+ loss is centered on reversing the underlying cause and restoration of ECF volume.

Dilutional hyponatremia is a result of excess H_2O retention and often can be detected during the physical examination as edema. In advanced renal failure, water is retained because of decreased filtration and H_2O excretion. When ECF is increased but the circulating blood volume is decreased, as occurs in hepatic cirrhosis or nephrotic syndrome, a vicious cycle is established. The decreased blood volume is sensed by baroreceptors and results in increased aldosterone and ADH, even though ECF volume is excessive. The kidneys reabsorb Na^+ and H_2O in response to increased aldosterone and ADH to restore the blood volume, resulting in further increases in ECF and further dilution of Na^+.

In hypoosmotic hyponatremia with a normal or euvolemic volume status, the most common causes are syndrome of inappropriate ADH (SIADH), primary polydipsia, and endocrine disorders such as secondary adrenal insufficiency and severe hypothyroidism (see Fig. 36.2). SIADH describes hyponatremia attributable to "inappropriate" ADH release, which stimulates excessive H_2O retention and increased urine osmolality. In secondary adrenal insufficiency, due to lack of adrenocorticotropic hormone (ACTH), glucocorticoid secretion is impaired. In the absence of cortisol and ACTH, increased corticotropin-releasing hormone (CRH) secretion stimulates ADH release and causes hyponatremia. Severe hypothyroidism impairs free H_2O excretion due to reduced cardiac output and glomerular filtration rate (GFR) in myxedema. Finally, euvolemic hyponatremia can also be found in primary polydipsia when water intake is greater than the renal capacity to excrete excess H_2O.

Hyperosmotic Hyponatremia

Hyponatremia in the presence of increased quantities of other solutes in the ECF is the result of an extracellular shift of water or an intracellular shift of Na^+ to maintain osmotic balance between ECF and ICF compartments. The most common cause of this type of hyponatremia is severe hyperglycemia but may also occur in the setting of mannitol diuresis (see Fig. 36.2). As a rule, Na^+ is decreased by ≈1.6 to 2.0 mmol/L for every 100 mg/dL (5.6 mmol/L) increase in glucose above 100 mg/dL (5.6 mmol/L).

Isosmotic Hyponatremia

If the measured Na^+ concentration in plasma is decreased, but measured plasma osmolality, glucose, and urea are normal, the most likely explanation is pseudohyponatremia caused by the *electrolyte exclusion effect* (see Chapter 24). This occurs when Na^+ is measured by an indirect ion-selective electrode in patients with severe hyperlipidemia or hyperproteinemia.

Hypernatremia

Hypernatremia (plasma Na^+ >150 mmol/L) is always hyperosmolar. Symptoms of hypernatremia are primarily neurologic (because of neuronal cell loss of H_2O to the ECF) and include tremors, irritability, ataxia, confusion, and coma. As with hyponatremia, the speed at which hypernatremia develops determines the plasma Na^+ concentration at which symptoms occur. Acute development may cause symptoms at 160 mmol/L, whereas in chronic hypernatremia, symptoms may not occur until Na^+ exceeds 175 mmol/L.

In many cases, the symptoms of hypernatremia may be masked by underlying conditions. Hypernatremia rarely occurs in an alert patient with a normal thirst response and access to water. Most cases are observed in patients with altered mental status or infants, both of whom may not be capable of rehydrating themselves.

Hypernatremia arises in the setting of (1) hypovolemia (excessive water loss or failure to replace normal water losses), (2) **hypervolemia** (a net Na^+ gain in excess of water gain), or (3) normovolemia. Again, assessment of TBW status by physical examination and measurement of urine Na^+ and osmolality are important steps in establishing a diagnosis (Fig. 36.3).

Hypovolemic Hypernatremia

Hypernatremia in the setting of decreased ECF is caused by renal or extrarenal loss of hypo-osmotic fluid, leading to dehydration. Once hypovolemia is established by physical examination, measurement of urine Na^+ and osmolality is used to determine the source of fluid loss. Patients with large extrarenal fluid losses will have concentrated urine (often >800 mOsmol/L) with low urine Na^+ (<20 mmol/L), reflecting a proper renal response to conserve Na^+ and water to restore ECF volume. Extrarenal causes include diarrhea, skin losses (burns, fever, or excessive sweating), and respiratory losses coupled with failure to replace the water. When extrarenal loss is excluded, and the patient has normal mental status and access to H_2O, a hypothalamic disorder (tumor or granuloma) inducing **diabetes insipidus** (DI) should be suspected. An additional consideration in the setting of poorly controlled diabetes mellitus is osmotic diuresis due to plasma glucose values greater than 600 mg/dL (33.3 mmol/L). This condition, referred to as hyperosmolar hyperglycemic nonketotic syndrome, occurs most commonly in elderly subjects with type 2 diabetes mellitus.

Normovolemic Hypernatremia

Hypernatremia in the presence of normal ECF volume typically arises due to insensible losses through the lungs or skin and is often characterized by concentrated urine as the kidneys conserve water. Normovolemic hypernatremia arising from insensible losses is frequently a prelude to hypovolemic

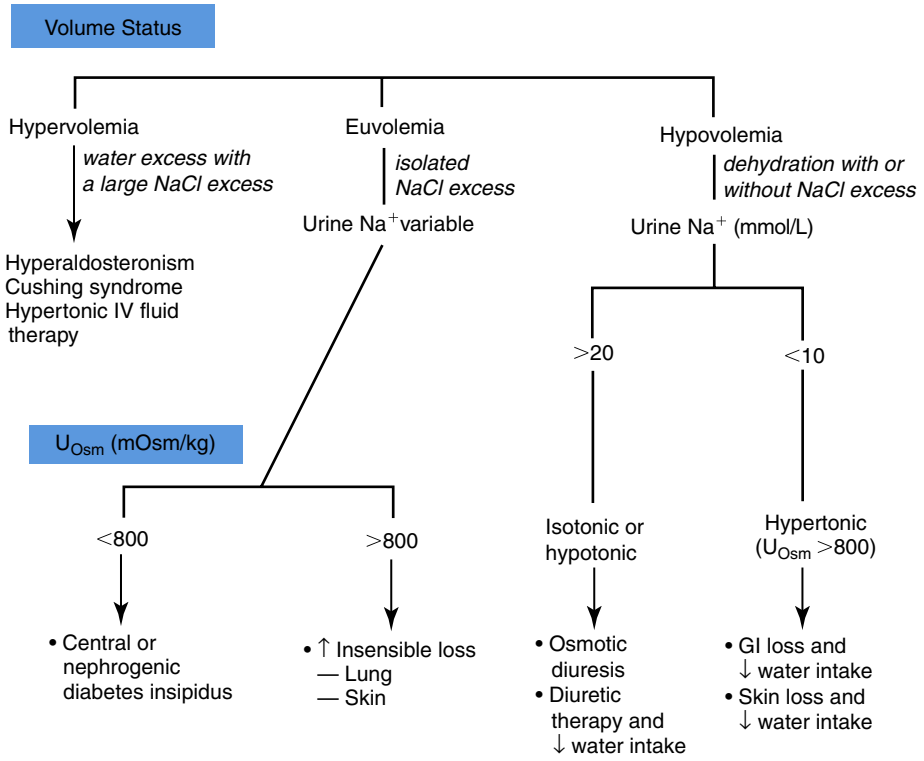


Fig. 36.3 Algorithm for the differential diagnosis of hypernatremia. (Modified from Kirkpatrick, W., & Kreisberg, R. [1986]. Acid-base and electrolyte disorders. In P. Liu [Ed.], *Blue Book of Diagnostic Tests* [pp. 239–254]. Philadelphia: WB Saunders.)

hypernatremia. Water or solute diuresis resulting in polyuria (generally defined as >3 L urine output/day) may also result in normovolemic hypernatremia (see Fig. 36.3). Solute diuresis is exemplified by the osmotic diuresis of diabetes mellitus and generally is characterized by urine osmolality greater than 300 mOsmol/L and hyponatremia (see previous discussion in this chapter). Water diuresis, a manifestation of DI, is characterized by dilute urine (osmolality <250 mOsmol/L) and hypernatremia. DI can be central or nephrogenic. Central DI is due to decreased or absent ADH secretion resulting from head trauma, hypophysectomy, pituitary tumor, or granulomatous disease. Nephrogenic DI is due to renal resistance to ADH as a result of drugs (e.g., lithium, demeclocycline, amphotericin); electrolyte disorders (e.g., hypercalcemia, hypokalemia); sickle cell anemia; or Sjögren syndrome. When thirst and access to water are uncompromised, many patients with DI will remain normonatremic since their free water losses are offset by intake. Such patients only display symptoms of polyuria and polydipsia. However, overt hypernatremia can develop with progression of underlying causes, impaired thirst, or restricted access to water. Administration of ADH (in the form of desmopressin) can be used to treat central DI, although patients with nephrogenic DI may be resistant to it.

Hypervolemic Hypernatremia

The presence of excess TBW and hypernatremia indicates a net gain of water and Na^+, with Na^+ gain in excess of water (see Fig. 36.3). This rare condition is observed most commonly in hospitalized patients receiving hypertonic saline or sodium bicarbonate.

POTASSIUM

The major intracellular cation is K^+, and only about 2% of total body K^+ is present in the ECF. Nevertheless, plasma K^+ is often a good indicator of total K^+ stores unless the K^+ concentration is abnormal due to cellular shifts. Disturbance of K^+ homeostasis has serious consequences. For example, a decrease in extracellular K^+ (hypokalemia) is characterized by muscle weakness, irritability, and paralysis. Plasma K^+ concentrations less than 3.0 mmol/L are often associated with neuromuscular symptoms and indicate a critical degree of intracellular depletion. At lower concentrations, tachycardia and cardiac conduction defects develop and can lead to cardiac arrest.

High extracellular K^+ (**hyperkalemia**) concentrations may produce symptoms of mental confusion, weakness, tingling, flaccid paralysis of the extremities, weakness of the respiratory muscles, bradycardia, and cardiac conduction defects. Prolonged, severe hyperkalemia greater than 7.0 mmol/L can lead to peripheral vascular collapse and cardiac arrest. Symptoms or electrocardiogram (EKG) abnormalities are usually present at K^+ concentrations greater than 6.5 mmol/L. Potassium concentrations greater than 10.0 mmol/L in most cases are fatal, though fatalities can occur at significantly lower values.

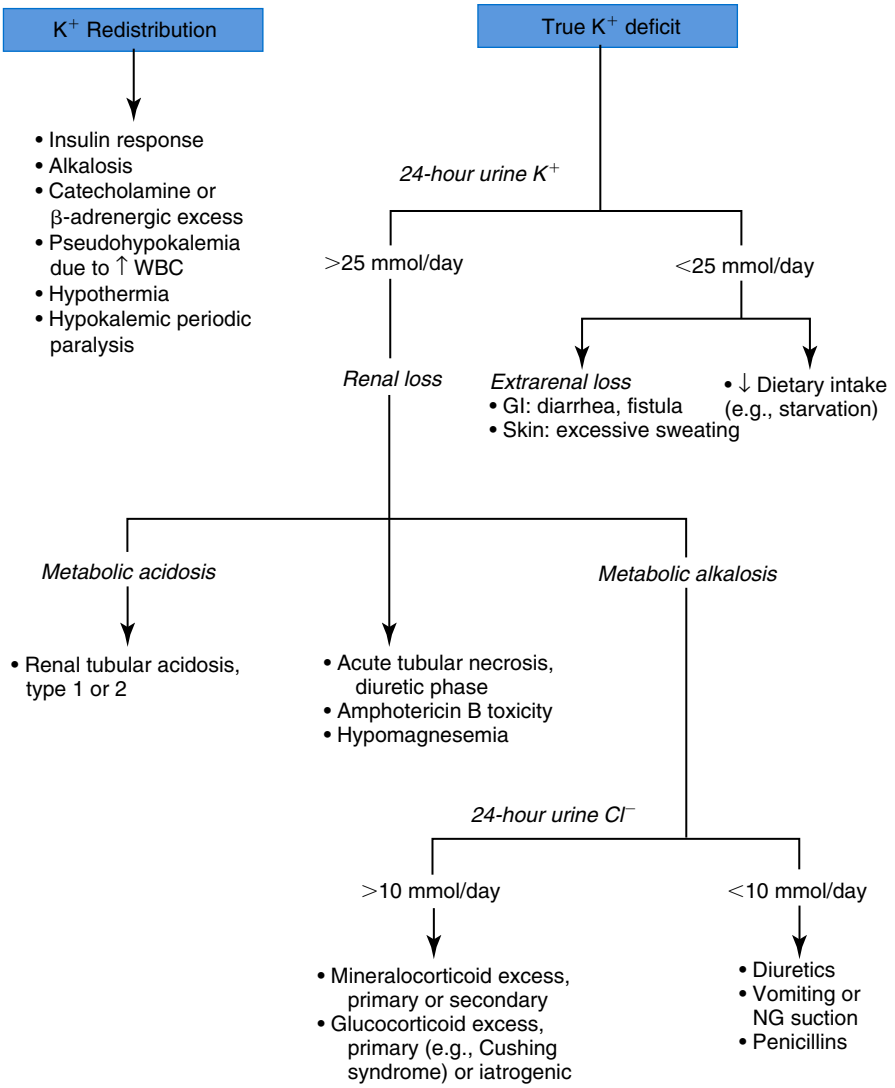


Fig. 36.4 Algorithm for the differential diagnosis of hypokalemia. *GI,* Gastrointestinal; *NG,* nasogastric; *WBC,* white blood cell. (Modified from Kirkpatrick, W., & Kreisberg, R. [1986]. Acid-base and electrolyte disorders. In P. Liu [Ed.], *Blue Book of Diagnostic Tests* [pp. 239–254]. Philadelphia: WB Saunders.)

Hypokalemia

Causes of **hypokalemia** (plasma K^+ <3.5 mmol/L) are classified as redistribution of extracellular K^+ into ICF, or true K^+ deficits, caused by decreased intake or loss of potassium-rich body fluids (Fig. 36.4).

Redistribution

Intracellular redistribution of K^+ is illustrated by the fall in plasma K^+ following insulin therapy for diabetic hyperglycemia. Insulin plays a crucial role in maintaining the intracellular distribution of K^+ through active cellular transport as well as glucose control. Redistribution hypokalemia is also a feature of alkalosis, in which K^+ moves from ECF into cells in exchange for H^+ efflux and renal conservation of H^+ at the expense of urinary K^+ loss. Other causes of intracellular redistribution are listed in Fig. 36.4. Clinically, redistributive hypokalemia is generally a transient phenomenon that is reversed once underlying conditions are corrected. Hypokalemia is highly prevalent in cancer patients but must be differentiated from pseudohypokalemia. Pseudohypokalemia can occur in settings of very high white blood cell or platelet counts. In some patients in leukemic blast crisis there can be a time-dependent transport of K^+ into the leukemic cells after the blood sample is drawn. In these settings, it is important to process the samples as quickly as possible.

True Potassium Deficit

Hypokalemia reflecting true total body deficits of K^+ can be classified into renal and nonrenal losses (see Fig. 36.4). If urine excretion of K^+ is less than 30 mmol/day, it can be concluded that the kidneys are properly functioning and are reabsorbing K^+. The cause may be decreased K^+ intake or extrarenal loss of K^+-rich fluid. Causes of decreased intake include chronic starvation and postoperative intravenous fluid therapy with K^+-poor solutions. GI loss of K^+ occurs most commonly with diarrhea and loss of gastric fluid through vomiting.

Urine excretion exceeding 30 mmol/day in a hypokalemic setting is inappropriate and indicates that the kidneys are

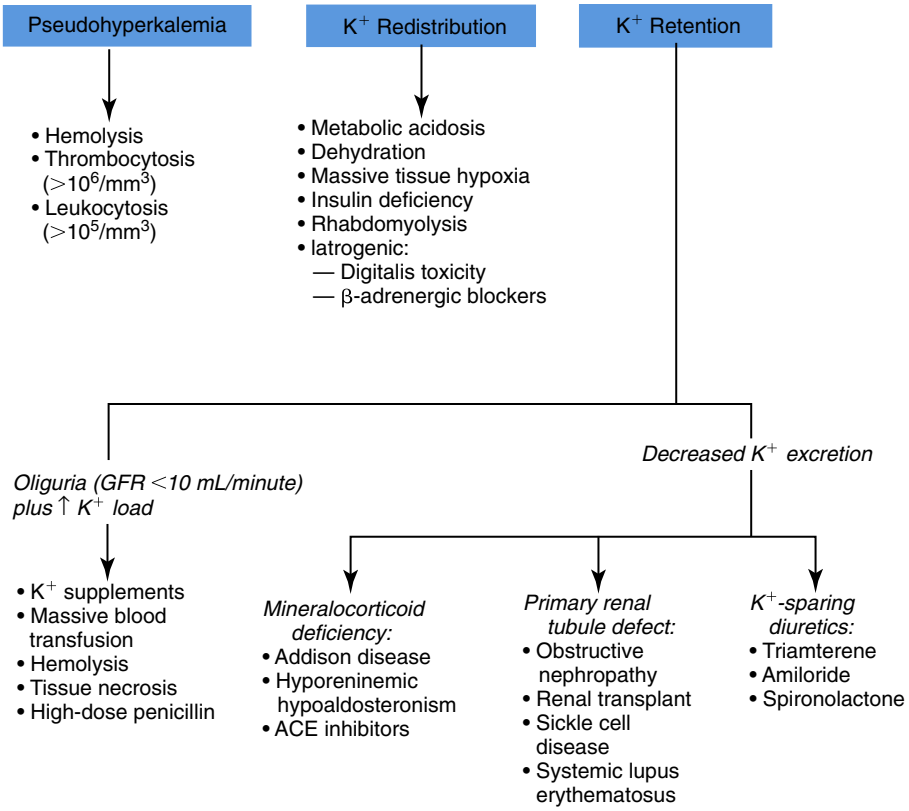


Fig. 36.5 Algorithm for the differential diagnosis of hyperkalemia. *ACE,* Angiotensin-converting enzyme. (Modified from Kirkpatrick, W., & Kreisberg, R. [1986]. Acid-base and electrolyte disorders. In P. Liu [Ed.], *Blue Book of Diagnostic Tests* [pp. 239–254]. Philadelphia: WB Saunders.)

the primary source of K^+ loss. Renal losses of K^+ may occur during the diuretic (recovery) phase of acute tubular necrosis and during states of excess mineralocorticoid (primary or secondary aldosteronism) or glucocorticoid (Cushing syndrome) when the distal tubules increase Na^+ reabsorption and K^+ excretion. Renal loss of K^+ is also caused by thiazide and loop diuretics. In addition to the redistribution of K^+ into cells in an alkalotic setting, K^+ can be lost from the kidneys in exchange for reclaimed H^+ ions. This cause of true hypokalemia is associated with low urine Cl^- and alkaline urine. True potassium deficit requires replacement of potassium. Although there are dietary sources of potassium, such as potatoes and tomatoes, significant K^+ losses may require oral or intravenous supplementation with potassium chloride.

Hyperkalemia

Hyperkalemia (plasma K^+ >5.0 mmol/L) is a result of (singly or in combination) (1) redistribution, (2) increased intake, or (3) increased retention. In addition, preanalytical conditions—such as hemolysis, thrombocytosis (>106/μL), and leukocytosis (>105/μL together with delayed sample analysis)—have been known to cause marked pseudohyperkalemia, as described in detail in Chapter 24.

Redistribution

The transfer of intracellular K^+ into ECF invariably occurs in acidemia as K^+ shifts outward as the result of pH-induced changes in Na^+/K^+-ATPase activity. When acidemia is corrected, normokalemia will be restored rapidly. Extracellular redistribution of K^+ may also occur in (1) tissue hypoxia; (2) insulin deficiency (e.g., diabetic ketoacidosis); (3) massive intravascular hemolysis; (4) severe burns; (5) intense and sustained muscular activity, as in status epilepticus; (6) rhabdomyolysis; (7) tumor lysis syndrome, as well as (8) iatrogenic causes. Redistributive hyperkalemia can be corrected by reversing the aberrations that cause K^+ to shift out of cells. Insulin and sodium bicarbonate are commonly used and have a quick onset of action, particularly in the diabetic or acidemic setting.

Potassium Retention

When glomerular filtration or renal tubular function is decreased, hyperkalemia will often occur. In the absence of severe renal failure, hyperkalemia may not develop, and if it does it is seldom prolonged. Decreased excretion of K^+ in moderate and acute renal disease and end-stage renal failure (with oliguria or anuria) are the most common causes of prolonged hyperkalemia (Fig. 36.5). Hyperkalemia occurs along with Na^+ depletion in adrenocortical insufficiency (e.g., Addison disease) because diminished Na^+ reabsorption results in decreased tubular K^+ secretion. Drugs that block the production of aldosterone, such as inhibitors of angiotensin-converting enzyme (ACE inhibitors; e.g., lisinopril), nonsteroidal anti-inflammatory drugs, and angiotensin II receptor blockers, may also cause hyperkalemia. Excess administration of potassium-sparing diuretics that block distal tubular K^+ secretion (e.g., triamterene, spironolactone) may also cause hyperkalemia.

CHLORIDE

In the absence of acid-base disturbances, Cl^- concentrations in plasma generally will follow those of Na^+. However, determination of plasma Cl^- concentration is useful in the differential diagnosis of acid-base disturbances and is essential for calculating the anion gap. Fluctuations in serum or plasma Cl^- have little clinical consequence but do serve as signs of an underlying disturbance in fluid or acid-base homeostasis. The specific replacement of chloride is rarely targeted at chloride deficit independently, but it is a corner stone of management for metabolic alkalosis.

Hypochloremia

In general, causes of hypochloremia parallel causes of hyponatremia. Persistent gastric secretion and prolonged vomiting result in significant loss of Cl^- and ultimately in hypochloremic alkalosis and depletion of total body Cl^- with retention of HCO_3^- by the nephron to reabsorb the filtered Na^+ load. **Respiratory acidosis**, which is accompanied by increased HCO_3^-, is another common cause of decreased Cl^- with normal Na^+.

Hyperchloremia

Increased plasma Cl^- concentration, similar to increased Na^+ concentration, occurs with dehydration, prolonged diarrhea with loss of sodium bicarbonate, DI, and overtreatment with normal saline solutions, which have a Cl^- content of 150 mmol/L. A rise in Cl^- concentration may also be seen in **respiratory alkalosis** because of renal compensation for excreting HCO_3^-.

Points to Remember

- Physical examination is important in assessing total body water status and hyponatremic or hypernatremic disorders.
- Urine electrolytes are important to determine if the kidneys are functioning properly in disorders of electrolyte and water balance.
- Patient history can be important in assessing electrolyte disorders. Examples include vomiting, diarrhea, water deprivation, excess perspiration, anuria, diabetes, and so on.

ACID-BASE PHYSIOLOGY

Normal metabolic processes result in the production of large amounts of metabolic acids. These products of metabolism are transported to the lungs and kidneys via the ECF and blood with no appreciable change in the ECF pH, and with only a minimal difference between arterial (pH 7.35 to 7.45) and venous (pH 7.32 to 7.42) blood. This is accomplished by the buffering capacity of blood and by respiratory and renal regulatory mechanisms.

ACID-BASE BALANCE AND ACID-BASE STATUS

Interpretation of the **acid-base balance** in physiology involves an accounting of the carbonic (H_2CO_3, HCO_3^-, CO_3^{2-}, and CO_2) and noncarbonic acids and conjugate bases in terms of input (intake plus metabolic production) and output (excretion plus metabolic conversion) over a given time interval. For example, during a 24-hour period, a 70 kg person exhales approximately 20 moles of CO_2 (volatile form of carbonic acid) through the lungs, and excretes about 70 to 100 mmol (or ≈1 mmol/kg) of nonvolatile acids (mainly sulfuric and phosphoric acids) through the kidneys.

Acid-Base Parameters—Definitions and Abbreviations

Acids are chemical substances that can donate protons (H^+ ions) in solution, and *bases* are substances that accept protons. Strong acids readily give up H^+, whereas strong bases readily accept H^+. Thus the conjugate base of a strong acid is a weak base and vice versa.

Acidemia is defined as an arterial blood pH less than 7.35, and **alkalemia** indicates an arterial blood pH greater than 7.45. *Acidosis* and *alkalosis* refer to pathologic states that often lead to acidemia or alkalemia.

pH and pK

The pH of a solution is defined as the negative logarithm of the hydrogen ion activity ($pH = -\log aH^+$). A decrease in 1 pH unit represents a 10-fold increase in H^+ activity. pH is measured using potentiometry (see Chapter 24) and is an assessment of H^+ ion activity. In the setting of pH assessment, an assumption is made that H^+ activity is equivalent to concentration. The H^+ concentration of blood is thus equal to 40 nmol/L at pH 7.4.

The p*K* (also known as p*K′* and p*Ka*) is the pH at which an acid is half-dissociated, existing as equal proportions of acid and conjugate base. Acids have p*K* values less than 7.0, whereas bases have p*K* values greater than 7.0. The lower the p*K*, the stronger the acid, and the higher the p*K*, the stronger the conjugate base.

The pH of plasma is the function of two independent variables:

1. PCO_2, which is regulated by the lungs and represents the acid component of the carbonic acid/bicarbonate buffer system
2. The concentration of titratable base (base excess or deficit, which is defined later), which is regulated by the kidneys

Bicarbonate and Dissolved Carbon Dioxide

Bicarbonate is the second largest fraction (behind Cl^-) of plasma anions. Conventionally, it is defined to include (1) plasma bicarbonate ion (HCO_3^-), (2) carbonate ion (CO_3^{2-}), and (3) CO_2 bound in plasma carbamino compounds (R-C-NH-COOH). Actual bicarbonate ion concentration is not measured in clinical laboratories. The analyte usually measured in plasma is total CO_2, which includes bicarbonate and dissolved CO_2 (dCO_2) but is often referred to as "serum bicarb".

Henderson-Hasselbalch Equation

The **Henderson-Hasselbalch equation** is described in detail in Chapter 24. However, it is important to review here because it enhances understanding of pH regulation of body fluids as it

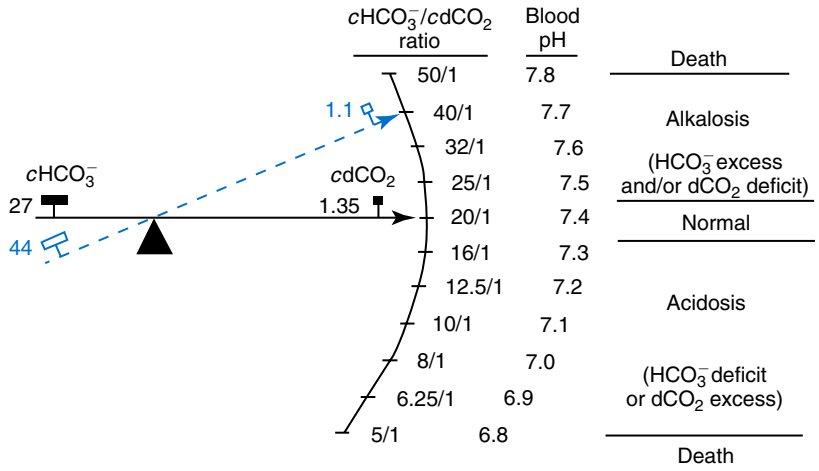


Fig. 36.6 Scheme demonstrating the relation between pH and the ratio of bicarbonate concentration to the concentration of dissolved CO_2. If the ratio in blood is 20:1 ($cHCO_2^-$ = 27 mmol/$cdCO_2$ = 1.35 mmol/L), the resultant pH will be 7.4, as demonstrated by the *solid beam*. The *dotted line* shows a case of uncompensated alkalosis (bicarbonate excess) with a bicarbonate concentration of 44 mmol/L and a $cdCO_2$ of 1.1 mmol/L. The ratio therefore is 40:1, and the resultant pH is 7.7. In a case of uncompensated acidosis, the pointer of the balance would point to a pH between 6.8 and 7.35, depending on the $cHCO_2^-/cdCO_2$ ratio. (From Weisberg, H. F. [1959]. A better understanding of anion-cation ["acid-base"] balance. *Surgical Clinics of North America*, *39*, 93–120.)

relates to compensatory mechanisms in acid-base disturbances. The equation described in Chapter 24 can be written as follows:

$$pH = 6.1 + \log \frac{cHCO_3^-}{cdCO_2},$$

where $cdCO_2$ is equal to α (0.0306 mmol/L per mm Hg) PCO_2 and 6.1 is the p*K*′ for the carbonic acid/bicarbonate system.

The average normal ratio of the concentrations of bicarbonate and dCO_2 in plasma is 25 (mmol/L)/1.25 (mmol/L) = 20/1. It follows then that any change in the concentration of bicarbonate or dCO_2 relative to each other must be accompanied by a change in pH. Such changes in this important ratio can occur through a change in $cHCO_3^-$ (the renal component) or in PCO_2 (the respiratory component). Clinical conditions characterized as *metabolic* disturbances of acid-base balance are classified as primary disturbances in $cHCO_3^-$. Those characterized as *respiratory* disturbances are classified as primary disturbances in $cdCO_2$ (PCO_2). Various compensatory mechanisms attempting to reestablish the normal ratio of $cHCO_3^-/cdCO_2$ may result in changes in bicarbonate concentration, dCO_2 concentration, or both. Application of the Henderson-Hasselbalch equation to human acid-base physiology are illustrated in Fig. 36.6.

BUFFER SYSTEMS AND THEIR ROLE IN REGULATING THE PH OF BODY FLUIDS

A buffer is a mixture of a weak acid and a salt of its conjugate base that resists changes in pH when a strong acid or base is added to the solution. If concentrations of the acid and base components of a buffer are equal, the pH will equal the p*K*. Generally, buffers work best at resisting pH changes in the interval of ±1 pH unit of its p*K* and are more effective at higher molar concentrations. The action of buffers in the regulation of body pH can be demonstrated by using the bicarbonate buffer system as an example. If a strong acid is added to a solution containing HCO_3^-, and H_2CO_3, the H^+ will react with HCO_3^- to form more H_2CO_3, and subsequently CO_2 and H_2O. The H^+ ions are thereby bound, and the increase in the H^+ concentration will be minimal.

$$H^+ + HCO_3^- \rightleftharpoons H_2CO_3 \rightleftharpoons CO_2 + H_2O$$

Bicarbonate and Carbonic Acid Buffer System

The most important buffer of plasma is the bicarbonate/carbonic acid pair even though its p*K* is 6.1, and normal plasma pH is 7.4. The normal bicarbonate/dCO_2 ratio is 20:1, which is outside the 10:1 or 1:10 ratio at which buffers work best. However, the effectiveness of the bicarbonate buffer is based on the fact that the lungs can readily dispose of or retain CO_2, and is present at higher concentrations than other buffers with the exception of Hb. In addition, the renal tubules can regulate bicarbonate through increased or decreased reclamation (see Chapter 35).

Phosphate Buffer System

At a plasma pH of 7.4, the ratio $cHPO_4^{2-}/cH_2PO_4^-$ is 4: 1 (pK′ = 6.8). The total concentration of this buffer in both erythrocytes and plasma accounts for about 5% of the nonbicarbonate buffer value of plasma. Organic phosphate, in the form of 2,3-diphosphoglycerate (present in erythrocytes in a concentration of approximately 4.5 mmol/L), accounts for about 16% of the nonbicarbonate buffer value of erythrocytes and plays a vital role in the titration and excretion of acids in urine.

Plasma Protein and Hemoglobin Buffer System

Proteins, especially albumin, account for the greatest portion (>90%) of the nonbicarbonate buffer value of plasma. The

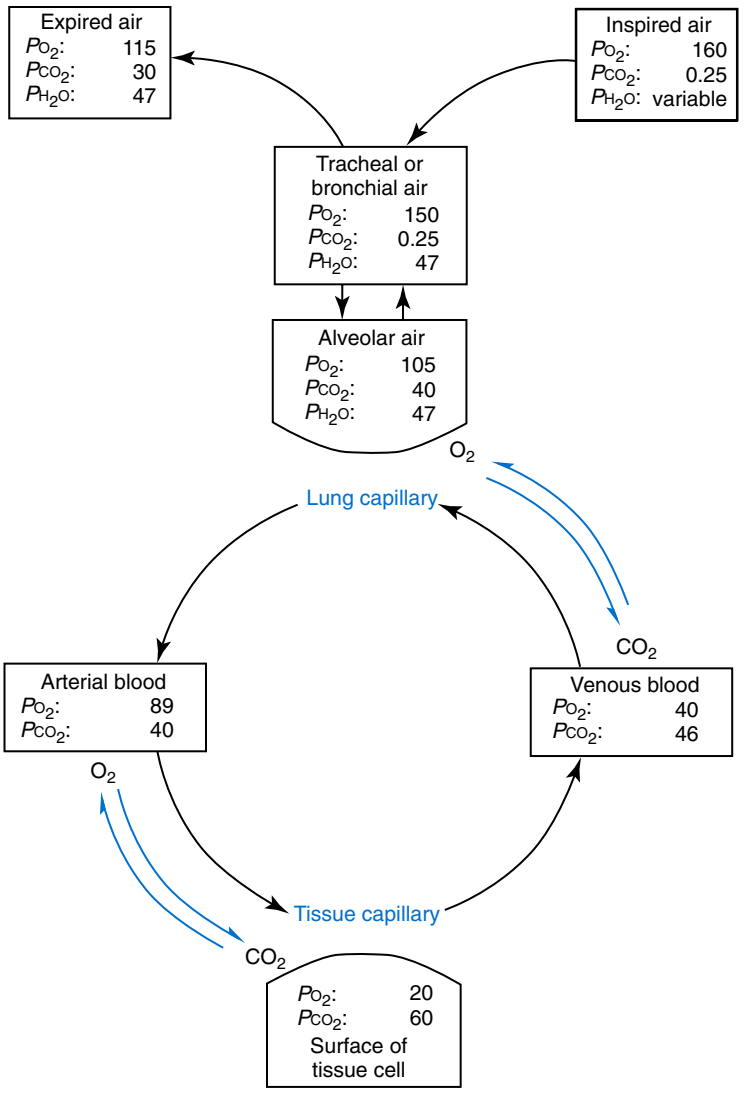


Fig. 36.7 Partial pressures of oxygen and carbon dioxide in air, blood, and tissue. Values shown are approximations in mm Hg and are calculated assuming a 5% shunt. *Blue arrows* show directions of gradients. (Modified from Tietz, N. W. [1987]. *Fundamentals of Clinical Chemistry* [3rd ed.]. Philadelphia: WB Saunders.)

imidazole groups of histidine (pK ~7.3) are the most important components of proteins in the physiological pH range encountered in plasma. For example, albumin contains 16 histidine residues per molecule.

The hemoglobin within erythrocytes also plays a significant role in the physiologic buffering system of whole blood through the diffusion of CO_2 produced by cellular respiration in peripheral tissues into the erythrocyte and its conversion to HCO_3^- and H^+ through the action of carbonic anhydrase. The H^+ initially decreases the erythrocyte fluid pH; however, this decrease in pH is rapidly compensated by the release of oxygen to the tissues from O_2Hb, which involves the conversion of stronger acid (O_2Hb) into weaker acid (HHb) that readily accepts the H^+. The reduced hemoglobin (HHb) also contributes to the buffering capacity of whole blood by binding CO_2 in the form of carbamino-CO_2. The HCO_3^- produced by the action of carbonic anhydrase is shunted out of the erythrocyte in exchange for plasma Cl^- to maintain electroneutrality referred to as the *chloride shift*. In the alveoli, the low PCO_2 and high PO_2 cause a reversal of reactions detailed briefly above to remove CO_2 (and thus acid) from circulation.

RESPIRATORY MECHANISM IN THE REGULATION OF ACID-BASE BALANCE

In addition to supplying O_2 to tissue cells for normal metabolism, the respiratory mechanism contributes to maintenance of normal body pH through elimination or retention of CO_2 in **metabolic acidosis** and **alkalosis**, respectively.

Exchange of Gases in the Lungs and Peripheral Tissues

Diffusion of O_2 and CO_2 across alveolar and cell membranes is governed by gradients in the partial pressure of each gas (Fig. 36.7). Dry air inspired at a pressure of 1 atm (760 mm Hg) consists of 21% O_2 ($PO_2 \approx 160$ mm Hg), 0.03% CO_2 (PCO_2

≈0.25 mm Hg), 78% nitrogen, and ≈0.1% other inert gases. As inspired air passes over the mucous membranes of the upper respiratory tract, it is warmed to 37°C, becomes saturated with water vapor, and mixes with air in the respiratory tree and alveoli, resulting in partial pressures of at the alveolar membrane of ≈105 mm Hg for O_2, ≈40 mm Hg for CO_2, and ≈47 mm Hg for H_2O. Venous blood on the opposite side of the alveolar membrane has PO_2 ≈40 mm Hg and PCO_2 ≈46 mm Hg. Thus the gradient for O_2 is inward, toward the blood, and for CO_2, it is outward, toward the alveoli. CO_2 removal is so efficient that the PCO_2 in expired air is more than 100 times the PCO_2 in inspired air (see Fig. 36.7). In arterial blood, the PO_2 is slightly lower than in alveolar air (90 to 100 vs. 105 mm Hg) as the result of shunting of about 5% of blood that does not equilibrate.

At the arterial end of capillaries in peripheral tissues, the PO_2 of 95 mm Hg is substantially higher than the average PO_2 at the surface of tissue cells (20 mm Hg), and the PCO_2 at 40 mm Hg is substantially lower than that in the cells (50 to 70 mm Hg). Thus, in the tissue capillary, the gradient for O_2 is inward to the cell; for CO_2, it is outward to the capillary blood.

RENAL MECHANISMS IN THE REGULATION OF ACID-BASE BALANCE

The average pH of plasma and of the glomerular filtrate is ≈7.4, whereas the average urinary pH is ≈6.0, reflecting renal excretion of nonvolatile acids. Various functions of the kidneys respond to different alterations in acid-base status. In the case of acidosis, excretion of acids is increased and base is conserved; in alkalosis, the opposite occurs. The pH of the urine changes correspondingly and may vary in *random* specimens from pH 4.5 to 8.0. The ability to excrete variable amounts of acid or base makes the kidney the final defense mechanism against changes in body pH. Renal excretion of acid and conservation of HCO_3^- occur through several mechanisms, including (1) Na^+-H^+ exchange, (2) production of ammonia and excretion of NH_4^+, and (3) reclamation of HCO_3^-.

Na^+-H^+ Exchange

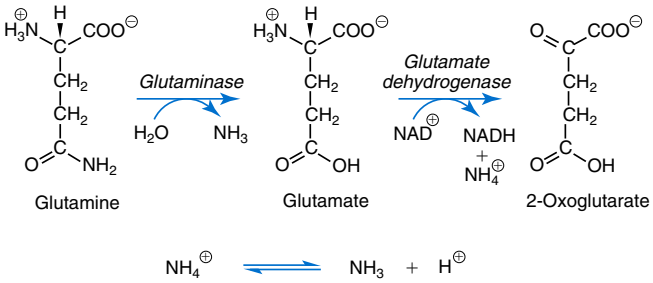
Nearly all mammalian cells contain a plasma membrane ATP-hydrolyzing protein capable of exchanging sodium ions for protons—the so-called Na^+-H^+ exchanger. In the renal tubules the Na^+-H^+ exchangers extrude H^+ ions into the tubular fluid in exchange for Na^+ ions, and their activity is enhanced in states of acidosis and is inhibited in alkalosis.

The proximal tubules, however, cannot maintain an H^+ gradient of more than ≈1 pH unit, whereas the distal tubules cannot maintain more than ≈3 pH units. Thus maximum urine acidity is reached at ≈pH 4.4.

Potassium ions compete with H^+ at the Na^+-H^+ exchangers in the renal tubule. If intracellular K^+ concentration of renal tubular cells is high, more K^+ and less H^+ are exchanged for Na^+. As a result, urine is less acidic, and body fluid pH decreases. If K^+ is depleted, more H^+ ions are exchanged for Na^+; urine becomes more acidic and body fluids more alkaline. In essence, hyperkalemia contributes to acidosis and hypokalemia to alkalosis.

Renal Production of Ammonia and Excretion of Ammonium Ions

Renal tubular cells are able to generate ammonia from glutamine and other amino acids derived from muscle and liver cells according to the following reaction:



The ammonium ion produced dissociates into ammonia and hydrogen ions to a degree dependent on the pH. Hydrogen ions secreted into the tubular lumen through the Na^+-H^+ exchangers lower the pH of tubular fluid, and at the acidic pH of urine, the equilibrium between NH_4^+ and NH_3 shifts markedly to the left, strongly favoring the formation of NH_4^+. Due to its positive charge, the NH_4^+ formed in the tubular lumen cannot readily cross cell membranes and is trapped in the tubule to be excreted in urine with phosphate, chloride, or sulphate anions.

Excretion of H^+ as $H_2PO_4^-$

Hydrogen ions secreted into the tubular lumen by the Na^+-H^+ exchanger may also react with HPO_4^{2-} to form $H_2PO_4^-$. Acidemia increases phosphate excretion and thus provides additional buffer for reaction with H^+. A decrease in the GFR results in a decrease in $H_2PO_4^-$ excretion.

Reclamation of Filtered Bicarbonate

The unmodified glomerular filtrate has the same concentration of HCO_3^- as plasma; however, with increasing acidification of proximal tubule fluid, the HCO_3^- concentration is decreased. The H^+ excreted into the proximal tubule by the Na^+-H^+ exchanger is buffered by HCO_3^- and converted to H_2CO_3 by carbonic anhydrase, which then dissociates to form CO_2 and H_2O. The increased CO_2 in the tubular fluid freely diffuses into tubular cells, where it reacts with H_2O through the action of cytoplasmic carbonic anhydrase form H_2CO_3 and ultimately HCO_3^- and H^+. The HCO_3^- formed in this process is used to restore or maintain normal plasma pH.

Normally, ≈90% of filtered HCO_3^- (or about 4500 mmol/day) is reclaimed in the proximal tubule. When plasma HCO_3^- concentration increases above ≈28 mmol/L, the capacity of the proximal and distal tubules to reclaim HCO_3^- is exceeded, and HCO_3^- is excreted in urine. Type II RTA is caused by a decreased ability to reabsorb HCO_3^- in the proximal tubules, leading to a decrease in blood pH.

TABLE 36.3 Classification and Characteristics of Simple Acid-Base Disorders

	Primary Change	Compensatory Response	Expected Compensation
Metabolic			
Acidosis	↓ $cHCO_3^-$	↓ PCO_2	$PCO_2 = 1.5\ (cHCO_3^-) + 8 \pm 2$ PCO_2 falls by 1–1.3 mm Hg for each mmol/L fall in $cHCO_3^-$ Last two digits of pH = PCO_2 (e.g., if PCO_2 = 28, pH = 7.28) $cHCO_3^-$ + 15 = last two digits of pH ($cHCO_3^-$ = 15, pH = 7.30)
Alkalosis	↑ $cHCO_3^-$	↑ PCO_2	PCO_2 increases 6 mm Hg for each 10-mmol/L rise in $cHCO_3^-$ $cHCO_3^-$ + 15 = last two digits of pH ($cHCO_3^-$ = 35, pH = 7.50)
Respiratory			
Acidosis			
Acute	↑ PCO_2	↑ $cHCO_3^-$	$cHCO_3^-$ increases by 1 mmol/L for each 10 mm Hg rise in PCO_2
Chronic	↑ PCO_2	↑ $cHCO_3^-$	$cHCO_3^-$ increases by 3.5 mmol/L for each 10 mm Hg rise in PCO_2
Alkalosis			
Acute	↓ PCO_2	↓ $cHCO_3^-$	$cHCO_3^-$ falls by 2 mmol/L for each 10 mm Hg fall in PCO_2
Chronic	↓ PCO_2	↓ $cHCO_3^-$	$cHCO_3^-$ falls by 5 mmol/L for each 10 mm Hg fall in PCO_2

From Narins, R. G., & Gardner, L. B. (1981). Simple acid-base disturbances. *Medical Clinics of North America, 65*, 321–346.

CONDITIONS ASSOCIATED WITH ABNORMAL ACID-BASE STATUS AND ABNORMAL ELECTROLYTE COMPOSITION OF BLOOD

Blood acid-base abnormalities are accompanied by characteristic changes in plasma electrolyte concentrations. Hydrogen ions cannot accumulate without concomitant accumulation of anions, such as Cl^- or lactate, or without exchange for cations, such as K^+ or Na^+. Consequently, plasma electrolyte composition is often determined with measurements of blood gases and pH to assess acid-base disturbances.

Acid-base disturbances are traditionally classified as (1) metabolic acidosis, (2) metabolic alkalosis, (3) respiratory acidosis, or (4) respiratory alkalosis with renal and respiratory compensatory responses (Table 36.3)

An acidosis can only occur as the result of one (or a combination) of three mechanisms: (1) increased addition of acid, (2) decreased elimination of acid, and (3) increased loss of base. Similarly, alkalosis occurs only by (1) an addition of base, (2) decreased elimination of base, and (3) increased loss of acid. This concept can be illustrated by depicting the body as a two-tank vat, one of acid and one of base, with inputs and outputs for each vat (Fig. 36.8).

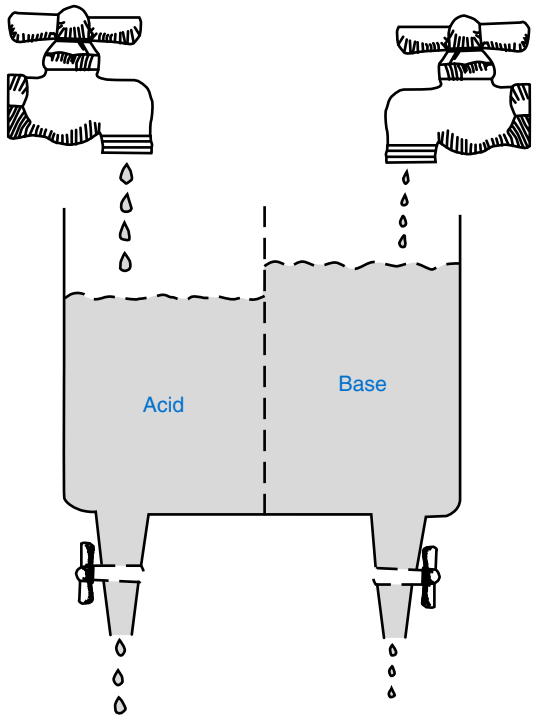


Fig. 36.8 Simple depiction of the body as a two-vat system of acid and base. At equilibrium, input and output from each vat are equal. (From Dufour, D. R. [1995]. Acid-base disorders. In D. R. Dufour & R. H. Christenson [Eds.], *Professional practice in clinical chemistry: a review* [pp. 604–635]. Washington, DC: AACC Press.)

Metabolic Acidosis (Primary Bicarbonate Deficit)

The primary perturbation in metabolic acidosis is decreased plasma bicarbonate. Bicarbonate is "lost" in buffering acid created through increased production (such as the generation of acetoacetic acid in diabetic ketoacidosis) or decreased elimination of acid (seen in renal failure or some types of RTA). Serum bicarbonate can fall with resultant metabolic acidosis if there is a net loss of bicarbonate in the urine (secondary to decreased tubular reabsorption) or in the GI tract (e.g., duodenal/pancreatic fluid loss in severe diarrhea).

The resulting drop in pH from decreased HCO_3^- stimulates respiratory compensation via hyperventilation, resulting in a lower PCO_2 and increased pH.

BOX 36.1 Conditions of Metabolic Acidoses With High and Normal Anion Gaps

Cause	Retained Acid(s)	Other Laboratory Findings
High Anion Gap (MUD PILES)		
Methanol	Formate	↑ Osmolal gap (>15 mOsmol/kg)
Uremia	Sulfuric, phosphoric, organic	↑ Urea and serum creatinine
Diabetes mellitus	Acetoacetate and β-hydroxybutyrate	↑ Plasma and urine glucose, hydroxybutyrate
Paraldehyde toxicity/**P**aracetamol (acetaminophen)	Acetate, chloracetate/pyroglutamate (5-oxoproline)	
Isoniazid or **I**ron	Organic, mainly lactate	
Lactic acidosis	Lactate	
Ethylene glycol	Hippurate, glycolate, oxalate	↑ Osmolal gap (>15 mOsmol/kg), urine oxalate crystals
Salicylate	Salicylate	Respiratory alkalosis
Normal Anion Gap		
Gastrointestinal fluid loss/diarrhea	Primary loss of bicarbonate	Hypokalemia
Acetazolamide	Bicarbonate wasting	
Renal Tubular Acidosis		
Type 1	Decreased H^+ secretion	Hypokalemia
Type 2	Bicarbonate wasting	Hypokalemia
Type 4	Aldosterone deficiency or resistance	Hyperkalemia
Pancreatitis		
Pancreatic fistula	Bicarbonate wasting	

Increased Anion Gap Acidosis (Organic Acidosis)

Metabolic acidoses are classified as those with an increased anion gap or a normal anion gap (Box 36.1). The concept of the **anion gap** was originally devised as a quality control rule when it was noted that if the sum of Cl^- and HCO_3^- values was subtracted from the Na^+ values [$Na^+ - (Cl^- + HCO_3^-)$], the difference, or "gap," averaged 12 mmol/L in healthy subjects. This *apparent* gap is due to unmeasured anions (e.g., proteins, SO_4^{2-}, $H_2PO_4^{2-}$) that are present in plasma. In reality, unmeasured cations (calcium, magnesium, organic cations) should be included in the equation with sodium (and occasionally potassium), but their plasma molar concentrations are relatively small compared with circulating sodium concentration. The anion gap should be assessed in the electrolyte profiles of all patients as it is increased, and is often the first indication of a metabolic acidosis. The gap is also slightly increased in the absence of acidosis by very low calcium, magnesium, or potassium concentrations, because lower concentrations of these "unmeasured" cations will result in lower values of anions (Fig. 36.9).

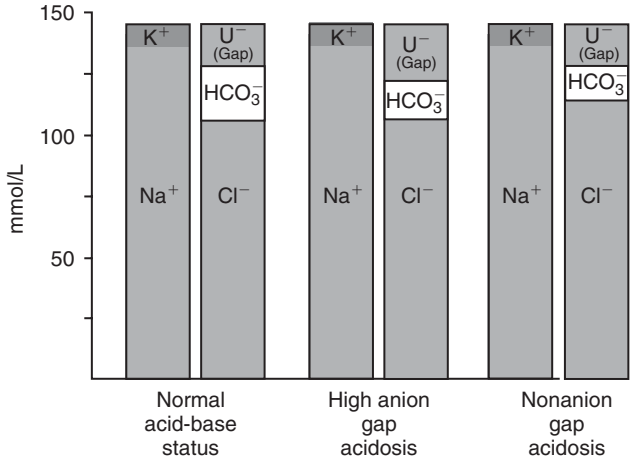


Fig. 36.9 Simple "Gamblegram" depiction of normal gap, anion gap acidosis, and nonanion gap acidosis. Cations, Na^+ and K^+, are in left bar for each condition, whereas measured (Cl^- and HCO_2^-) and unmeasured (U^-) anions are in right bar for each condition.

BOX 36.2 A New Mnemonic for High Anion Gap Acidosis (GOLD MARK)

G Glycols (ethylene and propylene)
O Oxoproline
L L-Lactate
D D-Lactate (short-bowel syndrome)
M Methanol
A Aspirin
R Renal failure
K Ketoacids

All anion gap metabolic acidosis, excluding inborn errors of metabolism, can be explained by one (or a combination) of eight underlying mechanisms included in the common mnemonic device MUDPILES (see Box 36.1) or the updated GOLD MARK (Box 36.2).

Methanol

Although nontoxic itself, methanol is metabolized by the liver to formaldehyde and formic acid, the accumulation of which results in a high anion gap acidosis. Ocular symptoms include optic papillitis, retinal edema, and blindness due to optic nerve atrophy. Coma may result as well. Methanol and other ingested alcohols such as ethylene glycol, ethanol, and isopropanol will increase plasma osmolality. Thus, in the presence of a high anion gap acidosis, determination of the osmolal gap (see Chapter 24) will help determine the source of the unmeasured anion and may indicate specific toxicologic analyses.

Uremia of Renal Failure

Loss of functional renal tubular mass results in decreased ammonia formation, decreased Na^+-H^+ exchange, and decreased GFR. All lead to decreased acid excretion (see Chapter 35), and acidosis usually develops if GFR falls below 20 mL/min.

Ketoacidosis

The pathogenesis of **ketoacidosis** is discussed in detail in Chapter 33. Ketoacids such as β-hydroxybutyrate accumulate during diabetic ketoacidosis, as well as in starvation and alcoholic malnutrition. Ketoacids represented unmeasured anions and their accumulation causes a decrease in HCO_3^- and a high anion gap. Treatment is directed at the process generating the ketoacids—most commonly insulin deficiency or starvation.

Paracetamol (Pyroglutamic Acid)

Chronic paracetamol (acetaminophen in the United States) use by patients with decreased glutathione stores can lead to an accumulation of pyroglutamic acid (5-oxoproline) that results in an anion gap. Pyroglutamic acid is a glutathione dependent intracellular intermediate in the γ-glutamyl cycle. An anion gap acidosis usually only develops in chronic paracetamol users that have other comorbidities such as malnutrition, chronic renal failure, or liver disease.

Isoniazid, Iron, or Ischemia

The "three I's" of high anion gap acidosis share a common feature: the accumulation of organic acids, especially lactic acid. Thus the "three I's" actually represent special cases in the general category of lactic acidosis; in the GOLD MARK mnemonic these three are represented by the "L" for lactic acidosis. Both isoniazid, an antimycobacterial agent commonly used in the treatment or prophylaxis of tuberculosis, and iron toxicity involve the production of toxic peroxides that act as mitochondrial poisons and interfere with normal cellular respiration. Ischemia results from many causes and leads to cellular hypoxia, anaerobic metabolism, and accumulation of organic acids, mainly lactate.

Lactic Acidosis

Lactic acid, present in blood as the lactate ion (pK = 3.86), is an intermediate of carbohydrate metabolism that is derived mainly from muscle cells and erythrocytes (see Chapter 22). It represents the end product of anaerobic metabolism and is normally metabolized by the liver. Therefore blood lactate concentration is affected by the rates of production and metabolism, both of which are dependent on adequate tissue perfusion. An increase in the concentration of lactate to greater than 3 mmol/L defines lactic acidosis.

Lactic acidosis caused by tissue hypoxia is seen most commonly in the setting of shock but can also be seen in severe anemia, sepsis, cardiopulmonary insufficiency, and due to drugs and toxins such as ethanol, methanol, and metformin, among others. Severe tissue oxygen deprivation blocks aerobic oxidation of pyruvic acid in the tricarboxylic acid cycle, resulting in the reduction of pyruvate to lactate. During vigorous exercise, an aerobic threshold may be met (depending on multiple factors) and anaerobic metabolism may develop, with lactate concentrations increasing significantly to ≈12 mmol/L. If the lactate-generating condition can be corrected, lactate is rapidly metabolized to CO_2, which then is eliminated via an intact respiratory system.

A special cause of lactic acidosis attributable to the bacterial production of D-lactate in the intestine is also possible. D-lactic acidosis is a rare cause of an anion gap acidosis and should only be considered in patients with "short-bowel" syndromes or following other GI surgeries resulting in "blind loops." Most lactate measurement methods do not detect D-lactate (for more detail, see Chapter 22).

Ethylene Glycol

Ethylene glycol is metabolized primarily to glycolic and oxalic acids, which leads to an acidosis with high anion and osmolal gaps. Accumulation of toxic metabolites may also contribute to lactic acid production that further contributes to the acidosis. Precipitation of calcium oxalate and hippurate crystals in the urinary tract may lead to acute renal failure. Clinically patients develop a variety of neurologic symptoms that may lead to coma.

Salicylate Intoxication

Acidosis generally occurs with blood salicylate concentrations above 30 mg/dL (2.1 mmol/L). Salicylate, itself an unmeasured anion, alters peripheral metabolism, leading to the production of various organic acids and a high anion gap acidosis. Salicylate toxicity can be complex, with a number of superimposed acid-base disturbances occurring simultaneously, such as respiratory alkalosis due to salicylate induced increased respiration rate.

Normal Anion Gap Acidosis (Inorganic Acidosis)

In contrast to high anion gap acidoses, in which bicarbonate is consumed from buffering excess H^+, the cause of a normal anion gap acidosis is the loss of bicarbonate-rich fluid from the kidney or the GI tract (see Box 36.1). As bicarbonate is lost, more Cl^- ions are reabsorbed with Na^+ or K^+ to maintain electrical neutrality, resulting in hyperchloremia. Normal anion gap acidosis can be divided into *hypokalemic*, *normokalemic*, and *hyperkalemic* acidoses, which can be helpful in the differential diagnosis of this type of disorder.

Gastrointestinal Losses

Diarrhea may cause acidosis as a result of loss of Na^+,K^+, and HCO_3^-. One of the primary exocrine functions of the pancreas is production of HCO_3^- to neutralize gastric contents on entry into the duodenum. If the water, K^+, and HCO_3^- in the intestine are not reabsorbed, a hypokalemic, normal anion gap metabolic acidosis will develop. The resulting hyperchloremia is due to replacement of lost bicarbonate with Cl^- to maintain electrical balance.

Renal Tubular Acidoses

These syndromes are characterized by loss of bicarbonate due to decreased tubular secretion of H^+ (distal or type I RTA) or decreased reabsorption of HCO_3^- (proximal or type II RTA), resulting in bicarbonate "wasting" in the urine. Because the major urine-acidifying power of the kidneys rests in the distal tubules, proximal and distal RTAs may be differentiated by measurement of urine pH. In proximal RTA, urine pH becomes less than 5.5, whereas in distal RTA, the distal tubules are compromised and urine pH is greater than 5.5. Type IV RTA is due to decreased aldosterone concentration or aldosterone resistance, leading to decreased Na^+ reabsorption and thus decreased H^+ and K^+ secretion. It can be differentiated from Type I by increased serum K^+.

Compensatory Mechanisms in Metabolic Acidosis

The buffer systems of blood (mainly the bicarbonate/carbonic acid buffer) minimize pH changes. The compensation mechanisms for metabolic acidoses seek to return arterial blood pH to 7.4. It is important to note that compensation does not overshoot, and in fully **compensated metabolic acidosis**, arterial blood pH will not exceed 7.4.

Respiratory Compensatory Mechanism

The respiratory compensatory mechanism responds immediately to a decrease in pH by increasing the rate and depth of respiration to increase CO_2 elimination. However, the maximal response may take several hours to develop. The increased rate and depth of respiration results in the elimination of carbonic acid as CO_2, a decrease in PCO_2 (hypocapnia), and thus a decrease in $cdCO_2$. Table 36.3 depicts expected compensation in various cases of acidoses and alkaloses, and corresponding laboratory values.

Renal Compensatory Mechanism

If possible, the kidneys respond to restore a normal pH through increased acid excretion and preservation of base (increased rate of Na^+-H^+ exchange, increased ammonia formation, and increased reabsorption of bicarbonate). When the renal compensating mechanisms are functioning, urine acidity and ammonia are increased.

POINTS TO REMEMBER

- In determining the differential for an anion gap acidosis, two mnemonic devices MUDPILES and GOLDMARK can be helpful (see Boxes 36.1 and 36.2).
- Volume depletion is very often accompanied by a hypochloremic metabolic alkalosis.
- Normal anion gap acidosis is due to loss of bicarbonate-rich fluid via the GI tract (diarrhea, pancreatic fistula) or the renal tubules.

METABOLIC ALKALOSIS (PRIMARY BICARBONATE EXCESS)

Alkalosis occurs when excess base is added to the system, base elimination is decreased, or acid-rich fluids are lost (see the At a Glance box). Respiratory compensation is achieved via hypoventilation and subsequent elevation of PCO_2. However, hypoxia limits compensation and prevents PCO_2 from increasing above 55 mm Hg. Above pH 7.55, tetany may develop, even in the presence of a normal serum total calcium concentration as the ionized calcium concentration decreases due to increased binding of calcium by albumin. Measurement of urine Cl^- can be helpful, as causes of metabolic alkalosis fall into Cl^- responsive, Cl^- resistant, and exogenous base categories (see the box "At a Glance" and Fig. 36.4).

At a Glance

Conditions Leading to Metabolic Alkalosis

Chloride Responsive (Urine Cl^- Less Than 10 mmol/L)
- Contraction alkaloses
  - Prolonged vomiting or nasogastric suction
  - Pyloric or upper duodenal obstruction
  - Prolonged or abusive diuretic therapy (loop diuretics)
  - Dehydration
- Posthypercapnic state
- Cystic fibrosis (systemic ineffective reabsorption of Cl^-)

Chloride-Resistant (Urine Cl^- Greater Than 20 mmol/L)
- Mineralocorticoid excess
  - Primary hyperaldosteronism (adrenal adenoma or, rarely, carcinoma)
  - Bilateral adrenal hyperplasia
  - Secondary hyperaldosteronism
  - Congenital adrenal hyperplasia (resulting from adrenal enzyme deficiencies in cortisol production [11β- or 17α-hydroxylase])
- Glucocorticoid excess
  - Primary adrenal adenoma (Cushing syndrome)
  - Pituitary adenoma secreting adrenocorticotropic hormone (Cushing disease)
  - Exogenous cortisol therapy
  - Excessive licorice ingestion
- Bartter syndrome (defective renal Cl^- reabsorption)

Exogenous Base
- Iatrogenic
  - Bicarbonate-containing intravenous fluid therapy
  - Massive blood transfusion (sodium citrate overload)
  - Antacids and cation-exchange resins in dialysis patients
  - High-dose carbenicillin or penicillin (associated with hypokalemia)
- Milk-alkali syndrome

Most causes of Cl^- responsive metabolic alkalosis occur as a result of *hypovolemia* (see the box "At a Glance"). The loss of acidic gastric secretions via prolonged vomiting or nasogastric suction, or the use of certain diuretics, can severely deplete the ECF and result in an acid-base disorder referred to as contraction alkalosis. Several proposed mechanisms

exist but Cl^- depletion is integral. Hydrochloric acid loss from the stomach is particularly important as both hypovolemia *and* hypochloremia can quickly develop. In this case, the kidneys preferentially reabsorb Na^+ to restore volume, and excess HCO_3^- is reabsorbed in the absence of sufficient Cl^- to maintain electrical neutrality. In addition, H^+ and K^+ are secreted in exchange for Na^+. Urine Cl^- will be less than 10 mmol/L in this setting (see Fig. 36.4). Treatment consists of correcting TBW deficit and chloride replacement, usually as NaCl.

Cl^- Resistant Metabolic Alkalosis

This condition is far less common than Cl^- responsive metabolic alkalosis and is almost always associated with an underlying disease (e.g., primary hyperaldosteronism, Cushing syndrome, Bartter syndrome) or with addition of exogenous base. In these conditions, urine Cl^- will usually be greater than 20 mmol/L.

In states of adrenocortical excess, K^+ and H^+ are "wasted" and Na^+ reabsorbed by the kidney due to the effect of aldosterone or cortisol. The hypokalemia produces an alkalosis through enhanced renal excretion of H^+ as NH_4^+ and increased HCO_3^- reabsorption. The decreased tubular K^+ concentration observed in diseases in which endogenous mineralocorticoids, glucocorticoids, or both are increased include primary and secondary hyperaldosteronism, bilateral adrenal hyperplasia, pituitary ACTH-producing adenoma (Cushing disease), and primary adrenal adenomas producing glucocorticoids (Cushing syndrome) or aldosterone.

Exogenous Base

Excess exogenous base states can include citrate toxicity, often following massive blood transfusion (citrate is metabolized to HCO_3^-), aggressive intravenous therapy with bicarbonate solutions, and ingestion of large quantities of antacids in the treatment of gastritis or peptic ulcer *(milk-alkali syndrome).*

Compensatory Mechanisms in Metabolic Alkalosis

The compensatory mechanisms for metabolic alkalosis include both respiratory and renal compensation. The increase in pH depresses the respiratory center, leading to hypoventilation and retention of PCO_2. This compensation is, however, limited by the resultant hypoventilation induced hypoxia. The kidneys respond to the alkalosis by decreased Na^+-H^+ exchange, decreased formation of ammonia, and decreased reclamation of bicarbonate. However, this response is blunted in the presence of hypokalemia and hypovolemia.

RESPIRATORY ACIDOSIS

Respiratory acidosis is produced only by conditions that decrease pulmonary elimination of CO_2, resulting in an increased PCO_2 (hypercapnia) and dCO_2 (and thus H_2CO_3, which dissociates to H^+ and HCO_3^-). This, in turn, causes a decrease in the $cHCO_3^-/cdCO_2$ ratio. Causes of decreased CO_2 elimination (see the box "At a Glance") are classified as acute or chronic.

At a Glance

Conditions Leading to Respiratory Acidosis

Factors That Directly Depress the Respiratory Center
- Narcotics and barbiturates
- CNS trauma, tumors, and degenerative disorders
- Infections of the CNS such as encephalitis and meningitis
- Comatose states such as cerebrovascular accident secondary to intracranial hemorrhage
- Primary central hypoventilation

Conditions That Affect the Respiratory Apparatus
- COPD (most common cause)
- Severe pulmonary fibrosis
- Status asthmaticus (severe)
- Disease of the upper airways such as laryngospasm or tumor
- Pulmonary infection (severe)
- Impaired lung motion secondary to pleural effusion or pneumothorax
- Acute respiratory distress syndrome
- Chest wall disease and chest wall deformity
- Neurologic disorders affecting the muscles of respiration
- Opioids

Others
- Abdominal distention, as in peritonitis and ascites
- Extreme obesity (pickwickian syndrome)
- Sleep disorders such as sleep apnea

Compensatory Mechanisms in Respiratory Acidosis

Compensation for respiratory acidosis occurs immediately via buffers, and over time via the kidneys and, if possible, the lungs. Excess $H_2CO_3^-$ present in blood is buffered to a great extent by hemoglobin and protein buffer systems. The kidneys respond to respiratory acidosis similarly to the way that they respond to metabolic acidosis—namely, with (1) increased Na^+-H^+ exchange, (2) increased ammonia formation, and (3) increased reclamation of bicarbonate. Renal compensation is not effective before 6 to 12 hours and is not maximal for 2 to 3 days. In chronic respiratory acidosis, such as in patients with **chronic obstructive pulmonary disease (COPD)**, full renal compensation may be seen even in patients with very high PCO_2 (>60 mm Hg).

The increased PCO_2, and resultant fall in pH, stimulates the CNS respiratory center, resulting in an increased rate and depth of respiration, provided that the primary defect is not CNS disease. Pulmonary elimination of CO_2 results in a decrease in $cdCO_2$; thus the ratio of $cHCO_3^-/cdCO_2$ and pH approach normal.

RESPIRATORY ALKALOSIS

A decrease in PCO_2 (hypocapnia) and the resulting primary deficit in $cdCO_2$ (respiratory alkalosis) are caused by an increased rate and/or depth of respiration. Excessive pulmonary elimination of CO_2 reduces PCO_2 and causes an increase in pH and the $cHCO_3^-/cdCO_2$ ratio. This shift also

results in a decrease in $cHCO_3^-$, which partially ameliorates the change in pH. Analogous to causes of respiratory acidosis, causes of respiratory alkalosis can be classified as those with a direct stimulatory effect on the respiratory center and those due to effects on the pulmonary system (see the box "At a Glance".)

POINTS TO REMEMBER

- Respiratory acidosis results in an increased PCO_2 secondary to the inability of the lungs to eliminate CO_2.
- Respiratory alkalosis is the result of excess elimination of CO_2 via hyperventilation.
- Respiratory compensatory mechanisms in primary metabolic acidosis or alkalosis settings should not be interpreted as respiratory disorders.

At a Glance

Factors Causing Respiratory Alkalosis

Nonpulmonary Stimulation of Respiratory Center
- Anxiety, hysteria
- Febrile state
- Gram-negative septicemia
- Metabolic encephalopathy (e.g., secondary to liver disease)
- CNS infection such as meningitis and encephalitis
- Cerebrovascular accident
- Intracranial surgery
- Hypoxia (e.g., severe anemia, high altitudes [acute condition])
- Drugs and agents such as salicylates, catecholamines, and progesterone
- Pregnancy, mainly third trimester
- Hyperthyroidism

Pulmonary Disorders
- Pneumonia
- Pulmonary emboli
- Interstitial lung disease
- Large right-to-left shunt (PCO_2 <50 mm Hg)
- Congestive heart failure
- Respiratory compensation after correction of metabolic acidosis

Others
- Ventilator-induced hyperventilation

Compensatory Mechanisms in Respiratory Alkalosis

The compensatory mechanisms respond to respiratory alkalosis in two stages. In the first stage, erythrocyte and tissue buffers provide H^+ ions that consume a small amount of HCO_3^-. The second stage becomes operational in prolonged respiratory alkalosis and depends on renal compensation as described for metabolic alkalosis (decreased reclamation of bicarbonate).

REVIEW QUESTIONS

1. Total body water (TBW) makes up approximately 60% of adult body weight. The majority of TBW is located in which one of the following compartments?
   a. Extracellular fluid (ECF)
   b. Intracellular fluid (ICF)
   c. Interstitial fluid
   d. Plasma
2. The anion gap, an estimate of anions not directly measured in serum, is correctly calculated by which one of the following formulae?
   a. $(Na) + (Cl + HCO_3)$
   b. $(Na) - (Cl + HCO_3)$
   c. $(Cl) + (HCO_3 + Na)$
   d. $(HCO_3 - Na) - (Cl)$
3. Hypernatremia commonly occurs in:
   a. decreased production of antidiuretic hormone (ADH).
   b. decreased aldosterone.
   c. the syndrome of inappropriate ADH (SIADH) secretion.
   d. decreased cortisol.
4. A decrease in PCO_2 and a resulting primary deficit in $cdCO_2$ is referred to as
   a. metabolic acidosis.
   b. metabolic alkalosis.
   c. respiratory acidosis.
   d. respiratory alkalosis.
5. Low blood volume will cause synthesis of this hormone by the kidneys, which in turn causes production of aldosterone, which will increase kidney reabsorption of sodium and retention of water.
   a. Insulin
   b. Renin
   c. Antidiuretic hormone (ADH)
   d. Erythropoietin
6. The use of spironolactone in addition to an *increased* dietary intake of this particular electrolyte can lead to this disorder characterized by mental confusion, bradycardia, and possible eventual cardiac arrest.
   a. Sodium; hypernatremia
   b. Sodium; hyponatremia
   c. Potassium; hyperkalemia
   d. Potassium; hypokalemia

7. Which one of the following is the primary cause of respiratory acidosis?
   a. Hyperventilation
   b. Overdose of antacids
   c. Chronic obstructive pulmonary disease (COPD)
   d. Uncontrolled diabetes mellitus
8. Respiratory compensation of a metabolic alkalosis involves hypoventilation to increase PCO_2. Which of the following is the maximum PCO_2 achievable, and what limits PCO_2 from increasing beyond that threshold?
   a. 55 mm Hg, hypoxia
   b. 65 mm Hg, CO_2 solubility
   c. 80 mm Hg, diffusion
   d. 15 mm Hg, tetany/seizures
9. In the absence of an acid/base disorder, the normal ratio of bicarbonate to dissolved carbon dioxide in whole blood is approximately
   a. 10:1
   b. 1:20
   c. 20:1
   d. 1:10
10. Of the following causes of metabolic acidosis, which will present with an anion gap within the reference interval?
   a. Uremia of kidney failure
   b. Renal tubular acidosis type II
   c. D-Lactic acidosis
   d. Acetaminophen ingestion

SUGGESTED READINGS

Adrogué, H. J., & Madias, N. E. (2000). Hypernatremia. *The New England Journal of Medicine*, *342*, 1493–1499.

Dufour, D. R. (1995). Acid-base disorders. In R. Dufour, & R. H. Christenson (Eds.), *Professional practice in clinical chemistry: a review* (pp. 604–635). Washington, DC: AACC Press.

Farese, R. V., Jr., Biglieri, E. G., Shackleton, C. H., Irony, I., & Gomez-Fontes, R. (1991). Licorice-induced hypermineralocorticoidism. *The New England Journal of Medicine*, *325*, 1223–1227.

Gandhi, M. J. (2007). *Refractory hyponatremia with lung cancer. Tietz's applied laboratory medicine* (2nd ed). 79–87.

Hoye, A., & Clark, A. (2003). Iatrogenic hyperkalaemia. *Lancet*, *361*, 2124.

Kadakia, S. C. (1995). D-lactic acidosis in a patient with jejunoileal bypass. *Journal of Clinical Gastroenterology*, *20*, 154–156.

Lyon, A. W., & Baskin, L. B. (2003). Pseudohyponatremia in a myeloma patient: direct electrode potentiometry is a method worth its salt. *Laboratory Medicine*, *5*, 357–360.

Mehta, A. N., Emmett, J. B., & Emmett, M. (2008). Gold mark: an anion gap mnemonic for the 21st century. *Lancet*, *372*, 892.

Mendu, D. A., Fleisher, M., McCash, S. I., Pessin, M. S., & Ramanathan, L. V. (2015). D-lactic acidosis mediated neuronal encephalopathy in acute lymphoblastic leukemia patient: an under diagnosis. *Clinica Chimica Acta*, *441*, 90–99.

Spasovski, G., Vanholder, R., Allolio, B., Annane, D., Ball, S., et al. (2014). Clinical practice guideline on diagnosis and treatment of hyponatremia. *European Journal of Endocrinology*, *170*, G1–G47.

37

Liver Disease

*William Rosenberg, Tony Badrick, Sudeep Tanwar**

OBJECTIVES

1. Define the following:
   - Acinus
   - Ascites
   - Bilirubin
   - Cholestasis
   - Cirrhosis
   - Gallbladder
   - Hepatic lobule
   - Hepatitis
   - Hepcidin
   - Jaundice
   - Model for end-stage liver disease (MELD) score
   - Portal hypertension
   - Portal triad
2. Describe the microscopic and macroscopic anatomy of the hepatic system, including functional units, lobules, hepatobiliary ducts, blood supply, intrinsic cell types, and functional organization.
3. Detail the major functions of the liver, including the mechanisms of bilirubin excretion, substances synthesized and their functions, and substances metabolized and their sources.
4. List the causes and consequences of each of the following clinical manifestations of liver disease:
   - Disordered hemostasis
   - Liver enzyme release
   - Jaundice
   - Portal hypertension
5. List the enzymes synthesized in the liver, their subcellular location, tissue specificity, and mechanisms of enzyme release and clearance.
6. Compare and contrast acute liver injury with chronic liver injury, including manifestations of injury, target cells involved, specific means of cell death, and laboratory analyses used to distinguish between them.
7. List and describe five types of viral hepatitis, including type of virus, mode of transmission, involvement in liver cancer, methods of prevention, and laboratory tests used to diagnose the presence of the virus.
8. Describe the following types of liver disease, including causes, clinical presentation, and laboratory findings:
   - Acute alcoholic hepatitis
   - Alcoholic liver disease
   - Autoimmune hepatitis
   - Cholestatic hepatitis
   - Cholestatic liver diseases
   - Cirrhosis
   - Drug-induced liver injury (DILI)
   - Ischemic hepatitis
   - Primary biliary cholangitis (PBC)
   - Primary sclerosing cholangitis (PSC)
   - Toxic hepatitis
9. State the formula for calculating a MELD score; calculate a MELD score given appropriate values and determine prognosis in cirrhosis.
10. Describe hepatocellular carcinoma (HCC) including prevalence, risk factors, screening tests, and laboratory findings.
11. Assess and solve case studies related to liver disease.

KEY WORDS AND DEFINITIONS

Alcoholic hepatitis An acute or chronic degenerative and inflammatory lesion of the liver in the alcoholic that is potentially progressive though sometimes reversible.

Alcoholic liver disease Liver injury caused by excessive ethanol (alcohol) ingestion.

Apoptosis Programmed cell death as signaled by the nuclei in normally functioning human and animal cells when age or state of cell health and condition dictates.

Ascites Serous fluid that accumulates in the abdominal cavity.

*The authors gratefully acknowledge the original contributions by Drs. D. Robert Dufour, Keith G. Tolman, and Robert Rej, upon which portions of this chapter are based.

Autoimmune hepatitis (AIH) A form of hepatitis, usually with hypergammaglobulinemia and serum autoantibodies.
Bile A greenish-yellow fluid secreted by the liver and stored in the gallbladder.
Cholestasis Suppression of the normal flow of bile.
Chronic hepatitis A collective term for a clinical and pathologic syndrome that has several causes and is characterized by varying degrees of hepatocellular necrosis and inflammation for at least 6 months.
Cirrhosis Liver disease characterized pathologically by loss of the normal microscopic lobular architecture with fibrosis and nodular regeneration.
Drug-induced liver injury (DILI) A disease of the liver that is caused by prescribed medications, over-the-counter medications, vitamins, hormones, herbs, illicit ("recreational") drugs, and environmental toxins.
Fulminant hepatic failure A condition characterized by the development of severe liver injury with impaired synthetic capacity and encephalopathy in patients with previous normal liver or at least well-compensated liver disease.
Fulminant hepatitis A rare and frequently fatal form of acute hepatitis B in which the patient's condition rapidly deteriorates with hepatic encephalopathy, necrosis of the hepatic parenchyma, coagulopathy, renal failure, and coma.
Gallstones Solid formations in the gallbladder, most commonly composed of cholesterol and bile salts.
Hemochromatosis A rare genetic disorder, due to abnormalities in genes that regulate iron metabolism.
Hepatic failure A condition of severe liver dysfunction that is accompanied by a loss of normal liver functions.
Hepatitis Inflammation of the liver. Typically divided into acute (duration of weeks to months) and chronic (lasting for more than 6 months).
Hepatocellular carcinoma (HCC) A cancer arising from hepatocytes.
Hepatocyte An epithelial liver cell that performs most of the synthetic and metabolic functions of the liver.
Hepcidin A hormone produced by liver cells, in response to signals from a complex pathway, that decreases iron mobilization across intestinal and macrophage membranes, preventing excess iron in the blood (see also Chapter 28).
Jaundice A clinical finding characterized by hyperbilirubinemia and deposition of pigment in the skin, mucous membranes, and sclera with resulting yellow appearance; also called *icterus*.
Model for End-Stage Liver Disease (MELD) Score A scoring system for assessing the severity of chronic liver disease.
Necrosis The sum of the morphologic changes indicative of cell death and caused by the progressive degradative action of enzymes; it may affect groups of cells or part of a structure or an organ.
Nonalcoholic fatty liver disease (NAFLD) A condition characterized by the buildup of extra fat in liver cells that is not caused by alcohol.
Nonalcoholic steatohepatitis (NASH) An inflammatory disease of the liver of uncertain pathogenesis.
Portal hypertension Any increase in the pressure in the portal vein (which carries venous blood from the intestines and spleen to the liver) due to anatomic or functional obstruction (e.g., cirrhosis) to blood flow in the portal venous system.
Primary biliary cholangitis (PBC) A rare form of liver disease, formerly called primary biliary cirrhosis, that results in the irreversible destruction of the small bile ducts within the liver.
Primary sclerosing cholangitis (PSC) A chronic, nonbacterial inflammatory narrowing of the bile ducts (usually the larger ducts outside of the liver).
Reye syndrome A sudden, sometimes fatal, disease of the brain (encephalopathy) caused by specific forms of acute injury to the liver.
Sjögren syndrome A systemic autoimmune disease in which immune cells attack and destroy the glands that produce tears and saliva.
Varices Enlarged and tortuous veins, most commonly found in the esophagus (termed *esophageal varices*).
Viral hepatitis Liver inflammation caused by viruses. Specific hepatitis viruses have been labeled A, B, C, D, and E.
Viral load The measurement of the amount of virus in the blood. It is expressed as the copies of virus per milliliter of body fluid and used to guide treatment decisions and monitor response to treatment.
Wilson disease An autosomal recessive disorder associated with excessive quantities of copper in the tissue, particularly the liver and central nervous system.

ANATOMY OF THE LIVER

The adult liver weighs approximately 1.2 to 1.5 kg. It is located beneath the diaphragm in the right upper quadrant of the abdomen and is protected by the ribs and held in place by ligamentous attachments.

Gross Anatomy

The liver is divided into left and right anatomic lobes by the falciform ligament (Fig. 37.1). It has a dual blood supply. The portal vein carries blood from the spleen and nutrient-enriched blood from the gastrointestinal (GI) tract supplying

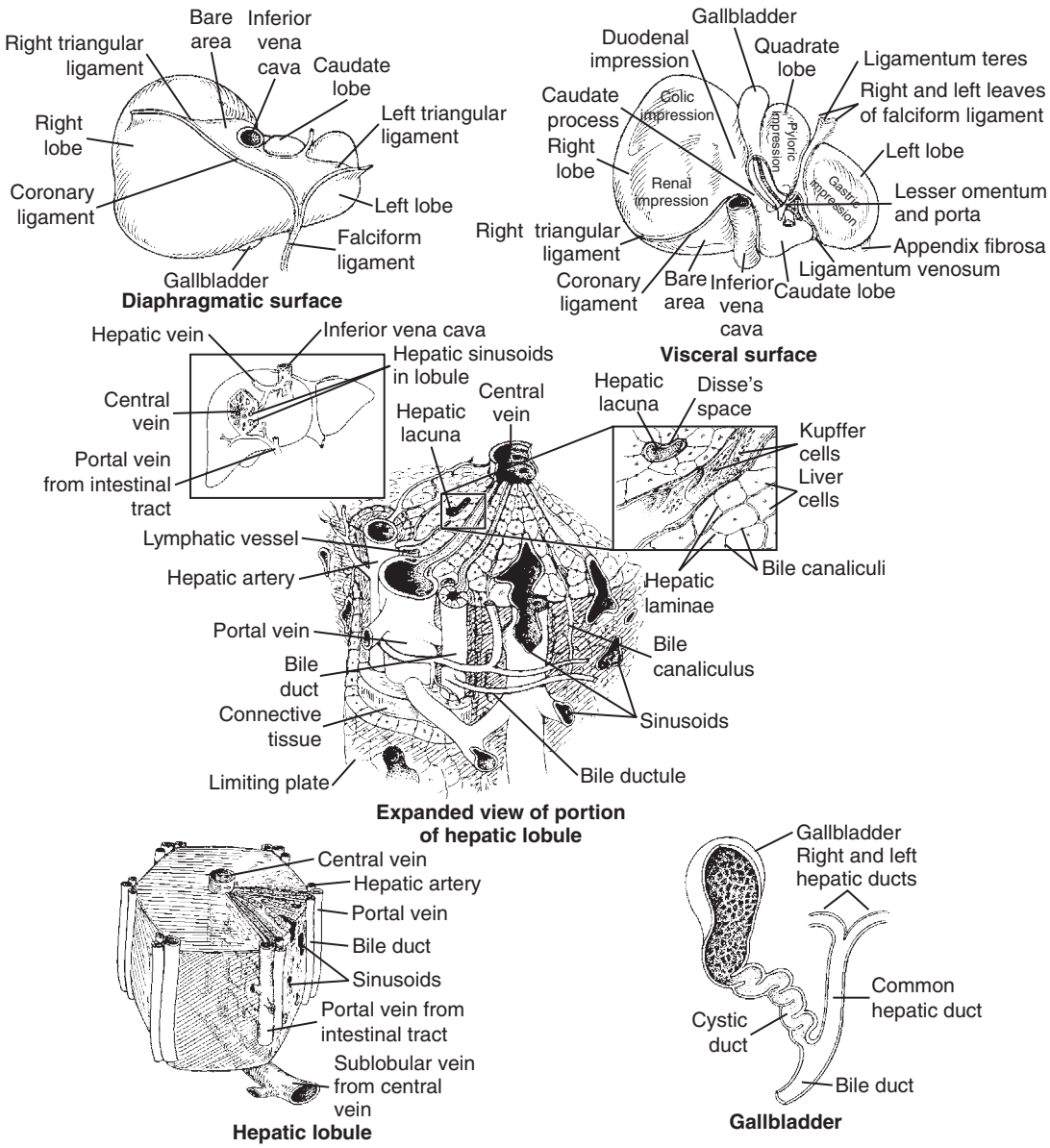


Fig. 37.1 Structure of the liver. (From Dorland, W. A. N. [2003]. *Dorland's Illustrated Medical Dictionary* [30th ed.]. Philadelphia: Saunders, plate 26.)

approximately 70% of the blood supply; the hepatic artery, a branch of the celiac axis, provides oxygen-enriched arterial blood. These two blood supplies merge and flow into the sinusoids that course between individual hepatocytes. Venous drainage from the liver ultimately converges into the right and left hepatic veins, which exit on the posterior surface of the liver and join the inferior vena cava near its entry into the right atrium.

Biliary drainage originates at the bile canaliculi that merge to form ductules that drain into the intrahepatic bile ducts, which ultimately join to form the right and left hepatic bile ducts to exit the liver at the porta hepatis forming the common hepatic duct. The hepatic duct is joined by the cystic duct that drains the gallbladder, to form the common bile duct (see Fig. 37.1) which then enters the duodenum (usually with the pancreatic duct) at the ampulla of Vater. The gallbladder, which is located on the undersurface of the right lobe of the liver, stores and concentrates 30 to 50 mL of **bile**.

Microscopic Anatomy

The functional anatomic unit of the liver is the acinus, adjacent to the portal triad (which consists of a branch of each of the portal vein, hepatic artery, and bile duct). Each acinus is supplied by a terminal branch of the portal vein and of the hepatic artery and is drained by a terminal branch of the bile duct. The blood vessels radiate toward the periphery, forming sinusoids, which perfuse the liver and ultimately drain into the central (terminal) hepatic vein (Fig. 37.2). The sinusoids are lined by fenestrated endothelial cells (which allow free filtration of blood) and phagocytic Kupffer cells (see Fig. 37.1) derived from blood monocytes.

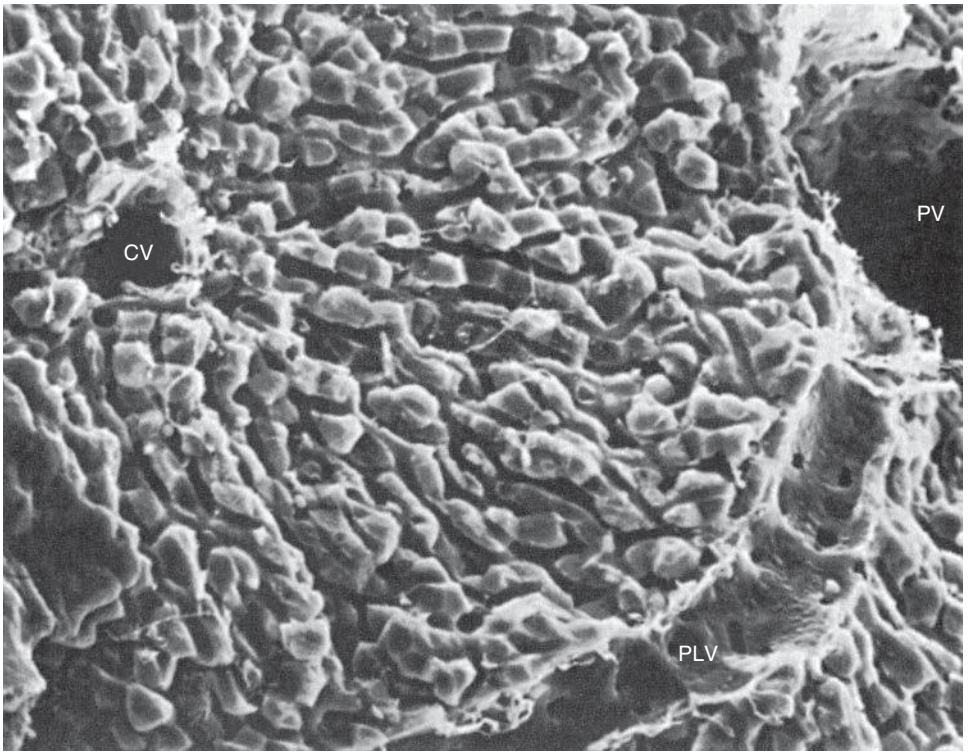


Fig. 37.2 A low-magnification scanning electron micrograph depicting a portion of a liver lobule from a rat liver. *CV*, Central vein; *PLV*, perilobular venules; *PV*, portal vein. (From Zakim, O., & Boyer, T. D. [1996]. *Hepatology: A Textbook of Liver Disease* [3rd ed., p. 9]. Philadelphia: WB Saunders.)

Hepatocytes are the major functioning cells in the liver and are responsible for approximately 70% of liver mass. They perform most of the metabolic and synthetic functions of the liver. Hepatic stellate cells are located between the endothelial lining of the sinusoids, and the hepatocytes are within a small cleft referred to as the space of Disse. In their normal, quiescent state, stellate cells serve as a site of storage for fat-soluble vitamins, particularly vitamin A. When stimulated, stellate cells are morphologically and functionally transformed. They synthesize collagen and are the cells responsible for fibrosis and, eventually, cirrhosis. They also synthesize nitric oxide, which helps to regulate intrahepatic blood flow.

The blood supply to each acinus is zonal (Fig. 37.3). Zone 1, the area immediately adjacent to the portal tract, is enriched with lysosomes and mitochondria and forms a base for hepatic regeneration. Zone 2 predominantly contains hepatocytes that perform the major metabolic functions of the liver. The periphery of the acinus, zone 3, is enriched with endoplasmic reticulum, is active metabolically, and has relatively low oxygen tension. This area is most susceptible to injury.

Ultrastructure of the Hepatocyte

Hepatocytes contain a well-developed organelle substructure (Fig. 37.4). Mitochondria, which constitute approximately 18% of hepatocyte volume, are the sites of oxidative phosphorylation and energy production. The smooth endoplasmic reticulum contains microsomes that are involved in bilirubin conjugation; detoxification (cytochrome P_{450}–dependent isoenzymes); and steroid, cholesterol, and bile acid synthesis. Several microsomal enzymes, including γ-glutamyltransferase (GGT), are induced by many drugs and inhibited by others. This is the site of most drug metabolism and many important drug interactions.

Peroxisomes are found near the smooth endoplasmic reticulum and contain oxidases that use molecular oxygen to modify a variety of substrates, leading to the production of hydrogen peroxide. They also contain catalase, which decomposes hydrogen peroxide. Peroxisomes catalyze the β-oxidation of fatty acids that have 7 to 18 chain lengths. Approximately 5% to 20% of the metabolism of ethanol also occurs in the peroxisomes. Lysosomes are dense organelles that contain hydrolytic enzymes that act as scavengers. Deposition of iron, lipofuscin, bile pigments, and copper occurs in the lysosomes. The Golgi apparatus lies near the canaliculus and is involved in the secretion of various substances, including bile acids and albumin.

BIOCHEMICAL FUNCTIONS OF THE LIVER

The liver is involved in various excretory, synthetic, and metabolic functions. Clinical laboratories perform numerous tests that are useful in the biochemical assessment of these functions.

Hepatic Excretory Function

Organic compounds of both endogenous and exogenous origin are extracted from the sinusoidal blood, biotransformed, and excreted into the bile or urine. Assessment of

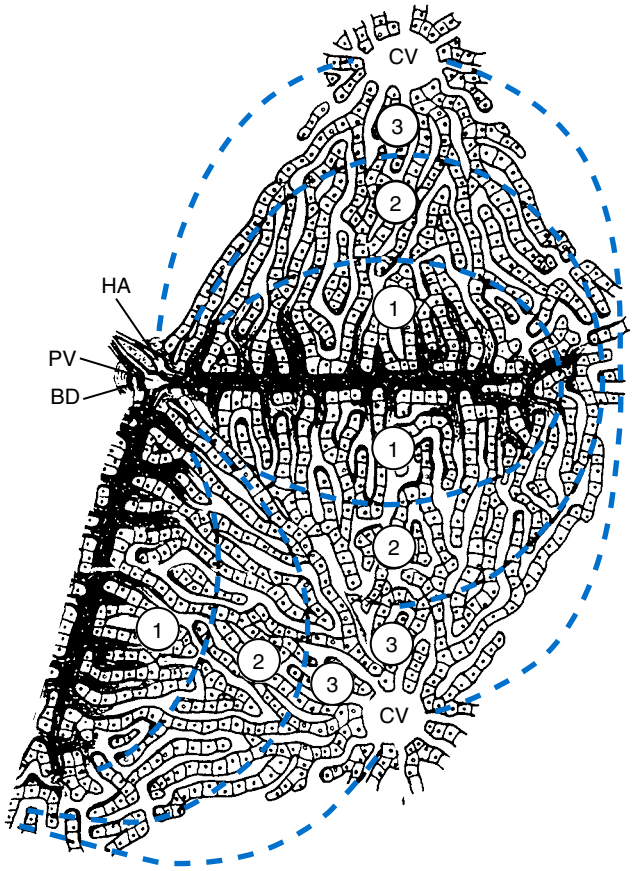


Fig. 37.3 Blood supply of the simple liver acinus. Zones 1, 2, and 3 indicate corresponding volumes in a portion of an adjacent acinar unit. Oxygen tension and the nutrient level in the blood in sinusoids decrease from zone 1 through zone 3. *BD*, Bile duct; *CV*, central vein; *HA*, hepatic artery; *PV*, portal vein. (From Zakim, O., & Boyer, T. D. [1996]. *Hepatology: A Textbook of Liver Disease* [3rd ed., p. 10]. Philadelphia: WB Saunders.)

this excretory function provides valuable clinical information. The most frequently used tests involve the measurement of plasma concentrations of endogenously produced compounds, such as bilirubin and bile acids. In specialist centers, these tests may be augmented by determination of the rate of clearance of exogenous compounds, such as aminopyrine, lidocaine, and caffeine.

Bilirubin

Bilirubin derived from heme is mainly a product of red blood cell turnover. It is extracted and biotransformed in the liver and excreted in bile and urine. The chemistry, biochemistry, and analytical methodology for bilirubin and related compounds are discussed in Chapter 28.

Bilirubin is transported from sites of production (mainly the spleen) and is loosely bound to albumin in its native, unconjugated form. Bilirubin is transported across the hepatocyte membrane and is rapidly conjugated with glucuronic acid to produce bilirubin glucuronides, which are then excreted into bile by an energy-dependent process.

The measurement of unconjugated (indirect) and conjugated (direct) bilirubin fractions can be used to identify the likely cause of hyperbilirubinemia into prehepatic, hepatic, and posthepatic causes.

Bile Acids

Regulation of bile acid metabolism is a major function of the liver. Cholesterol homeostasis is in large part maintained by the conversion of cholesterol to bile acids and subsequent regulation of bile acid metabolism. Bile acids facilitate both hepatic excretion of cholesterol and solubilization of lipids for intestinal absorption.

Hepatic Synthetic Function

The liver has extensive synthetic capacity and plays a major role in the regulation of protein, carbohydrate, and lipid metabolism (see Chapters 18, 22, and 23). The liver is vital to the regulation of blood glucose and to the synthesis of proteins, triglycerides, fatty acids, cholesterol, and bile acids.

Protein Synthesis

The liver is the primary site of the synthesis of most plasma proteins (see Chapter 18) in the rough endoplasmic reticulum of hepatocytes, followed by release into the hepatic sinusoids. Many factors may affect plasma protein concentrations. The extensive functional reserve and relatively long half-life of proteins such as albumin mean that reduction in protein production is indicative of extensive liver damage. For this reason, the sensitivity and specificity of protein concentrations for diagnosis of liver disease are far from ideal. Short-lived hepatic proteins (such as transthyretin and prothrombin) fall quickly, whereas those of proteins with longer half-lives are normal or minimally changed. Serial determination of plasma proteins provides prognostic information; for example, worsening of prothrombin time (PT) during acute **hepatitis** suggests a poor prognosis.

Plasma proteins

Albumin. Albumin, the most commonly measured plasma protein, is synthesized exclusively by the liver. With liver disease, hypoalbuminemia is noted primarily in cirrhosis, **autoimmune hepatitis (AIH)**, and **alcoholic hepatitis**. One important consideration in measurement of albumin is the inaccuracy of dye-binding methods in patients with liver disease. Although bromocresol green measurements tend to overestimate albumin concentration at low concentrations, bromocresol purple methods give falsely low values in patients with **jaundice** because of the interference of bilirubin at the site of binding.

Transthyretin. This protein has a short half-life of 24 to 48 hours, making it a sensitive indicator of current synthetic ability. Transthyretin is typically decreased in cirrhosis (among other conditions) as a result of decreased synthesis. It is more commonly used as a measurement of nutritional status.

Immunoglobulins. Plasma immunoglobulin (Ig) concentrations are commonly increased in cirrhosis, AIH, and **primary biliary cirrhosis (PBC)**, but they are normal in most other types of liver disease. IgG is increased in AIH and cirrhosis; IgM is increased in PBC. IgA tends to be increased

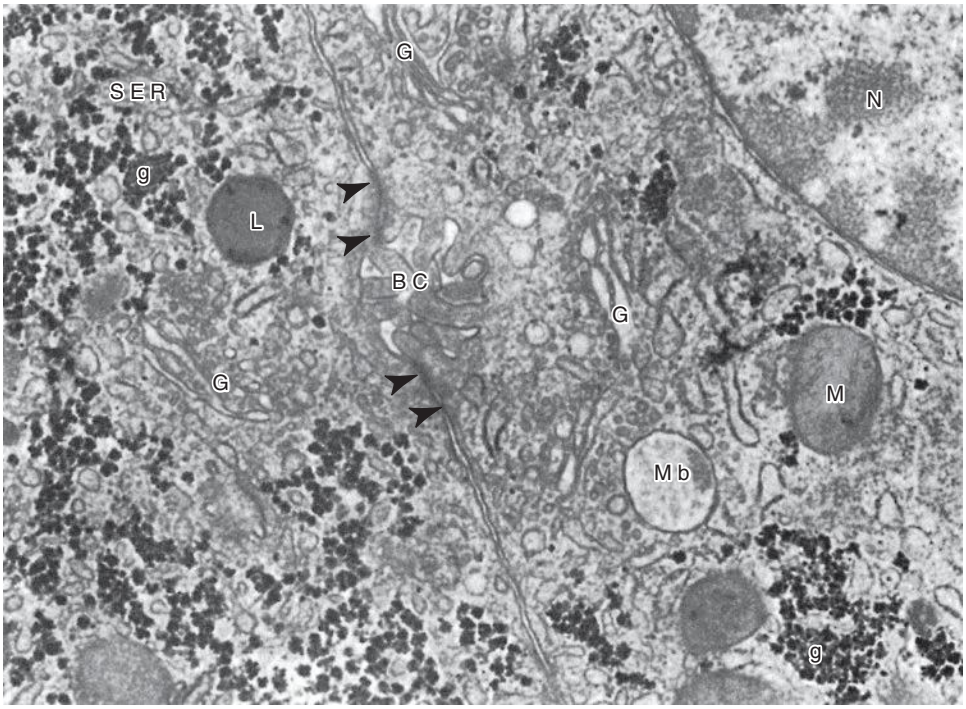


Fig. 37.4 Portions of two human liver cells showing the relationship of the organelles and a typical bile canaliculus *(BC). Arrowheads* indicate light junctions. *g*, Glycogen; *G*, Golgi; *L*, lysosome; *M*, mitochondria; *Mb*, microbody; *N*, nucleus; *SER*, smooth endoplasmic reticulum. (From Zakim, O., & Boyer, T. D. [1996]. *Hepatology: A Textbook of Liver Disease* [3rd ed., p. 20]. Philadelphia: WB Saunders.)

in all types of cirrhosis. None of these findings are specific, and they are seldom used in the diagnosis of liver disease, but serial measurement of IgG levels can be used to monitor disease activity in AIH.

Ceruloplasmin. The concentration of this protein is decreased in Wilson disease, cirrhosis, and many causes of chronic hepatitis, but it may be increased by inflammation, cholestasis, **hemochromatosis**, pregnancy, and estrogen therapy (see section on Wilson disease).

Alpha$_1$-antitrypsin. Concentrations of this protein, which is the major serine protease inhibitor (serpin) in plasma, is decreased in homozygous deficiency and cirrhosis, and is increased by acute inflammation. It is discussed in greater detail later in the section on alpha$_1$-antitrypsin (AAT) deficiency.

Alpha fetoprotein. The concentration of this protein, a normal component of fetal blood, falls to adult values by 1 year of age. Mild increases are seen in patients with acute and chronic hepatitis and indicate hepatocellular regeneration. It is present at higher concentrations in **hepatocellular carcinoma (HCC)**.

Coagulation proteins. The coagulation proteins, inhibitors of the coagulation system, including antithrombin, protein C, and protein S, are synthesized in the liver. Some of the coagulation factors (II, VII, IX, and X) or coagulation inhibitors (protein C and S) require vitamin K for posttranslational carboxylation within the hepatocyte. Activated protein C in plasma inhibits coagulation by inactivating factors V and VIII. Parenchymal liver disease of sufficient severity to impair protein synthesis or obstructive liver disease, sufficient to impair intestinal absorption of vitamin K, is therefore a potential cause of bleeding disorders. Because of the great functional reserve of the liver, failure of hemostasis usually does not occur except in severe or long-standing liver disease.

Lipid and lipoprotein synthesis. The liver plays a key role in the metabolism of lipids and lipoproteins (see Chapter 23). On a daily basis, approximately 33% of the fatty acids originating from adipose tissue enter the liver, where they undergo esterification into triglycerides or are oxidized. Oxidation is favored in the fasting state and esterification is favored in the nonfasting state. Excessive esterification results in fatty liver, a disorder in which excess triglycerides are deposited in large vacuoles that displace other cellular components. Most cholesterol is synthesized endogenously, in the liver. Hepatic cholesterol and cholesterol of dietary origin enter the hepatic pool, where they are converted to bile acids, incorporated into lipoproteins, or used in the synthesis of liver cell membranes. The relative rates of secretion of bile acids, cholesterol, and lecithin are important factors in the pathogenesis of cholesterol gallstones.

Urea synthesis. Patients with end-stage liver disease may have low concentrations of urea in plasma (see Chapter 35). The rate of urea excretion in urine is lower in these patients than that in healthy individuals. In addition, plasma concentrations of urea precursors—ammonia and amino acids—are elevated. Lower specific activities of enzymes involved in urea synthesis are also seen. These findings suggest that patients with liver disease have an impaired ability to metabolize protein nitrogen and to synthesize urea. The rate of hepatic urea

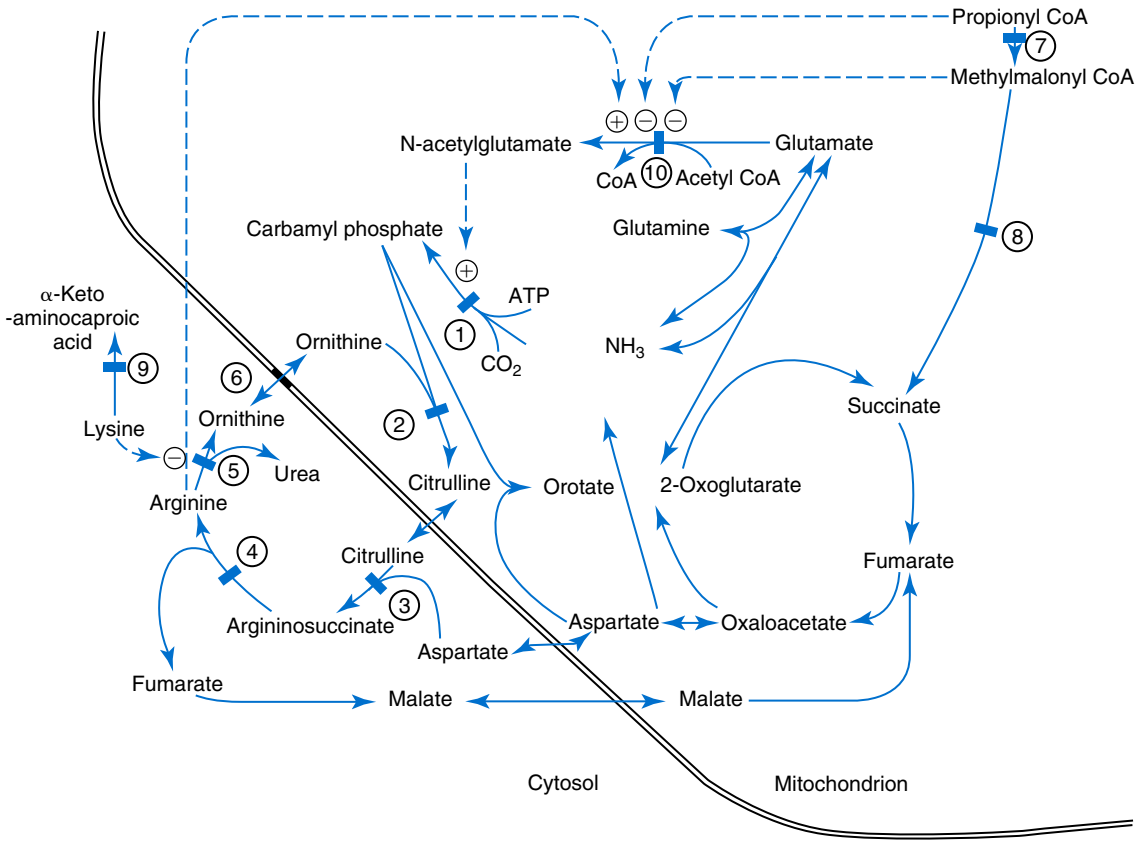


Fig. 37.5 Major metabolic pathways for the use of ammonia by the hepatocyte. *Solid bars* indicate the sites of primary enzyme defects in various metabolic disorders associated with hyperammonemia: *(1)* carbamyl phosphate synthetase 1, *(2)* ornithine transcarbamylase, *(3)* argininosuccinate synthetase, *(4)* argininosuccinate lyase, *(5)* arginase, *(6)* mitochondrial ornithine transport, *(7)* propionyl coenzyme-A *(CoA)* carboxylase, *(8)* methylmalonyl CoA mutase, *(9)* L-lysine dehydrogenase, and *(10)* N-acetyl glutamine synthetase. *Dotted lines* indicate the site of pathway activation (+) or inhibition (–). (From Flannery, O. B., Hsia, Y. E., & Wolf, B. [1982]. Current status of hyperammonemia syndromes. *Hepatology, 2,* 495–506.)

synthesis also depends on exogenous intake of nitrogen and on endogenous protein catabolism.

Hepatic Metabolic Function

A recurring theme is the central importance of the liver in metabolic and regulatory pathways. The functional expression of the complex, integrated organelle structure includes the metabolism of drugs (activation and detoxification) and the disposal of exogenous and endogenous substances, such as galactose and ammonia. In addition, metabolic abnormalities due to specific inherited enzyme deficiencies can affect the liver. A classic example is galactosemia. In this condition, congenital absence of galactose 1-phosphate uridyltransferase allows accumulation of the toxic metabolite galactose 1-phosphate, which causes injury to the liver, brain, and kidneys.

Ammonia Metabolism

Biochemistry and physiology. The major source of circulating ammonia is the GI tract. Plasma ammonia concentration in the hepatic portal vein is typically 5-fold to 10-fold higher than that in the systemic circulation. It is derived from the action of bacterial proteases, ureases, and amine oxidases on the contents of the colon and from the hydrolysis of glutamine in both the small and large intestines. Under normal circumstances, most of the portal vein ammonia load is metabolized to urea in hepatocytes through the Krebs-Henseleit (urea) cycle during the first pass through the liver; this process includes intramitochondrial and cytosolic enzyme-catalyzed steps (Fig. 37.5).

Ammonia enters the tissue of the central nervous system by passive diffusion. The rate of entry increases in proportion to the plasma concentration and is dependent on pH. Ammonia crosses the blood-brain barrier more readily than the ammonium ion. Fasting venous plasma ammonia concentration is useful in the differential diagnosis of encephalopathy to determine if it is hepatic in origin. In acute liver failure, ammonia concentrations in excess of 200 μmol/L are associated with cerebral edema and a poor prognosis.

Carbohydrate Metabolism

Because the liver is a major processor of dietary and endogenous carbohydrates, liver disease affects carbohydrate metabolism in a variety of ways (see Chapters 22 and 33). However, none of the conventional modes of evaluating carbohydrate metabolism have value in the diagnosis of liver disease. Because the liver is the major site of both glycogen storage

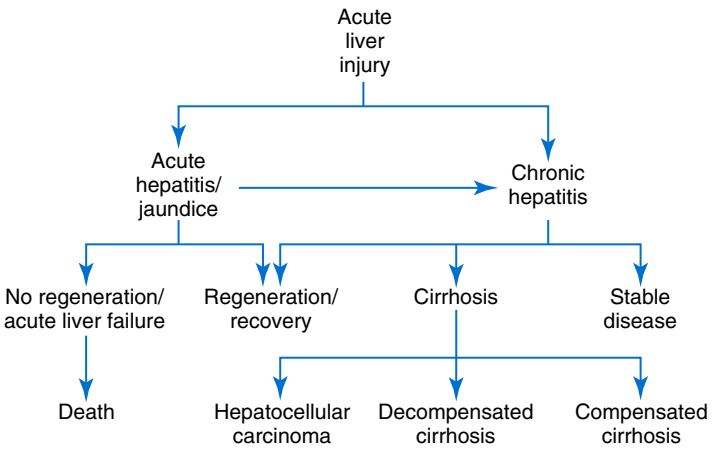


Fig. 37.6 Natural history of liver disease. With acute injury to the liver, several outcomes are possible. In many individuals, damage is clinically inapparent and recovery occurs with clearance of the causative agent. In some, clinical acute hepatitis occurs. In most of these, clearance of the causative agent results in complete recovery; in a small minority, damage is so severe that acute liver failure (fulminant hepatitis) develops, which is usually fatal without liver transplantation. A variable percentage of persons with acute liver injury (dependent on the cause) progress to chronic hepatitis. In some, recovery eventually occurs naturally or following treatment of the underlying cause. Among those in whom chronic hepatitis persists, many will never progress to cirrhosis. Most of those who do will remain well for many years, but approximately 3% per year develop decompensated cirrhosis (bleeding varices, ascites, hepatic encephalopathy) or hepatocellular carcinoma. These are the most common causes of death from liver disease.

and gluconeogenesis, hypoglycemia is a common complication in certain liver diseases, particularly **Reye syndrome**, **fulminant hepatic failure**, advanced cirrhosis, and HCC.

Hepatic Storage Function

The liver serves as a major site for storage of energy-rich carbohydrate substrates. Hepatic glycogen is released to other tissues when the need exists.

Nutritional and Metabolic Abnormalities

Severe metabolic and nutritional derangements have been observed in cirrhotic patients, including alterations in glucose metabolism caused by insulin resistance, and hypokalemia caused by secondary hyperaldosteronism. In addition, hypoalbuminemia is frequently present because of decreased production and sinusoidal leakage of albumin in patients with **portal hypertension**. In patients with chronic cholestasis, impaired delivery of bile salts to the duodenum may result in malabsorption of lipids and fat-soluble vitamins, leading to deficiencies in vitamins A, D, E, and K (see Chapters 23 and 27).

Disordered Hemostasis in Liver Disease

The liver manufactures most of the soluble clotting factors (except for factor VIII and von Willebrand factor) and a number of inhibitors of clotting (proteins C and S, antithrombin III). The liver also clears activated clotting factors from the circulation. Bile acids are necessary for vitamin K absorption and are needed to produce the active forms of several clotting factors, as well as proteins C and S. Disorders of fibrinogen also occur in liver disease.

Enzymes Released from Diseased Liver Tissue

Because hepatic function is often normal in many patients with liver disease, the plasma activities of several cytosolic, mitochondrial, and membrane-associated enzymes are measured because they are increased in many forms of liver disease.

DISEASES OF THE LIVER

The liver has a limited number of ways of responding to injury. Acute injury to the liver may be asymptomatic but often presents as jaundice. The major acute liver disorders are acute hepatitis and cholestasis. Chronic liver injury generally takes the clinical form of chronic hepatitis; its long-term complications include cirrhosis and HCC.

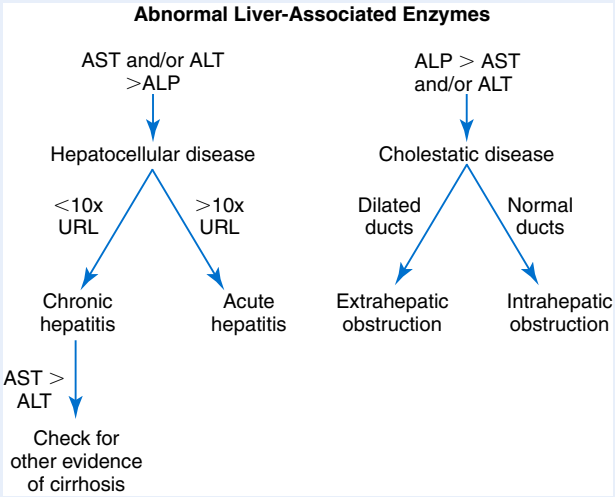
Mechanisms and Patterns of Injury

Cell death occurs by **necrosis** (death of cell), **apoptosis** (programmed cell death), or both. The target cell determines the pattern of injury, with hepatocyte injury leading to hepatocellular disease and biliary cell injury leading to cholestasis. All cellular injury induces fibrosis as an adaptive or healing response, with the duration of injury and genetic factors determining whether cirrhosis and ultimately carcinoma occur (Fig. 37.6).

In general, the aminotransferase enzymes and alkaline phosphatase (ALP) are used to distinguish the pattern (see the following "At a Glance" box), the plasma albumin determines the chronicity, and the PT or factor V concentration determines the severity. Classically, liver fibrosis has been assessed using liver biopsy, but in recent years several noninvasive methods of assessment of liver fibrosis have been described and evaluated. Although liver biopsy yields incomparable information about the nature, severity, and chronicity of liver disease, noninvasive tests may provide more accurate assessment of liver fibrosis, and in particular, disease severity and prognosis.

AT A GLANCE

Algorithm for using abnormal liver function tests to classify and diagnose various types of liver disease.



ALP, Alkaline phosphatase; *ALT*, alanine aminotransferase; *AST*, aspartate aminotransferase; *URL*, upper reference limit.

Initial evaluation is best accomplished by examining the pattern of liver-associated enzymes. If elevation primarily affects one of the aminotransferases, then hepatocellular disease is likely; an increase primarily in *ALP* suggests a cholestatic disorder. If only ALP is elevated, then it is appropriate to consider nonhepatic sources before further investigation (using measurement of other canalicular enzymes such as *GGT* or ALP isoenzymes). If the liver is the source of elevated ALP, then an imaging study to evaluate the ducts is the next test performed; dilated ducts establish a mechanical cause of obstruction, and normal ducts indicate intrahepatic cholestasis requiring further evaluation, as discussed in the text. Predominant increases in aminotransferases suggest hepatocellular injury; values more than 10 times the upper reference limit *(URL)* usually indicate acute hepatitis, and lower values are typical of chronic hepatitis. If aspartate aminotransferase *(AST)* is higher than alanine aminotransferase *(ALT)*, common causes include early hepatic injury, nonhepatic injury (such as muscle injury), and with mildly increased values, cirrhosis.

Disorders of Bilirubin Metabolism

Defects in bilirubin metabolism resulting in jaundice are known to occur at each step in the metabolic pathway (see Chapter 28).

Hepatic Viral Infection

Five viruses (hepatitis A, B, C, D, and E) have been identified as causes of infections that primarily target the liver. In addition, certain other viruses may infect the liver as part of a more generalized infection, of which the more important causes include cytomegalovirus (CMV), Epstein-Barr virus (EBV), and herpes simplex virus (HSV). Several other viruses have been proposed as causes of liver injury; these include hepatitis G virus (HGV), transfusion-transmitted virus, and the closely related SEN virus. Although all three are bloodborne chronic viral infections, and in the case of transfusion-transmitted virus and SEN, have been known to replicate in the liver, none of these viruses appear to cause acute or chronic liver injury. The various hepatitis viruses are outlined in Table 37.1.

Hepatitis A Virus

The incidence of hepatitis A virus (HAV) has fallen over the past four decades due to the introduction of effective vaccination. Hepatitis A is most common in young adult men, particularly in people exposed to sewage, those who eat raw seafood or inject drugs, and in men who have sex with men. It tends to be most virulent in middle-aged and older people. Epidemics have been associated with waterborne and foodborne contamination. It is estimated 50% to 70% of infected adults develop jaundice (more than hepatitis B virus [HBV] or hepatitis C virus [HCV]), and mortality is almost 2% with infection in those older than 60 years of age. HAV infection in children is rarely associated with jaundice and so is usually not detected clinically. Hepatitis A may cause severe hepatitis and death in those who have chronic hepatitis, particularly hepatitis B or C.

HAV is not cytopathic but causes liver injury by stimulating both cellular and humoral immune responses. The incubation period of hepatitis A is 15 to 50 days. The clinical course of acute hepatitis A is usually that of a mild flulike illness that lasts for a few days to a few weeks. There is no chronic form of hepatitis A, but cholestasis (manifested by several weeks of jaundice and pruritus) may occur in some adults. Relapse has been known to happen 1 to 3 months after the acute illness in up to 5% of patients.

Diagnosis of hepatitis A is based primarily on serologic tests for antibodies to HAV. Total anti-HAV is believed to be protective and occurs with natural exposure to HAV and to HAV vaccine. With natural exposure, HAV antibodies appear to persist for life. IgM antibodies to HAV are always present at the time of diagnosis of acute hepatitis A and generally remain present for 3 to 6 months, although they may persist for longer in approximately 14% of individuals. False-positive results can be identified through reference to the clinical context.

Three types of effective vaccines are available. A monovalent vaccine against HAV, a combined HAV and HBV vaccine, and a combined HAV and typhoid vaccine. Vaccination followed by a booster at 12 months will provide immunity for up to 20 years.

Hepatitis B Virus

HBV is the commonest cause of acute hepatitis and the most common chronic viral infection worldwide. It is estimated 350 million individuals are chronically infected with HBV, and several times as many individuals have been exposed to HBV. The prevalence of chronic HBV infection varies worldwide, being highest in Central and Southeast Asia, Central Africa, and Southern Europe (prevalence >8% of the population) and intermediate (2% to 8%) in most of the rest of Asia, Africa, and South America. In endemic areas, the incidence

TABLE 37.1 Types of Viral Hepatitis

	A	B	C	D	E	G
Type	RNA	DNA	RNA	Partial	RNA	RNA
Incubation period, days	45–50	30–150	15–160	30–150	20–40	Unknown
Transmission						
Fecal-oral	Yes	No	Min	No	Yes	No
Household	Yes	Min	Min	Yes	Yes	No
Vertical	No	Yes	Min	Yes	No	Yes
Blood	Rare	Yes	Yes	Yes	Unknown	Yes
Sexual	No	Yes	Min	Yes	Unknown	Yes
Diagnosis	Anti-HAV IgM	HBsAg, PCR, anti-HBc IgM	Anti-HCV, PCR	Anti-HDV	Anti-HEV	Anti-HGV
Carrier state	No	Yes	Yes	Yes	Yes	Yes
Risk of chronic hepatitis	No	Depends on age, immune status	50%–70%	Yes	Rare[a]	No
Risk of liver cancer	No	Yes	Yes	No	No	No
Prevention						
Vaccine	Yes	Yes	No	Yes[b]	No	No
Immunoglobulin	Yes	Yes	No	Yes[b]	No	No
Response to interferon	Not used	30%	40%–80%	Yes	Not used	Yes

[a]Only with severe immunosuppression.
[b]Vaccination and passive immunization against HBV protects against HDV infection.
HAV, Hepatitis A virus; *HBc*, hepatitis B core antigen; *HBsAg*, hepatitis B surface antigen; *HCV*, hepatitis C virus; *HDV*, hepatitis D virus; *HEV*, hepatitis E virus; *HGV*, hepatitis G virus; *IgM*, immunoglobulin M; *Min*, minimal; *PCR*, polymerase chain reaction.

of new infection has decreased markedly where HBV vaccination is practiced. HBV is transmitted through body fluids, primarily by parenteral or sexual contact; it can be transmitted from mother to child, usually at or after delivery (termed vertical transmission). In parts of the world with high rates of chronic infection, much of the transmission is vertical. The residual risk from transfusion is estimated to be 1 in 600,000.

Hepatitis B is caused by a 42-nm DNA virus of the Hepadnavirus family. The S gene codes for several different length variants of surface protein; the smallest form, hepatitis B surface antigen (HBsAg), is produced independently of and in excess of the amount needed for viral replication; the largest form (S1) makes up the surface coat of circulating viral particles. The C gene encodes the hepatitis B core antigen (HBcAg), which is part of the infectious core of the virus. The X gene codes for a transactivating factor that may be involved with viral replication and the development of malignancy. The precore and basal core promoter regions code for production of hepatitis B e antigen (HBeAg), a protein found only in those with circulating viral particles. The final major viral protein is a polymerase, which has several different enzymatically active sites and functions including reverse transcription. This error-prone reproductive strategy and an extremely high replication rate results in a high frequency of mutation in HBV.

Hepatitis B e antigen and hepatitis B e antigen mutants. The most common HBV mutations involve the regions that code for production of HBeAg. These mutants are associated with undetectable HBeAg and are usually found in patients with detectable levels of anti-HBe. They may be present at the time of infection or may develop during the course of disease. Infection with these mutant strains is associated with a higher risk of development of HCC, and risk is stronger for the basal core promoter mutations.

Polymerase mutants. Treatment with antiviral agents that inhibit the reverse transcriptase domain of HBV polymerase is currently the most widely used therapy for chronic HBV.

Hepatitis B surface antigen mutants. Mutations in the "a" determinant of HBsAg are rare but important HBsAg mutants because antibody to HBsAg, which is developed by natural exposure or by the HBV vaccine, is primarily directed against the "a" determinant. Mutations in HBsAg can lead to failure of detection by antibodies used in HBsAg assays. Use of assays with antibodies to multiple epitopes improves detection of such mutant strains.

Diagnostic tests for hepatitis B. The diagnostic evaluation of HBV infection requires the integration of results from biochemical, immunologic, and molecular tests (see the following "At a Glance" box).

HBsAg, the most widely used marker for detecting current hepatitis B infection, is detected by kits using antibody to HBsAg. False-positive results occur rarely, particularly during pregnancy; a neutralization assay is available. False-negative results can occur with mutants in the surface antigen, and they occur more commonly in early HBV infection. Most assays are qualitative. Increasingly quantitative measurement of HBsAg is being used to guide treatment.

Antibody to the HBcAg (anti-HBc) is the most commonly detected antibody against HBV. Two assays are usually used: IgM and total anti-HBc. IgM anti-HBc assays typically use a large dilution of plasma (1:100) before analysis to reduce the likelihood of positivity in individuals with chronic HBV. The total antibody assay measures both IgM and IgG antibodies. Anti-HBc appears to last longer than anti-HBs in natural infection and is still present in 97% of previously infected individuals more than 30 years after exposure. Isolated anti-HBc is a relatively common finding. Although this may represent a false-positive result, particularly as a transient phenomenon after influenza vaccination, current guidelines on hepatitis B recommend consideration of individuals with isolated anti-HBc as having been exposed to HBV.

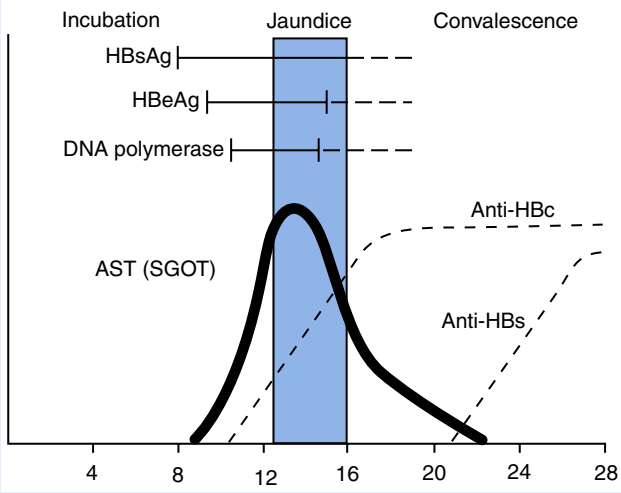
Antibody to the HBsAg (anti-HBs) is considered evidence of immunity to hepatitis B and is the only marker found in those have received the hepatitis B vaccine. The World Health Organization (WHO) has developed reference material that contains 10 IU/mL of anti-HBs. Current guidelines suggest that immunocompetent individuals who achieve an anti-HBs of ≥10 IU/mL have effective immunity to hepatitis B.

The HBeAg and antibody to the e antigen (anti-HBe) are typically used only in the setting of chronic HBV infection. HBeAg typically appears at approximately the same time as HBsAg in acute hepatitis. In chronic infection, HBeAg has historically been used as a marker of persistence of infectious virus; its clearance and the appearance of anti-HBe have been used as indicators of conversion to the nonreplicating state and remain goals of antiviral treatment. With widespread availability of HBV DNA assays with low detection limits, the discordance between HBeAg and the presence of infectious viral particles has become apparent. Although most untreated patients with HBV who are HBeAg positive have high viral loads (usually $>10^6$ IU/mL), detectable HBV DNA is also found (usually with lower viral load) in approximately 70% of those who are HBeAg negative, some of whom may have hepatitis that may be associated with elevated transaminases. When HBeAg-positive individuals are treated with polymerase inhibitors, loss of HBV DNA occurs in the majority. Loss of HBeAg during treatment, with development of anti-HBe, indicates a high likelihood that viral suppression will be maintained after discontinuation of treatment. In contrast, in those with HBV viremia who were HBeAg negative (and anti-HBe positive) before treatment, discontinuation of treatment usually leads to recurrence of viremia except in those with HBsAg levels less than 1000 IU/mL. Thus HBeAg remains an important marker for monitoring therapy but must be assessed in conjunction with HBV DNA for detection of those who harbor the infectious virus.

Hepatitis B viral DNA is now routinely measured using amplification techniques. The WHO has established an international reference material for HBV DNA and results are typically reported in international units per milliliter; conversion from copies per milliliter differs on the basis of viral load and is different for various assays. An approximate conversion factor is 5 copies/mL = 1 IU/mL. Data (primarily from Taiwan) have shown that risk of progression to cirrhosis or HCC increases at viral loads at more than 10,000 copies/mL (2000 IU/mL). Current treatment guidelines suggest that this number should be used as one criterion in treatment decision making.

Hepatitis B mutants and genotypes are usually determined by direct sequencing or with the use of line probes.

AT A GLANCE



Course of acute type B hepatitis with recovery

(1) Onset of hepatitis with jaundice 3 months after exposure; (2) detection of hepatitis B surface antigen *(HBsAg)* 2 to 8 weeks after exposure, followed by appearance of its antibody *(anti-HBs)* 2 to 4 weeks after HBsAg is no longer detectable; (3) detection of hepatitis B e antigen *(HBeAg)* shortly after HBsAg disappears (this is usually followed by the appearance of antibody to HbeAg *[anti-HBe]*, which persists); (4) detection of hepatitis B core antibody *(anti-HBc)* at the time of onset of disease 2 to 3 months after exposure. Anti-HBc immunoglobulin M will be detectable in high levels for approximately 5 months.

AST, Aspartate aminotransferase; *SGOT,* serum glutamic-oxaloacetic transaminase.

From Balistreri, W. F. (1984). Viral hepatitis: unique aspects of infection during childhood. *Consultant, 24,* 131–153.

Hepatitis C virus. The HCV is the cause of most cases previously known as non-A, non-B hepatitis. It was recognized in 1989 and fully characterized 2 years later. It is the commonest cause of chronic **viral hepatitis** in North America, Europe, and Japan and is estimated to infect approximately 170 million individuals worldwide. HCV infection primarily occurs through plasma; major risk factors are injection drug use and transfusion. Because of its mode of spread, HCV infection is rare in children. Vertical transmission may occur from an infected mother (estimated to occur in ≈5% of infected women). Unsterile medical and dental procedures and ritual practices continue to account for new infections.

HCV is a single-stranded enveloped RNA virus of the flavivirus family, which includes other hepatitis viruses (yellow fever virus) and viruses that cause unrelated disease (such

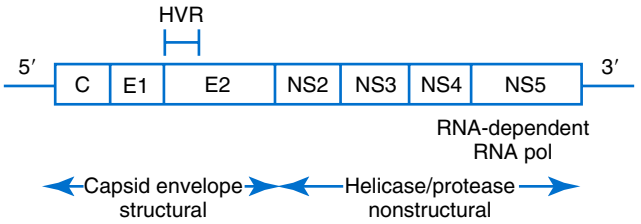


Fig. 37.7 Structure of the hepatitis C genome.

as West Nile virus). HCV RNA contains one reading frame (Fig. 37.7). The resulting polypeptide is cleaved to core and envelope antigens, and a number of nonstructural proteins, including a polymerase and a protease. As an RNA virus, HCV is subject to a high rate of spontaneous mutation, giving rise to large numbers of variants. According to various global reviews, genotype 1 (G1) is the most common (46%; affecting ≈83 million cases, one-third of which are in East Asia), followed by G3 (22% to 31%; ≈54 million), G2 (13%; ≈22 million), and G4 (13%; ≈22 million).

Chronic HCV infection is associated with evidence of chronic liver injury in most cases. Elevations in aminotransferases are usually mild and fluctuate in most infected individuals. In 15% to 20% of cases, cirrhosis becomes evident within 30 years. HCC may develop once cirrhosis is present, at an average rate of 1.5% to 3% of cases per year. In North America, Europe, and Japan, HCV is the most common risk factor in the development of HCC. Various extrahepatic manifestations of chronic HCV infection may be noted; the most common are cryoglobulinemia and porphyria cutanea tarda (see Chapter 29). Epidemiologic evidence has linked HCV to increased risk of lymphoma and type 2 diabetes mellitus.

Diagnostic tests for hepatitis C. Measurement of the antibody to HCV (anti-HCV) is the principal screening test for HCV exposure. Current widely used assays incorporate enzyme-linked immunosorbent assays (ELISAs) using antigens from four different regions of the HCV genome. Tests become positive 9 to 12 weeks after exposure.

Reverse immunoblot assays (RIBAs) use similar principle to Western blotting. HCV antigens are blotted onto a membrane and reactivity is detected after incubation with serum. Results are interpreted as negative if there is less than 1+ reactivity with any of the four antigens, indeterminate if there is 1+ or greater reactivity to only a single antigen (or to more than one antigen along with the nonspecific yeast marker superoxide dismutase), and positive with 1+ or greater reactivity to multiple antigens.

HCV RNA measurement, typically performed using polymerase chain reaction (PCR) amplification, has become the most widely used test to detect current HCV infection. Rapid separation of serum from a clot is critical for accurate measurement of HCV RNA. Separation of serum from cells by centrifugation and refrigeration preserve the integrity of HCV RNA, and samples can be kept indefinitely if frozen. Samples collected in ethylenediaminetetraacetic acid are stable for 24 hours, even if plasma is not separated from red cells.

HCV RNA assays have been standardized against international reference material for quantification of HCV RNA in international units per milliliter. Until recently, qualitative assays had significantly lower detection limits than quantitative assays, but quantitative assays using real-time PCR have equivalent or lower detection limits compared with qualitative assays.

Hepatitis C core antigen (HCV Ag) is produced by the most constant part of the HCV genome and has a similar time course to that of HCV RNA in both acute and chronic HCV infection; the currently available assay for HCV Ag becomes reliably positive when HCV RNA is 20,000 IU/mL or greater which is observed in 95% of untreated HCV RNA–positive individuals. In contrast to HCV RNA, HCV Ag is stable in storage. Currently, no commercial HCV Ag assays are available in North America.

Hepatitis C genotype shows regional diversity and is currently important for determining the selection of antiviral therapy, although pan-genotypic treatments are emerging. Several methods are currently used to determine the infecting genotype. Serologic assays are available but their correlation with direct tests is approximately 90%, and a significant minority of infected individuals have antibodies to more than one genotype. The most reliable method involves direct sequencing of regions of the genome that show characteristic patterns with specific genotypes and subtypes. Commercial assays using the 5′-untranslated region are the most widely used, although assays using the NS5b region are now available. Available assays show good agreement on genotype, but they differ in their detection limits. Sequencing methods have the advantage that they can be used to identify treatment resistance–associated variants that may develop under selection pressure from the new directly acting antiviral agents. Line probe assays are also widely used and show good agreement with direct sequencing assays.

Hepatitis D Virus (Delta Agent)

Hepatitis D virus (HDV) is an incomplete, 36-nm single-stranded antisense RNA virus that cannot replicate on its own. It is coated with HBsAg and is dependent on HBV for its activation. It is thus a satellite virus similar to that seen in plants. It is very infectious and strongly associated with intravenous drug use; approximately 10 million individuals have been infected worldwide, although the incidence is declining with the fall in incidence of HBV infection. It occurs as simultaneous infection with hepatitis B (coinfection) or as a superimposed infection in someone with chronic hepatitis B (superinfection). Coinfection usually runs the same time course as acute hepatitis B, and HDV is spontaneously cleared as the hepatitis B resolves, but the risk of **fulminant hepatitis** is higher than in HBV infection alone, and mortality is higher. Superinfection typically results in chronic HDV infection, suppression of HBV DNA replication, and more rapid progression to cirrhosis (estimated 4%/year) and HCC (estimated 3%/year). It should be assessed in all patients with HBV infection, due to its seriousness, and suspected in patients with HBV infection whose condition worsens. Although it is traditionally diagnosed serologically by detection of anti-HDV (total or IgM) and/or HDV Ag, HDV RNA measurements are often used as evidence of current infection.

TABLE 37.2 Laboratory Features of Different Forms of Acute Hepatitis

Type	AST/ALT	ALP	Bilirubin	PT (s)	Serology	Other
Viral	8–50× URL	<3× URL	5–15 mg/dL (86–256 μmol/L)	<15	Positive	
HAV					IgM anti-HAV	
HBV					HBsAg, IgM anti-HBc	
HCV					HCV RNA ± anti-HCV	
Alcoholic	<8× URL	>3× URL in 25%	5–15 mg/dL (85–256 μmol/L)	<15	Negative	AST > ALT
Toxic	>50× URL	Normal	<5 mg/dL <85 μmol/L)	>15	Negative	Toxin usually detectable; acute renal failure common
Ischemic	>50× URL	Normal	<5 mg/dL (<85 μmol/L)	>15	Negative	Acute renal failure common
Drug induced	8–50× URL	>3× URL in 50%	5–15 mg/dL (85–256 μmol/L)	<15	Negative	Eosinophilia, skin rash common
Autoimmune	8–50× URL	<3× URL	5–15 mg/dL (85–256 μmol/L)	<15	Positive ANA or ASMA	Low albumin, high globulins
Wilson	8–50× URL	Low normal or decreased	5–15 mg/dL (85–256 μmol/L)	<15	Negative	Hemolytic anemia, renal failure, low ALP common; low ceruloplasmin often absent

ALP, Alkaline phosphatase; *ALT*, alanine aminotransferase; *ANA*, antinuclear antibody; *ASMA*, anti–smooth muscle (or antiactin) antibody; *AST*, aspartate aminotransferase; *HAV*, hepatitis A virus; *HBc*, hepatitis B core antigen; *HBsAg*, hepatitis B surface antigen; *HBV*, hepatitis B virus; *HCV*, hepatitis C virus; *IgM*, immunoglobulin M; *PT*, Prothrombin time; *URL*, upper reference limit.

Hepatitis E Virus

Hepatitis E virus (HEV) is a 34-nm, single-stranded, unenveloped RNA virus. It accounts for sporadic and epidemic hepatitis in tropical and semitropical countries and in people returning from these areas. HEV RNA is frequently isolated from city sewage treatment plants in nonendemic areas. It is enterically transmitted and viral RNA has been detected in plasma and in stools. The specificity of antibody tests for open reading frame 2 antigen of HEV is high. The prevalence of antibodies to HEV is in the region of 21% in the United States. HEV has been isolated from a number of animals, notably rats and pigs, and HEV has been linked to ingestion of pork.

Detection of IgM anti-HEV antibodies is considered diagnostic of acute infection by HEV, but false-positive results have been reported with hepatitis due to CMV and EBV. Molecular tests for HEV RNA can be used to diagnose infection and monitor clearance. The clinical course is similar to that of HAV infection. HEV typically has a self-limited course, but increasingly cases of chronic infection are reported, particularly in organ transplantation recipients. A peculiar feature of this disease is its virulent course in late pregnancy in India, with mortality generally in the range of 20% to 25%, but rates as high as 50% have been reported. Mortality during pregnancy is not increased in other parts of the world. Mortality is increased among the elderly and in those with chronic liver disease.

Hepatitis G Virus

HGV, also known as GBV-C, is an RNA virus of the Flavivirus family and is closely related to HCV. It is most commonly transmitted by plasma but vertical transmission (i.e., from mother to offspring) has also been reported. HGV infection appears to have no adverse consequences, and, although it has been called a hepatitis virus, viral RNA cannot be isolated from the liver in chronic infection.

Acute Hepatitis

Acute hepatitis refers to an acute injury directed against the hepatocytes. The injury may be mediated directly, which occurs with certain drugs, such as acetaminophen or with ischemia, or indirectly, which occurs with immunologically mediated injury from most of the hepatitis viruses and most drugs, including ethanol. Direct injury typically causes a rapid rise in cytosolic enzymes, such as AST, ALT, and lactate dehydrogenase (LDH), followed by a rapid fall, with rates of decline similar to known half-lives of the enzymes. Immunologic injury typically causes a gradual rise in cytosolic enzymes, followed by a plateau phase and gradual resolution of enzyme elevation. Although jaundice is a key clinical finding in acute hepatitis, it is often absent. An increase in AST activity to greater than 200 U/L or in ALT activity to greater than 300 U/L has sensitivity and specificity greater than 90% for acute hepatitis.

ALP usually is mildly elevated and typically is less than 3 times the URL in 90% of cases of acute hepatitis. Increased plasma concentration of bilirubin, when present, typically is predominantly due to direct reacting bilirubin. The distribution of direct bilirubin percentage is identical in acute hepatitis and bile duct obstruction. The features helpful in the differential diagnosis of acute hepatitis are summarized in Table 37.2.

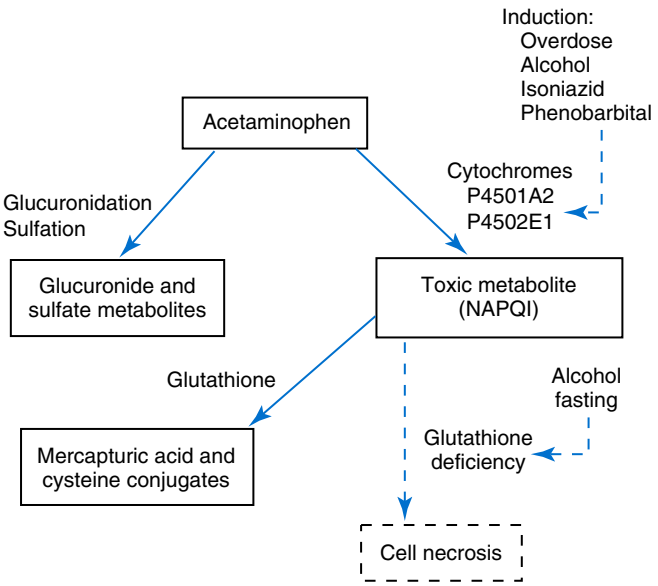


Fig. 37.8 Metabolism of acetaminophen by the liver.

The outcome of acute hepatitis is variable. In most cases, complete recovery occurs, and liver regeneration leads to normal structure and function. In a small percentage of cases, massive destruction of the liver leads to acute (fulminant) hepatic failure, which is associated with high mortality unless liver transplantation is performed.

Acute Viral Hepatitis

All forms of acute viral hepatitis have similar pathology and a similar clinical course. They are all diagnosed on the basis of marked elevations in serum aminotransferase activities, usually to between 8 and 50 times the upper reference intervals, with only slight elevations in ALP and little or no effect on hepatic synthetic function. ALT is typically higher than AST because of slower clearance. Enzyme elevations typically peak before peak bilirubin occurs and remain increased for an average of 4 to 5 weeks (longer for ALT than AST because of its longer half-life). Bilirubin elevation is variable. The incidence, prevalence, and management of acute viral hepatitis are covered in the practice guideline of the World Gastroenterology Organisation.

Toxic Hepatitis

Toxic hepatitis refers to direct damage of hepatocytes by a toxin or toxic metabolite. Toxic reactions are usually predictable and are directly related to the dose of the agent ingested. In North America and Europe, the most common cause of toxic hepatitis (and the most common cause of acute liver failure) is acetaminophen. The metabolism of acetaminophen is affected by dose, induction of metabolic enzymes, and concentrations of glutathione (Fig. 37.8). When a large dose of acetaminophen is ingested (the average lethal dose as a single ingestion is 15 g), the metabolic pathways are overwhelmed, glutathione is depleted, and toxic intermediates accumulate, causing liver damage. In glutathione depletion due to starvation or metabolic enzyme induction (such as by ethanol), toxicity can occur with relatively small doses of acetaminophen (total doses of 2 to 4 g). Toxicity can also occur with excessive cumulative doses of acetaminophen. Diagnosis is often based on history and increased acetaminophen concentrations. In late presentation or when a history cannot be obtained, measurement of acetaminophen-protein adducts allows diagnosis. The first laboratory abnormality to appear is an increase in PT, followed by increased activity of cytosolic enzymes, with AST tending to be higher than ALT. Peak activities (typically >100 times the URLs) usually occur by 24 to 48 hours, followed by rapid clearance at rates approximating the known half-lives of the enzymes. PT elevations are typical and are more than 4 seconds greater than the control value in most cases. Prognosis is related most closely to the prolongation of PT; persistent elevation of PT 4 days after ingestion is associated with a poor prognosis. Other markers of risk include development of acute renal failure and the presence of lactic acidosis, particularly if the pH is less than 7.30 ($[H^+]$ >50 nmol/L).

Ischemic Hepatitis (Shock Liver)

Hepatic hypoperfusion (ischemic hepatitis) is one of the most common causes of elevated cytosolic enzymes, which may exceed 10,000 IU/L; in hospital patients, it is the cause of most cases of acute hepatitis. Ischemic hepatitis may follow any cause of shock; the most common causes are septic and cardiogenic shock. Bilirubin elevations typically are minimal, and they usually peak several days after enzyme activity reaches its greatest point. Laboratory findings are similar to those seen in toxic hepatitis, and acute renal failure is a common complicating factor. Prognosis is primarily related to the underlying cause of hypotension. Individuals with prolonged elevation of bilirubin appear to have a poor prognosis.

AT A GLANCE

Assessment of Viral Hepatitis

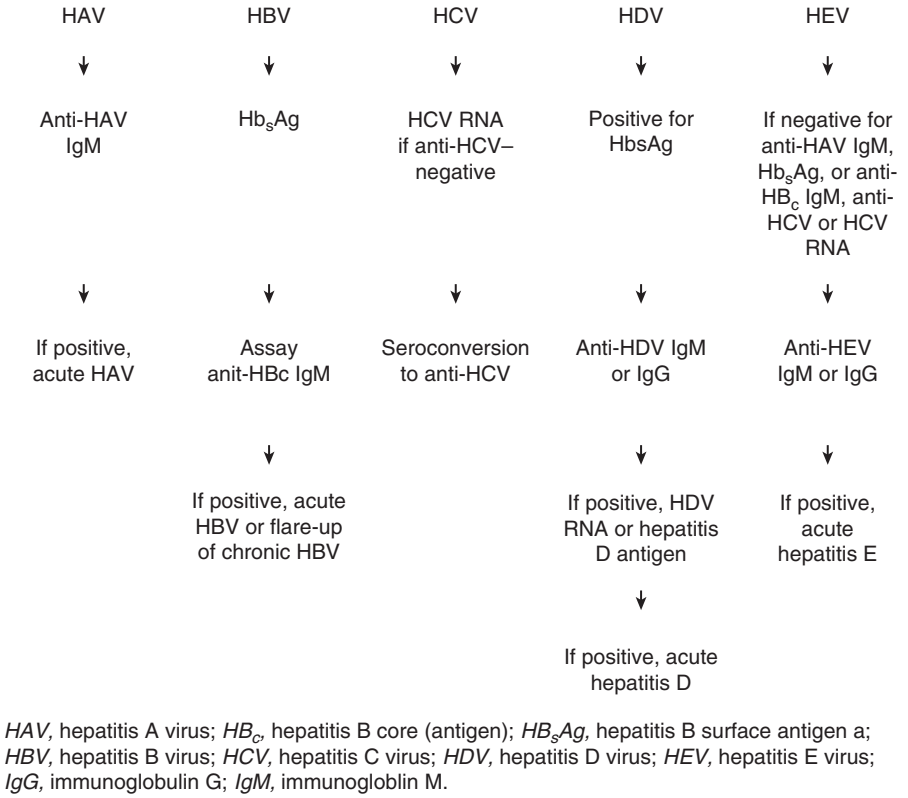
- In all cases of viral hepatitis, it is prudent to test patients for each of the major bloodborne viruses: HBV, HCV, and human immunodeficiency virus.
- In viremic patients, testing should include measurement of viral nucleic acid (HBV DNA and HCV RNA).
- Hepatic inflammation is indicated by the extent of elevation of aminotransferases, but these are neither sensitive nor specific.
- The staging of severity of chronic viral hepatitis is determined by the severity of liver fibrosis. Liver fibrosis can be assessed using noninvasive blood tests, fibroelastography, imaging or liver biopsy, or any combination of these tests.

Chronic Hepatitis

Chronic hepatitis is defined as chronic inflammation of the liver that persists for at least 6 months or signs and symptoms of chronic liver disease in the presence of elevated cytosolic enzymes. Common causes of chronic hepatitis and tests used to make a specific etiologic diagnosis are listed in Table 37.3.

The clinical features of chronic hepatitis are highly variable. Most patients are asymptomatic, but nonspecific features, such

DIAGNOSTIC ALGORITHM OF VIRAL HEPATITIS



HAV, hepatitis A virus; *HB_c,* hepatitis B core (antigen); *HB_sAg,* hepatitis B surface antigen a; *HBV,* hepatitis B virus; *HCV,* hepatitis C virus; *HDV,* hepatitis D virus; *HEV,* hepatitis E virus; *IgG,* immunoglobulin G; *IgM,* immunogloblin M.

From the World Gastroenterology Organisation Practice Guidelines–Management of Acute Viral Hepatitis. (2003). http://www.worldgastroenterology.org/guidelines/global-guidelines/management-of-acute-viral-hepatitis/acute-viral-hepatitis-english.

TABLE 37.3 Causes of Chronic Hepatitis and Diagnostic Strategies

Cause	Diagnosis
Hepatitis B	History, HBsAg, anti-HBs, anti-HBc, HBV DNA
Hepatitis C	Anti-HCV, HCV RNA by PCR
Autoimmune type 1	ANA, anti–smooth muscle antibody
Autoimmune type 2	SLA, anti-LKM1
Wilson disease	Ceruloplasmin
Drugs	History
AAT deficiency	AAT phenotype
Nonalcoholic fatty liver disease	Metabolic syndrome, liver ultrasound, liver biopsy
Idiopathic	Liver biopsy, absence of markers

AAT, Alpha$_1$-antitrypsin; *ANA,* antinuclear antibody; *HBc,* hepatitis B core antigen; *HBsAg,* hepatitis B surface antigen; *HBV,* hepatitis B virus; *HCV,* hepatitis C virus; *LKM1,* liver-kidney microsomal antigen type 1; *PCR,* polymerase chain reaction; *SLA,* soluble liver antigen.

as fatigue, lack of concentration, and weakness may be present. Most patients are diagnosed because of an unexplained abnormality in aminotransferase activities or detection of positive results on a screening test for a cause of chronic hepatitis. Moderate elevations in plasma aminotransferase activities are characteristic (an average of approximately twofold, and in most cases, less than fivefold). Normal aminotransferase activities do not exclude chronic hepatitis, especially in the presence of chronic viral hepatitis or **nonalcoholic fatty liver disease (NAFLD)**. Characteristically, ALT is elevated to a greater degree than AST. Reversal of the AST/ALT activity ratio to more than 1 suggests coexisting alcohol abuse or cirrhosis. Although ALT is relatively specific for the liver, skeletal muscle sources for AST and ALT should always be considered, especially in physically active young individuals. Persistent elevation of aminotransferase activity should lead to an evaluation for chronic hepatitis using the tests outlined in Table 37.3. A liver biopsy may be helpful in determining the cause and in assessing disease severity. A specific etiologic diagnosis is essential to determine prognosis and treatment. The most common causes of chronic hepatitis are chronic HBV, HCV, and NAFLD, but a variety of other disease processes may cause chronic hepatitis.

Chronic Hepatitis B

Worldwide, HBV infection is the most important cause of chronic hepatitis. According to the WHO, approximately 350 million individuals worldwide have chronic HBV infection; most cases are found in Asia, Africa, and Southern Europe.

HBV is not cytopathic, rather injury results from an immune-mediated inflammatory attack against hepatocytes. Chronic hepatitis results when the virus is not eliminated

TABLE 37.4 Patterns of Chronic Hepatitis B Virus Infection

Type	AST/ALT	HBsAg	HBeAg	Anti-HBc	HBV DNA
Occult	Normal	Negative	Negative	Positive	Negative[a]
Immune control	Normal	Positive	Negative	Positive	Negative[a]
Immune tolerant	Normal	Positive	Positive/negative	Positive	Positive
Immune active	Increased	Positive	Positive/Negative	Positive	Positive, viral load >2000 IU/mL
HBeAg-negative chronic hepatitis	Increased	Positive	Negative	Positive	Positive, viral load usually >2000 IU/mL but $<10^6$ IU/mL

[a]May have very low level (usually $<10^2$ IU/mL) in serum.
ALT, Alanine aminotransferase; *AST*, aspartate aminotransferase; *HBc*, hepatitis B core antigen; *HBeAg*, hepatitis B e antigen; *HBsAg*, hepatitis B surface antigen; *HBV*, hepatitis B virus.

from infected cells resulting in a continuing cycle of viral replication, reinfection of regenerating hepatocytes, and immune damage to newly infected cells that is inadequate to clear infection.

The clinical presentation may be complicated by various extrahepatic complications (which occur in 1% to 10% of those with HBV), including polyarteritis, glomerulonephritis, polymyalgia rheumatica, cryoglobulinemia, myocarditis, and Guillain-Barré syndrome. These conditions are associated with circulating immune complexes containing HBsAg. Immunocompromised persons, such as HBV/HIV coinfected individuals, typically have higher replication markers, less hepatic inflammation, and poorer survival than those with HIV alone. The natural history of chronic hepatitis B (defined by the persistence of HBsAg) varies. Features of the different stages of chronic HBV are given in Table 37.4. The evolution of HBV infection is best defined by the nature of the host immune response and inflammatory activity. The stages of chronic HBV infection can only be classified with reference to eAg and eAb status, HBV DNA concentrations, along with ALT activities and histology. HBV DNA levels are commonly very high (exceeding 10^6 IU/mL) during the immune-tolerant phase of infection. In the immune active and immune control phases of HBV infection, viral DNA levels exceeding 2000 IU/mL are considered to be indicative of levels of replication associated with liver damage. All phases of chronic HBV are associated with the presence of HBsAg in serum. Occult HBV infection is not associated with inflammation in the liver or elevated aminotransferases.

In general, chronic HBV passes through stages in untreated persons. For those infected early in life, the immune tolerant phase is usually followed by an immune active phase that may vary in duration and may cause significant liver damage. This phase usually gives way to a protracted period of immune control, with HBsAg positivity or occult infection if HBsAg is cleared. For those infected individuals who go on to have chronic infection as older children or adults, the immune tolerant phase may be a short period or absent, leading straight into the immune active phase. Approximately 8% to 10% of persons per year will transition from the immune active to the immune control phase of chronic HBV. A variable, but low, proportion of those in the immune control phase will revert to the immune active phase that is manifest by levels of HBV DNA greater than 2000 IU/mL and elevated transaminases. Approximately 0.5% to 1% of persons per year will convert from HBsAg positive to HBsAg negative, entering the occult phase of infection (or, in a minority, resolved HBV). Each of these transitions can be associated with an acute rise in aminotransferase activities and a clinical picture that mimics acute viral hepatitis.

Individuals who have chronic replicating infection are at risk of developing cirrhosis and HCC. An estimated 20% to 30% of individuals with chronic hepatitis B will develop cirrhosis over a 20-year follow-up period; the risk is directly related to the amount of HBV DNA, with risk progressively increasing at viral loads of more than 2000 IU/mL (10,000 copies/mL). Once cirrhosis has developed, the annual risk of HCC is 1.5% to 5%. Although the risk of HCC is lower in individuals with HBV infection who do not have cirrhosis, risk is directly related to viral load and rises at quantities greater than 2000 IU/mL. Even a person in the nonreplicating stage of infection have a 10-fold higher risk of HCC than unexposed individuals. Worldwide, hepatitis B infection is the most common cause of liver cancer.

Efficacy of treatment is typically measured by response of ALT and/or AST and HBV DNA; goals of treatment include normalization of ALT and suppression of HBV DNA to less than the limits of detection of assays, ideally using assays with detection limits of approximately 20 to 50 IU/mL. With polymerase inhibitors, approximately 70% to 80% of patients will achieve these goals within 1 year of treatment. Duration of treatment is largely dependent on HBeAg status before the start of therapy. For those who are HBeAg positive before treatment, loss of HBeAg and development of anti-HBe indicate a high likelihood of maintenance of HBV DNA control and normalization of ALT once treatment is stopped (after 6 to 12 months of further treatment once anti-HBe appears). For those who are HBeAg negative prior to treatment, suppression of HBV DNA for more than 3 years with HBsAg levels less than 1000 IU/mL is associated with good control after stopping treatment. Treatment should be maintained indefinitely in patients who do not achieve these goals. With interferon therapy, response rates to 1 year of treatment are somewhat lower than with polymerase inhibitors, but treatment requires only 6 to 12 months. Loss of HBsAg is uncommon; with most agents, the likelihood of loss of HBsAg usually is not higher than that in untreated persons, although it may be higher with interferon (IFN) and tenofovir.

Chronic Hepatitis C

Approximately 170 million individuals worldwide have been diagnosed with chronic HCV infection; most cases are found in North America, Northern Europe, and Japan. Chronic HCV infection develops in 75% to 85% of cases of acute hepatitis C. Clearance of acute HCV infection does not appear to confer lasting immunity. Once viremia becomes established beyond 6 months after initial exposure, infection never resolves spontaneously. It is estimated that approximately 20% to 30% of patients with hepatitis C will progress to cirrhosis over a period of 20 years. The frequency of progression appears to be increased by age older than 40 years at the time of infection, male sex, alcohol abuse, and immunosuppression, but it is less than 5% after 20 years of infection in those infected during the first 20 to 30 years of life. In those who develop recurrent HCV after liver transplantation, the rate of progression to cirrhosis is faster than in primary infection. The likelihood of progression to HCC is between 1.5% and 5% per year in those with cirrhosis.

Infection with HCV is characterized by fluctuating ALT activities over time. Only approximately one-third of those with chronic HCV have continually increased ALT, and many of these individuals show variation in ALT activity. A significant minority of individuals with chronic hepatitis C and normal transaminases have advanced fibrosis or cirrhosis.

Dramatic advances in the treatment of chronic hepatitis C have revolutionized outcomes for patients and changed the requirements for laboratory testing in the management of patients on treatment. Directly acting antiviral agents that target essential HCV proteins have been shown to cure infection in most patients. The critical tests required for the evaluation and monitoring of HCV infection are HCV genotype and viral load measurement, both of which depend on molecular tests. Although this remains a rapidly evolving field, it is highly likely that it will be possible to cure infection in almost all cases within the next decade in any patient capable of adhering to a full course of treatment. The assessment of HCV genotype, the presence of resistance associated sequences, and viral load may influence the choice of treatment and so access to molecular tests of HCV is mandatory in the treatment of hepatitis C. The extent to which liver fibrosis reverses in patients who attain clearance of HCV still needs to be established, but studies of patients who have been cured using INF-based therapies suggest that the long-term risks of liver cancer and bleeding **varices** are reduced even in cirrhotic patients who attain clearance of HCV.

Nonalcoholic Fatty Liver Disease and Nonalcoholic Steatohepatitis

NAFLD is now a major cause of chronic liver disease and is increasing in prevalence. NAFLD encompasses a range of pathology from simple steatosis to **nonalcoholic steatohepatitis (NASH)** characterized by an inflammatory response to hepatic fat. Fibrosis is a feature of NAFLD that is triggered by NASH and may culminate in cirrhosis. The histologic features of NASH are identical to those of alcoholic hepatitis (including hepatocyte ballooning, presence of Mallory hyaline, and neutrophil infiltration) but can be attributed to fatty liver disease in those who have no history of heavy alcohol intake. Many individuals with NAFLD have features of the metabolic syndrome (obesity, hypertension, and dyslipidemia). As many as 20%–30% of the population in North America and Europe have NAFLD, making it far and away the most common form of liver disease and an extremely common condition in the developed world. Approximately 10% of those with NAFLD have NASH. Almost half of individuals who meet the criteria for metabolic syndrome have NAFLD. The frequency in obese or diabetic individuals is much higher, with NAFLD in 60% to 75% and NASH in 20% to 25%. Prospective studies of patients with liver disease have confirmed that NASH is a common cause of elevated liver enzymes in an unselected population of patients referred to gastroenterologists or seen in primary care settings. The frequency of cirrhosis in NASH is not well established, but it has been suggested that NASH may be a major cause of cryptogenic cirrhosis. Because weight loss develops with chronic illness, fat may disappear from the liver, leaving only fibrosis.

Current evidence suggests that accumulation of fat in NAFLD is a consequence of insulin resistance. However, the pathogenesis is likely to be more complicated because a variety of other factors lead to increased fat accumulation in the liver, including increased carbohydrate intake, certain drugs, and mutations in lipid synthesis, but they have not been associated with the development of NASH. Laboratory abnormalities may include raised ferritin levels, signs of diabetes, and dyslipidemia.

Autoimmune Hepatitis

AIH represents a rapidly progressive form of chronic hepatitis, with up to 40% 6-month mortality in untreated individuals; it is associated with the presence of autoimmune markers. It is relatively uncommon, with an annual incidence of 1.9 cases per 100,000 population in the United States, but it is responsible for 3% to 6% of all liver transplantations; the disease recurs in approximately 30% of patients after transplantation. As with most autoimmune diseases, there is a strong female predominance. Forms of AIH have been found in individuals of all ages, with no racial or ethnic predilection. It has been associated with specific human leukocyte antigen (HLA) haplotypes, notably DR3 and DR4, as is true for many other autoimmune diseases. For details on autoimmune chronic hepatitis, refer to the AASLD (American Association for the Study of Liver Diseases) practice guidelines.

AIH is associated with the presence of liver and nonliver autoantibodies in plasma. These are helpful in diagnosis but are not likely to be the cause of liver injury. The most important antibodies for diagnosis include antinuclear antibody (ANA), anti–smooth muscle (or anti-actin) antibody (ASMA), anti-liver–kidney microsomal antigen type 1, and anti-soluble liver antigen (SLA), which is insensitive but highly specific. A summary of the most common autoantibodies, their associations, and their molecular targets (when known) are given in Table 37.5. Individuals who are negative

TABLE 37.5 Serologic Markers of Autoimmune Liver Disease

Antibody Name	Antigen Target	Associations
Antiactin	Actin	AIH type 1; more specific than ASMA, poor response to corticosteroids, early age of onset
Anti-asialoglycoprotein receptor	Transmembrane antigen-binding protein	AIH, correlate with activity, disappear with successful treatment
Anti–LKM1	Cytochrome P450 2D6	AIH type 2; seen in only 4% of US cases; usually in children
Anti–liver-specific cytosol	Enzyme (possibly formiminotransferase cyclodeaminase or argininosuccinate lyase)	AIH in younger patients, often with anti-LKM1, primary sclerosing cholangitis; vary with activity of disease
Antimitochondrial antibody	Dihydrolipoamide acyltransferase	Primary biliary cholangitis
Antineutrophil cytoplasmic antibodies	Bactericidal/permeability protein, cathepsin G, lactoferrin	PSC (50%–70%), ulcerative colitis (50%–70%), AIH; nonspecific
Antinuclear antibody	Multiple targets (centromere, ribonucleoproteins); may not be detected by ELISA	AIH type 1, some PSC cases
ASMA	Actin, tubulin, vimentin, desmin, keratin	AIH type 1, seen in other autoimmune diseases in lower titers
Antisoluble liver antigen/liver pancreas	UGA tRNA suppressor–associated transfer protein	AIH type 3; specific for AIH, correlate with relapse after corticosteroid withdrawal

AIH, Autoimmune hepatitis; *ASMA*, anti–smooth muscle antigen; *ELISA*, enzyme-linked immunosorbent assay; *LKM1*, liver-kidney microsome; *PSC*, primary sclerosing cholangitis.

for common autoantibodies, but who otherwise meet criteria for diagnosis, have a similar prognosis and response to treatment.

Inherited Liver Disease Presenting as Chronic Hepatitis

Inherited liver diseases that present as chronic hepatitis are relatively uncommon but important to recognize and include hemochromatosis, Wilson disease, and AAT deficiency.

Hemochromatosis. Hereditary hemochromatosis (HH) is an autosomal recessive disorder of iron metabolism that results in excessive iron absorption and accumulation in tissues, specifically in the parenchymal cells of the liver, heart, pancreas, and other organs HH is caused by mutations in the *HFE* gene that affect any of the proteins that control the entry of iron into the circulation. These proteins include **hepcidin**, HFE, transferrin receptor 2 (Tfr2), hemojuvelin (HJV) (these proteins all sense iron accumulation that hepcidin acts to correct) and the protein ferroportin (Fpn), which is a cellular transporter of iron down-regulated by hepcidin (for more detail refer to Chapter 28).

HH is primarily a disease in men and postmenopausal women with common clinical features, including lethargy and weakness, arthralgia, loss of libido, upper abdominal discomfort, hepatomegaly, grey/bronze skin pigmentation, testicular atrophy, and joint swelling and/or tenderness. Untreated, HH can lead to serious complications, including liver fibrosis and cirrhosis with HCC in approximately 30% of patients with cirrhosis, diabetes mellitus, and cardiomyopathy, and arrhythmias.

Liver disease due to hemochromatosis becomes more common after the age of 30 years. Liver function tests are frequently normal in asymptomatic patients. The most useful tests are fasting transferrin saturation and serum ferritin. A transferrin saturation value more than 45% is the most sensitive marker of early iron overload, but neither a raised fasting transferrin saturation nor ferritin concentration is diagnostic of HH. Ferritin is also an acute-phase reactant and can be raised nonspecifically in the presence of alcohol consumption, NAFLD, other liver disease and any systemic inflammatory condition. Serum ferritin is considered to be abnormal when it is more than 250 μg/L in premenopausal women and more than 300 μg/L in men and postmenopausal women. *HFE* gene testing can be used to confirm that diagnosis of HH and to evaluate familial risks. Cirrhosis is unlikely if the ferritin is less than 1000 μg/L, plasma AST activities are normal, and there is no hepatomegaly. Cirrhotic patients rarely regress to normal and have a lifelong risk of HCC.

Wilson disease. **Wilson disease** is an autosomal recessive disorder of copper metabolism. It has a gene frequency of 1 in 200 and a disease frequency of 1 in 30,000. It is due to mutations in a gene on chromosome 13 that codes for a copper-transporting adenosine triphosphatase (ATP7B) found mainly in the liver and involved in movement of copper into bile; deficiency leads to accumulation of copper in the liver and eventually in other tissues.

Wilson disease usually manifests at age younger than 30 years. Patients usually present either with the hepatic or the neuropsychiatric form of the disease. Hepatic involvement tends to predominate in children, whereas in adolescents and adults, the neuropsychiatric form becomes more common. Patients presenting with neuropsychiatric manifestations commonly have advanced liver disease at presentation, whereas those presenting with liver disease may have little in the way of neurologic damage. An important presentation to recognize is antibody negative hemolytic anemia resulting from hepatic release of copper into the circulation. Several laboratory tests are available for the diagnosis of Wilson disease; ceruloplasmin (see Chapter 18) and copper

measurement (see Chapter 27). Classic findings of Wilson disease include decreased plasma ceruloplasmin, decreased total plasma copper, increased plasma-free (or non-ceruloplasmin) copper, increased urine copper excretion, and increased hepatic copper content. Ceruloplasmin is a ferroxidase that typically is measured by enzymatic activity or by immunoassay. Controversy is ongoing over which assay format is preferable, and guidelines have not specified one type. Molecular diagnosis is complicated by the numerous gene mutations known to cause Wilson disease.

Alpha$_1$-antitrypsin deficiency. AAT is the most important of the serine protease inhibitors (collectively termed serpins; see also Chapter 18). AAT inhibits trypsin and other proteolytic enzymes, including neutrophil-derived elastase, cathepsin G, and proteinase 3. The gene for AAT *(SERPINA1)* is located on chromosome 14. Several genetic variants of AAT (differing by a single amino acid) have been classified on the basis of their electrophoretic mobility; the slowest migrating of these was termed the Z variant. The most severe forms of disease have been associated with homozygosity for the Z variant, which is found in 1 in 1000 to 2000 individuals in Europe and North America. However, it is estimated that only approximately 10% of those with AAT deficiency develop clinical disease.

The effects of AAT deficiency on the liver are controversial. In neonates, AAT deficiency is often associated with hepatitis, with up to 25% 1-year mortality. However, in those who survive the first year, evidence of liver injury diminishes and usually resolves by age 12 years.

AAT is estimated by protein electrophoresis, in which it constitutes most of the α_1-globulin band. It can also be quantified by a variety of other techniques. Determination of phenotype was typically accomplished by isoelectric focusing and had been recommended as the diagnostic test of choice in one guideline, but phenotyping cannot distinguish true homozygotes from heterozygotes who have a null genotype on the other *SERPINA1* gene. Molecular tests are now available to determine *SERPINA1* genotype. Because the prognosis may vary between those who are actually homozygous for the Z variant and those with null phenotype, molecular testing of the *SERPINA1* gene is considered preferable.

Drug-Induced Liver Injury

As discussed earlier, most cases of **drug-induced liver injury (DILI)** present as acute hepatitis. Less commonly, drugs have produced chronic liver injury in a pattern that mimics chronic hepatitis or other chronic liver injury (chronic cholestasis and hepatic granulomas). The drugs most commonly linked to chronic hepatitis are nitrofurantoin, methyldopa, and hydroxy-3-methylglutaryl-coenzyme-A (CoA) reductase inhibitors; however, a large number of drugs have been associated with liver injury, and herbal medications have been linked to chronic hepatitis. In individuals with increased activities of aminotransferases and no obvious cause, prescription drug use was significantly more likely to be present than in those with a known cause for elevated enzyme activities. As with acute drug reactions, establishing drugs as the cause of chronic hepatitis is difficult; temporal relationships to drug ingestion are not as clear as with acute hepatitis, and reactions can be seen first in those who have been taking the medication for many months. Most chronic drug reactions resolve when administration of the drug is discontinued.

POINTS TO REMEMBER

Most Common Causes of Chronic Liver Disease
Alcohol
Fatty liver disease
Chronic viral hepatitis B or hepatitis C

AT A GLANCE

Causes of Chronic Liver Disease
Often the cause of chronic liver disease is not immediately obvious. In this situation, a systematic approach to differential diagnosis should be taken, gathering history, and the findings of clinical examination, and blood and urine tests, as well as imaging and biopsy.

The following etiologies of chronic liver disease should be considered:
Toxins: alcohol, drugs
Viruses: HBV, HCV
Metabolic liver diseases: fatty liver disease, hemochromatosis, AAT deficiency, Wilson disease
Immune-mediated liver diseases: AIH, PBC, primary sclerosing cholangitis
Infiltration of the liver
Tumors: benign and malignant (primary, secondary)

Cirrhosis

Cirrhosis, which is defined anatomically as diffuse fibrosis with nodular regeneration, represents the end stage of scar formation and regeneration in chronic liver injury. Fibrosis is the common response to all forms of liver injury. Common causes of cirrhosis and their therapies are listed in Table 37.6. Virtually all chronic liver diseases are known to lead to cirrhosis (see Fig. 37.6).

In the early stages of cirrhosis when the functions of the damaged liver are sufficiently well preserved (termed compensated cirrhosis), no signs or symptoms of liver damage may be present. Laboratory abnormalities usually appear before clinical findings, such as **ascites**, gynecomastia, palmar erythema, and portal hypertension develop. The earliest laboratory abnormalities to develop in cirrhosis are a fall in platelet count, an increase in PT, a decrease in the plasma albumin-to-globulin concentration ratio to less than 1, and an increase in the AST/ALT activity ratio to more than 1. In general, cirrhosis progresses to decompensation slowly, at a rate of approximately 3% per year; 10-year survival with compensated cirrhosis is 90%. Once decompensation occurs, 10-year survival is only approximately 20%. However, prognosis varies with etiology and may be influenced dramatically by response to treatment (as is the case in viral hepatitis B and C) or by abstinence from alcohol in **alcoholic liver disease**. Jaundice is a late finding in decompensated cirrhosis. Plasma

TABLE 37.6 Causes and Treatment of Cirrhosis

Cause	Treatment
Viral	
Hepatitis B	Administration of nucleoside or nucleotide HBV DNA polymerase inhibitors or pegylated α-interferon
Hepatitis C	Directly acting antiviral inhibitors of HCV
Toxic	
Alcohol	Abstinence, liver transplantation
Metabolic	
Hemochromatosis	Phlebotomy
Wilson disease	Penicillamine, zinc trientine
α_1-antitrypsin deficiency	Gene therapy, protein administration
Nonalcoholic fatty liver disease	Diet, exercise, insulin sensitizers
Biliary	
Primary biliary cholangitis	Ursodeoxycholic acid, obeticholic acid
Primary sclerosing cholangitis	Liver transplantation
Autoimmune hepatitis	Corticosteroids, azathioprine, other immunosuppressants
Idiopathic	Consider immunosuppression
Advanced cirrhosis, irrespective of cause	Liver transplantation

HBV, Hepatitis B virus; *HCV*, hepatitis C virus.

TABLE 37.7 Child-Pugh System for Classifying Severity of Cirrhosis

Feature	1 Point	2 Points	3 Points
Encephalopathy	None	Grade 1–2	Grade 3–4
Ascites	None	Slight	Moderate-severe
Albumin, g/dL	>3.5	2.8–3.5	<2.8
Prothrombin time, seconds prolonged	<4	4–6	>6
Bilirubin, mg/dL (μmol/L)	<4 (<68)	4–10 (68–170)	>10 (>170)

Scoring: <7 points: class A; 7 to 9 points: class B; >9 points: class C.

activities of aminotransferases are variable in cirrhosis and reflect underlying necroinflammatory activity. Clinical scales for assessing cirrhosis in widespread use include the Child-Pugh system (Table 37.7). Current thinking advocates the concept of "acute on chronic" liver failure in which acute exacerbations of liver disease due to any one of a variety of causes may cause acute worsening of liver function in cirrhosis. Tests commonly used for the assessment of hepatic function are summarized in Table 37.8.

TABLE 37.8 Tests of Hepatic Function and Injury

Test	Utility
Bilirubin	Diagnosing jaundice, modest correlation with severity
Alkaline phosphatase	Diagnosing disorders of metabolism and disorders of the newborn
Bilirubin fractionation	Diagnosing cholestasis and space-occupying lesions
AST	Sensitive test of hepatocellular disease; AST > ALT in alcoholic disease
ALT	Sensitive and more specific test of hepatocellular disease
γ-Glutamyltransferase	Prognostic indicator for increased cardiovascular and all-cause mortality
Albumin	Indicator of chronicity and severity
Prothrombin time	Indicator of severity of cholestasis

ALT, Alanine aminotransferase; *AST*, aspartate aminotransferase.

Biomarkers of Fibrosis

Due to the limitations of liver biopsy, several noninvasive methods have been developed to detect progressive liver fibrosis and stage liver disease. Over the past two decades, numerous serum markers or biomarkers have been identified and described. In contrast to a liver biopsy, biomarkers are less invasive, with minimal associated procedural morbidity. Serum biomarkers of fibrosis can be categorized into direct serum markers (measuring parameters directly related to both the fibrolytic and fibrogenic processes involved in liver matrix turnover) and indirect serum markers (combinations of serum parameters that are related to liver function, including AST and ALT).

Direct Biomarkers

Direct biomarkers exhibit biologic plausibility because they represent alterations of extracellular matrix (ECM) composition that occur during hepatic fibrosis. Combinations of direct biomarkers can result in superior diagnostic performance for the detection of fibrosis compared with their constituent components. The Enhanced Liver Fibrosis (ELF) test is a combination of hyaluronic acid (HA), tissue inhibitor of matrix metalloproteinase 1, and the amino terminal fragment of procollagen III that has been validated for the detection of fibrosis in a variety of liver diseases, including viral hepatitis B and C, NAFLD in children and adults, and PBC and **primary sclerosing cholangitis (PSC)**.

Indirect biomarkers. By contrast, indirect biomarkers reflect parameters that are altered because of changes in hepatic function that arise in the context of a particular stage of liver fibrosis. Indirect biomarkers include biochemical or hematologic variables that are synthesized or regulated by the liver (e.g., clotting factors, cholesterol, and bilirubin), or indicate inflammation (e.g., aminotransferases). Hitherto, the AST-to-platelet ratio index (APRI) has been one of the most

studied indirect biomarkers of liver fibrosis. The AST-to-platelet ratio is calculated by dividing AST by the URL of AST (for the sex of the patient) and the platelet count and multiplying by 100. Another indirect biomarker is Fib-4, which is a combination of AST, ALT, age, and platelet count.

Hybrid biomarkers. Indirect biomarkers can also be combined with direct biomarkers to form combination or hybrid biomarkers. Fibrometer is a hybrid biomarker that uses age, platelets, prothrombin index, AST, α_2-macroglobulin, HA, and urea. Hepascore is a combination of bilirubin, GGT, HA, α_2-macroglobulin, age, and sex. Fibrotest is a combination of GGT, bilirubin, haptoglobin, α_2-macroglobulin, apolipoprotein A1, age, and sex.

Hepatic Glycogenoses

The glycogenoses are a group of rare inherited disorders that are characterized by excessive and/or aberrant glycogen storage in various tissues. All are transmitted as autosomal recessive traits, except for type IV, which is sex linked. The hepatic glycogen storage diseases and their enzyme defects are discussed in more detail in Chapter 22.

Cholestatic Liver Disease

Cholestasis (stoppage or suppression of the flow of bile) is characterized by retention of bile within the excretory system. The term obstruction is often used inappropriately, because cholestasis frequently occurs without mechanical obstruction to the biliary tract, as with many drug reactions. Mechanical obstruction must always be excluded with imaging in the first instance before other causes are considered. The major cholestatic diseases include mechanical obstruction of the bile ducts, PBC, and PSC. There are specific types of cholestasis that can occur in pregnancy; these are discussed in Chapter 44.

Mechanical Bile Duct Obstruction

The most common cause of cholestasis is biliary tract obstruction by space-occupying lesions. Extrahepatic bile duct obstruction occurs most commonly as the result of gallstones in the common bile duct or because of tumors in the head of the pancreas or duodenum. Other causes of extrahepatic obstruction include (1) bile duct strictures, (2) extrinsic compression of the bile ducts by enlarged lymph nodes, (3) congenital biliary atresia, and (4) PSC. Extrahepatic obstruction is commonly associated with jaundice, especially when obstruction is complete. Transient increases in aminotransferases are more common with choledocholithiasis than with other causes of extrahepatic obstruction. Transient increases in carbohydrate antigen (CA) 19-9 occur with bile duct obstruction; this is an important consideration because CA 19-9 is often used as a diagnostic test for pancreatic and bile duct carcinoma (for more details see Chapter 20 on tumor markers).

Intrahepatic cholestasis caused by mechanical obstruction is also common but is rarely associated with jaundice or with visibly dilated ducts on imaging studies, although it may be associated with increased direct bilirubin. The commonest cause is drug-induced cholestasis, but viral infections can also cause intrahepatic cholestasis. Obstructive intrahepatic cholestasis leads to jaundice with lesions that are large or are located near the porta hepatis, where they may obstruct both hepatic ducts. Common causes of intrahepatic obstruction include tumors (particularly metastases), granulomatous diseases (such as sarcoidosis and tuberculosis), and infiltrative processes (such as lymphoma, leukemia, and extramedullary hematopoiesis).

Primary Biliary Cholangitis (PBC, Formerly Primary Biliary Cirrhosis)

PBC, or nonsuppurative destructive cholangitis, is an uncommon autoimmune disorder that targets intrahepatic bile ducts. Its prevalence is approximately 2 to 8 per 100,000 population in the developed world but is much lower in developing areas. The median age at onset is 50 years, and the female-to-male ratio is approximately 6:1. An association with HLA class II antigen DR8 has been noted in some populations. A family history of PBC is seen in 1% to 4% of cases. In up to 80% of cases, the condition is associated with other autoimmune processes, most commonly **Sjögren syndrome** and hypothyroidism (which often develops before the onset of PBC).

The pathogenesis of PBC is not well understood. However, it is known that destruction of the bile duct is mediated by T cells in the presence of upregulation of HLA class I antigens on hepatocytes and HLA class II antigens on biliary epithelial cells. At least 95% of patients have antimitochondrial antibodies that react against the dihydrolipoamide acyltransferase component of the pyruvate decarboxylase complex. Part of this complex is found on the apical surface of biliary epithelial cells, suggesting a role for this antigen as an immune target. In individuals with coexisting Sjögren syndrome, the antigen is also expressed on the surface of salivary gland cells.

Biochemical tests and direct markers of liver fibrosis have been studied as surrogate markers of disease severity and prognostic markers of survival in PBC. Of these, the ELF test (see earlier section on direct biomarkers of liver fibrosis) has been shown to be the most accurate in determining prognosis.

Primary Sclerosing Cholangitis

PSC is a chronic inflammatory disease of the biliary tree that most commonly affects the extrahepatic bile ducts; involvement of intrahepatic ducts, with extrahepatic involvement or as an isolated finding, is also possible. In contrast to PBC, PSC has a male predominance (60% to 70%) and a younger median age at onset (30 years). In 70% to 90% of patients, PSC is associated with ulcerative colitis, which usually (but not always) precedes onset of PSC; conversely, only approximately 2% to 4% of patients with ulcerative colitis develop PSC. This has led to speculation that bacterial antigens in portal blood might be involved in the pathogenesis of PSC.

An autoimmune component is likely because 97% of patients with PSC have one or more autoantibodies present in their plasma. The prevalence of PSC is similar to that of PBC, but it is most common in Northern Europe, where PSC is the most common indication for liver transplantation. A markedly increased prevalence of HLA antigens B8 and DR3 has been noted.

Drug-Induced Cholestasis

Drugs are a common cause of cholestasis, causing approximately 15% of cases. Drug reactions are especially common in older individuals, among whom up to 50% of individuals have increased enzymes because of medications. Nonmedicinal drugs are also increasingly recognized as a cause of cholestasis. Drugs can cause a cholestatic picture by two major mechanisms. In some cases, only conjugated bilirubin is increased, whereas canalicular enzymes are not elevated. This picture, which is often seen with estrogen and anabolic steroids, appears to be due to inhibition of production of the multidrug resistance protein-2, an ATP-binding cassette (ABC) transporter that transports drug-conjugates and divalent bile salt conjugates into bile. More commonly, drugs induce a cholestatic hepatitis, as discussed earlier. The National Institutes of Health website, http://livertox.nih.gov, and reviews provide further information and excellent support in the investigation of suspected drug-induced liver toxicity.

Gallstones

Gallstones are solid formations in the gallbladder that are composed of cholesterol and bile salts. Although they vary in chemical composition, they generally contain a mixture of cholesterol, bilirubin, calcium, and mucoproteins. In the United States, 70% to 85% of all gallstones are predominantly cholesterol, and more than 10% of the adult population is affected.

Three major types of gallstones are cholesterol, pigmented, and, the most common, mixed. These stones form whenever bile is supersaturated with cholesterol or unconjugated bilirubin. Whenever an increase in cholesterol or a decrease in bile acids or lecithin occurs, bile becomes lithogenic and cholesterol may precipitate. Factors that predispose to cholesterol hypersecretion include obesity, aging, certain drugs such as clofibrate and nicotine, and certain hormones such as estrogen. Factors that decrease bile acid secretion include terminal ileal disease and cholestatic diseases, such as PBC, PSC, and cystic fibrosis. Genetic factors also appear to be involved. Within racial groups, women are more frequently affected than men. Diet may play a role because it appears that people who ingest diets high in polyunsaturated fats have a higher incidence, whereas those with a diet high in fiber have a decreased incidence.

Pigmented gallstones are associated with conditions in which the bilirubin concentration is increased, such as hemolytic anemia, or when bilirubin becomes insoluble (i.e., deconjugated), such as occurs in cholestasis or chronic biliary infection.

Hepatic Tumors

The liver is host to a wide variety of benign and malignant primary tumors. It is the second most common site of metastases; metastatic tumors account for 90% to 95% of all hepatic malignancies. Primary tumors may arise from many cell lines in the liver, but they arise most commonly from parenchymal and biliary epithelial cells and from mesenchymal cells. The two most important primary liver tumors are HCC and cholangiocarcinoma.

Hepatocellular Carcinoma

HCC is the fifth most common cancer worldwide and a leading cause of cancer death; more than 500,000 cases occur annually, with a similar number of deaths. Wide geographic and ethnic variations are noted in the incidence, suggesting that both host and environmental factors are involved in its origin. Seventy-five percent of HCC cases occur in Asia, with an annual incidence of HCC in China of approximately 30 cases per 100,000 males. Worldwide, the incidence is twofold to threefold higher among men than among women. The incidence of HCC has been increasing in the United States and much of Europe because of the increasing frequency of cirrhosis caused by HCV; however, incidence has declined in many parts of the world because of the success in prevention of infection by HBV. Although cirrhosis is present in most patients with HCC, it is absent in approximately 25% to 30% of cases, often in association with HBV. More importantly, the presence of cirrhosis had been recognized before diagnosis of HCC in only one-third of cases. The incidence of HCC in cirrhosis varies greatly with the cause of cirrhosis, but viral hepatitis B and/or C is the commonest cause worldwide.

The tumor marker most widely used for screening purposes is alpha-fetoprotein (AFP); it is typically quantified using assays that measure its total concentration. Although it appears to be relatively sensitive, elevation of AFP is common in individuals with chronic hepatitis and cirrhosis, which is the group at highest risk for HCC.

LIVER TRANSPLANTATION

Clinical biochemistry, as part of the multidisciplinary transplant team, plays an important role in the management of patients before and after transplantation. Algorithms incorporating biochemical and other clinical parameters have been developed for the assessment of the need for transplantation in patients with acute and chronic liver disease. These include the Child-Pugh-Turcotte, **Model for End-Stage Liver Disease (MELD) Score**, and United Kingdom End-stage Liver Disease Score classifications. Post transplantation, the main focus of management is on detecting graft failure, organ rejection, and monitoring the efficacy of immunosuppression. Laboratories are closely involved in monitoring liver enzymes, tests of hepatic synthetic function, therapeutic drug monitoring, and monitoring for viruses, including recurrence of HBV or HCV and CMV, which may complicate the course of recovery in immune-suppressed patients.

REVIEW QUESTIONS

1. In the liver, bilirubin is conjugated to:
   a. vinyl groups.
   b. methyl groups.
   c. hydroxyl groups.
   d. glucuronide.
2. Functions of the liver include the synthesis of all of the following *except*:
   a. albumin.
   b. immunoglobulins.
   c. glycogen.
   d. coagulation factors.
3. In the liver, the small grooves between adjacent hepatocytes that carry bile to the gallbladder are the:
   a. cords.
   b. canaliculi.
   c. lobules.
   d. sinusoids.
4. Hepatocellular carcinoma (HCC) can be directly related to:
   a. an acute viral hepatitis infection.
   b. cholestasis.
   c. a chronic hepatitis B infection.
   d. the synthetic function of the liver.
5. Which type of viral hepatitis is usually spread parenterally by transfusion, shared needles, or dialysis and is considered the most common chronic viral infection in North America?
   a. Hepatitis B
   b. Hepatitis C
   c. Hepatitis A
   d. Cirrhosis
6. Laboratory tests that are initially run to determine the presence of any liver disease include:
   a. liver enzymes only.
   b. viral antigens and antibodies, serum cholesterol.
   c. hepatitis antigens and antibodies, coagulation times, and serum proteins.
   d. bilirubin, liver enzymes, prothrombin time (PT), and albumin.
7. Blockage of the bile ducts or blockage of bile flow from within the liver due to inflammation will stop normal bile flow. This is referred to as:
   a. hepatitis.
   b. hepatocellular carcinoma (HCC).
   c. cholestasis.
   d. cirrhosis.
8. A genetic disorder that is associated with elevated amounts of copper in the liver and other tissues and leads to decreased ceruloplasmin concentration in blood is:
   a. Wilson disease.
   b. Reye syndrome.
   c. cholestasis.
   d. autoimmune hepatitis.
9. A woman visits her physician with symptoms of jaundice, hepatic pain, and chalky-appearing stools. Lab values indicate elevated conjugated bilirubin, alkaline phosphatase (ALP) and γ–glutamyltransferase (GGT). These findings are most likely due to:
   a. hemolytic anemia.
   b. ineffective erythropoiesis.
   c. cholestasis due to gallstones.
   d. Reye syndrome.
10. The type of portal hypertension seen in the majority of cases is sinusoidal hypertension, which is most commonly caused by:
    a. blockage of the portal veins.
    b. hepatic vein occlusion.
    c. congestive heart failure.
    d. cirrhosis.

SUGGESTED READINGS

Choo, Q. L., Kuo, G., Weiner, A., et al. (1989). Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. *Science*, *244*, 359–362.

Day, C. P. (2011). Non-alcoholic fatty liver disease: a massive problem. *Clinical Medicine*, *11*, 176–178.

Dufour, D., Lott, J., Nolte, F., et al. (2000). Diagnosis and monitoring of hepatic injury. I. Performance characteristics of laboratory tests. *Clinical Chemistry*, *46*, 2027–2049.

EASL. (2012). Clinical practice guidelines. *Management of Chronic Hepatitis B*. <http://www.easl.eu/research/our-contributions/clinical-practice-guidelines/detail/management-of-chronic-hepatitis-b-virus-infection>.

Fontana, R., & Lok, A. (2002). Noninvasive monitoring of patients with chronic hepatitis C. *Hepatology*, *36*, S57–S64.

Friedman, S. L. (2010). Evolving challenges in hepatic fibrosis. *Nature Review Gastroenterology Hepatology*, *7*, 425–436.

Gow, P., Smallwood, R., Angus, P., et al. (2000). Diagnosis of Wilson's disease: an experience over three decades. *Gut*, *46*, 415–419.

Guha, I. N., Parkes, J., Roderick, P., et al. (2008). Non-invasive markers of fibrosis in nonalcoholic fatty liver disease: validating the European Liver Fibrosis panel and exploring simple markers. *Hepatology*, *47*, 455–460.

Hennes, E., Zeniya, M., Czaja, A., et al. (2008). Simplified criteria for the diagnosis of autoimmune hepatitis. *Hepatology*, *48*, 169–176.

Kamath, P., Wiesner, R., Malinchoc, M., et al. (2001). A model to predict survival in patients with end-stage liver disease. *Hepatology*, *33*, 464–470.

Lam, B. P., Jeffers, T., Younoszai, Z., et al. (2015). The changing landscape of hepatitis C virus therapy: focus on interferon-free treatment. *Therapeutic Advances Gastroenterology*, *8*, 298–312.

Lee, W. (2008). Acetaminophen-related acute liver failure in the United States. *Hepatology Research*, *38*, S3–S8.

Lindor, K., Gershwin, M., Poupon, R., et al. (2009). Primary biliary cirrhosis. *Hepatology*, *50*, 291–308.

Lucey, M., Marthurin, P., & Morgan, T. (2009). Alcoholic hepatitis. *New England Journal of Medicine*, *360*, 2758–2769.

McMahon, B. (2009). The natural history of chronic hepatitis B virus infection. *Hepatology*, *49*, S45–S55.

Pietrangelo, A. (2010). Hereditary hemochromatosis: pathogenesis, diagnosis and treatment. *Gastroenterology*, *139*, 393–408.

Rosenberg, W. M., Voelker, M., Thiel, R., et al. (2004). Serum markers detect the presence of liver fibrosis: a cohort study. *Gastroenterology*, *127*, 1704–1713.

Zimmerman, H. (2000). Drug-induced liver disease. *Clinical Liver Disease*, *4*, 73–96.

Gastric, Pancreatic, and Intestinal Function

Roy A. Sherwood, Natalie E. Walsham, Ingvar Bjarnason

OBJECTIVES

1. Define the following terms:
   - Digestion
   - Absorption
   - *Helicobacter pylori*
   - Celiac disease
   - Steatorrhea
   - Diarrhea
2. List and describe the three phases of digestion.
3. Describe the structure and function of the stomach, intestinal tract, and pancreas.
4. List the main enzymes involved in the digestion of carbohydrates, fats, and proteins.
5. List the tests used to investigate possible celiac disease, lactase deficiency, and bacterial overgrowth and describe the test principles.
6. State the use of the following tests:
   - Serum gastrin
   - Fecal elastase
   - Antibodies against tissue transglutaminase
   - Fecal calprotectin

KEY WORDS AND DEFINITIONS

Acute pancreatitis An acute episode of destruction of the pancreas due to the release of active pancreatic enzymes into pancreatic tissue.

Breath tests Tests that detect products of bacterial metabolism in the gut or products of human metabolism by measuring compounds in breath.

Celiac disease A disease caused by the destructive interaction of gluten with intestinal mucosa causing malabsorption.

Chronic pancreatitis An inflammatory disease characterized by persistent and progressive destruction of the pancreas.

Crohn disease A chronic inflammatory bowel disease that may affect any part of the digestive tract.

Cystic fibrosis An inherited disease caused by genetic alteration of a transmembrane conductance regulator protein (CFTR) that can lead to chronic pancreatitis and pulmonary disease.

Diarrhea The passage of liquid stools more than three times daily and/or a stool weight greater than 200 g/day.

Digestion The conversion of food in the intestines into compounds capable of being absorbed.

Gastrin A group of peptide hormones secreted by gastrointestinal mucosal cells that stimulates the production of acid.

Gastritis Inflammation of the mucosa of the stomach.

Helicobacter pylori A bacterium found in the stomach that can be associated with inflammation.

Lactose intolerance A condition due to lactase deficiency leading to the malabsorption of lactose resulting in symptoms following consumption of foods containing lactose.

Malabsorption An abnormality in the absorption of nutrients.

Maldigestion An abnormality of the digestive process due to dysfunction of the pancreas or small intestine.

Steatorrhea A condition of excessive fat in feces.

Ulcerative colitis An inflammatory bowel disease predominantly localized to the large intestine.

INTRODUCTION TO THE ANATOMY AND PHYSIOLOGY OF THE GASTROINTESTINAL TRACT

The major organs of the gastrointestinal (GI) tract include the stomach, small and large intestines, pancreas, liver (see Chapter 37), and gallbladder, all of which are involved in the digestive processes that commence with the ingestion of food and water and culminate with the excretion of waste products as feces.

Anatomy

The GI tract is a hollow tube approximately 8 m in length beginning with the mouth and ending with the anus. The esophagus, which is approximately 25 cm in length, is a muscular tube connecting the pharynx to the stomach. The laboratory has little role in the investigation of disorders of the esophagus.

Stomach

The stomach consists of three major zones: the cardiac zone, the body, and the pyloric zone (Fig. 38.1). The upper cardiac zone includes the fundus and contains mucus and pepsinogen II–secreting surface epithelial cells and endocrine secreting cells. The body of the stomach contains cells of many different types: surface epithelial cells that secrete mucus, parietal cells that secrete hydrochloric acid and intrinsic factor, chief cells that secrete groups I and II pepsinogens, enterochromaffin cells that secrete serotonin, and other endocrine cells. The pyloric zone is subdivided into the antrum (approximately the distal third of the stomach), the pyloric canal, and the sphincter. The cells of the pyloric zone secrete mucus, group II pepsinogens, serotonin, and **gastrin**, but not hydrochloric acid. In the stomach, food is converted into a semifluid, homogeneous material (chyme) that passes through the pyloric sphincter into the small intestine.

Small Intestine

The small intestine consists of three parts: the duodenum, jejunum, and ileum. In adults, the small intestine is approximately 6 m in length and its cross section decreases as it proceeds distally. The duodenum is approximately 25 cm long and is the shortest part of the small intestine. The jejunum and ileum make up the remainder of the small intestine, with the ileum constituting the distal 3.5 m.

The wall of the small intestine consists of four layers: mucosa, submucosa, muscular layer, and serosa/adventitia. The internal surface of the upper part of the small intestine contains valve-like circular folds (valvulae conniventes or plicae circulares) that project 3 to 10 mm into the lumen. Covering the entire mucous surface are small (1-mm) finger-like projections (villi). The luminal surface of each epithelial cell consists of some 1700 microvilli projecting approximately 1 μm from the cell. The folds, villi, and microvilli increase the absorptive surface area 600-fold to approximately 250 m^2, comparable to the area of a doubles lawn tennis court.

Large Intestine

The large intestine is approximately 1.5 m in length, extending from the ileum to the anus and including the cecum, appendix, colon, rectum, and anal canal. The cecum is a blind pouch that begins the large intestine; it is connected to the terminal ileum via the ileocecal sphincter (or valve). The colon is approximately 1 m long and is divided into the ascending, transverse, descending, and sigmoid sections. The sigmoid colon connects to the rectum, which is approximately 15 cm long and connects to the anal canal.

Pancreas

The pancreas is 12 to 15 cm long and lies across the posterior wall of the abdominal cavity. The head is located in the duodenal curve, with the body and tail extending to the left to the spleen (Fig. 38.2). The pancreas secretes a juice containing digestive enzymes and bicarbonate; this enters the duodenum through the ampulla of Vater and the sphincter of Oddi to mix with the bolus of food coming from the stomach.

Phases of Digestion

The processes of **digestion** can be divided into neurogenic, gastric, and intestinal phases.

Neurogenic Phase

The neurogenic, or cephalic, phase is initiated by the intake of food into the mouth; the sight, smell, and taste of food stimulates the cerebral cortex and subsequently the vagal nuclei.

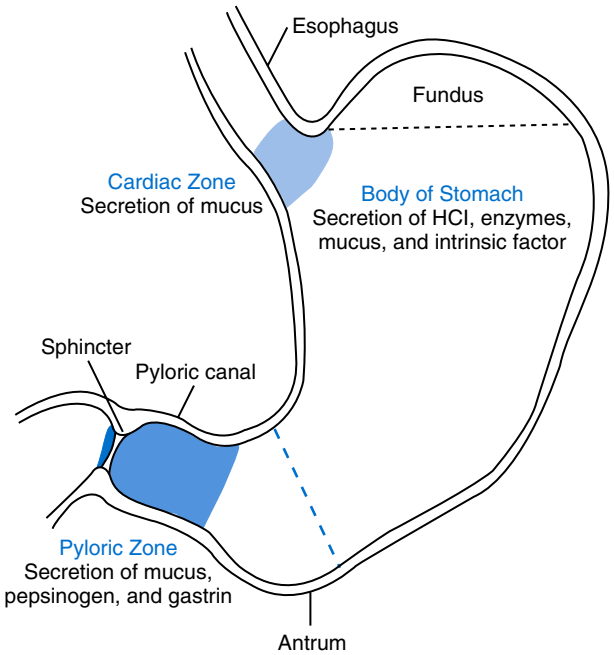


Fig. 38.1 Schematic drawing of the stomach with major zones.

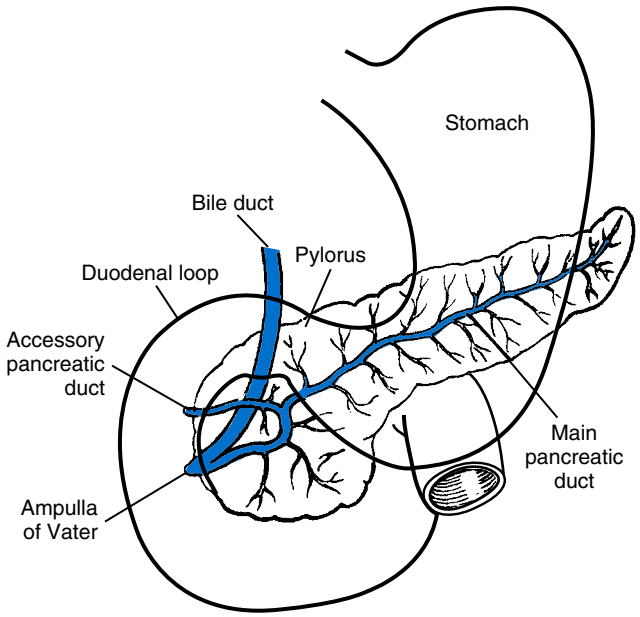


Fig. 38.2 Cross section through the pancreas.

Gastric Phase

When food enters the stomach, the resulting distention initiates the gastric phase of digestion, which is mediated by local and vagal reflexes. Hydrochloric acid release is caused by direct vagal stimulation of the parietal cells, local distention of the antrum, and vagal stimulation of antral cells to secrete gastrin, which in turn stimulates the release of hydrochloric acid from parietal cells and release of gastrin stimulated by the near neutralization (pH 5.0 to 7.0) of the stomach contents entering the pyloric zone. Gastrin also stimulates antral motility, the secretion of pepsinogens and of pancreatic fluid rich in enzymes, and the release of hormones such as secretin, insulin, acetylcholine, somatostatin, and pancreatic polypeptide. As a result of the acid environment, pepsinogen is rapidly converted to the active proteolytic enzyme pepsin. Food is mixed by contractions of the stomach and is partially degraded into chyme by the chemical secretions of the stomach. The pylorus plays a role in emptying chyme into the duodenum by virtue of its strong musculature.

Intestinal Phase

The intestinal phase of digestion begins when the weakly acidic digestive products of proteins enter the duodenum. Many hormones and other regulatory peptides are released by both neural and local stimulation and act within the GI tract to regulate digestion and absorption. Digestion, absorption, and storage functions are stimulated or inhibited by different hormones, creating a control system that regulates the action of intestinal hormones and provides for the secretion of bile acids, bicarbonate, and numerous enzymes involved in the digestion of food.

During the intestinal phase, carbohydrates, proteins, and fats are broken down and absorbed. Most nutrients, including vitamins and minerals, have been absorbed by the time the food passes from the jejunum and ileum into the large intestine. In the large intestine, water is actively absorbed, electrolyte balance is regulated, and bacterial actions take place. These processes result in the formation of feces.

Processes of Digestion and Absorption

The total quantity of fluid absorbed each day by the gut is estimated to be approximately 9 L, which is composed of 2 L oral intake, 1.5 L saliva, 2.5 L gastric juice, 0.5 L bile, 1.5 L pancreatic juice, and 1 L intestinal secretions. More than 90% of this fluid is absorbed in the small intestine. The maximal absorptive capacity for fluid is probably at least 20 L. Several hundred grams of carbohydrates, 100 g of fat, and 50 to 100 g of amino acids are absorbed daily in the small gut, but the maximal absorptive capacity may be 10 times higher. This considerable reserve capacity may explain the lack of symptoms from mild disease processes. The efficiency of absorption is due to the unique features of the absorptive surface of the bowel and the relationship of the epithelial cells to the underlying rich vascular plexus and the lymphatic vessels.

Digestion of food starts in the mouth through the action of salivary amylase and lingual lipase, but it principally takes place both within the lumen of the small intestine and at the mucosal (brush-border) surface. Defects of digestion may occur at one or more stages of this process. The terms **maldigestion** and **malabsorption** refer to different functional abnormalities. *Maldigestion* is a dysfunction of the digestive process that may occur at various sites in the GI tract. For example, a reduction in the acidity in the stomach will reduce peptic digestion of protein, whereas hyperacidity of the duodenum can inactivate pancreatic enzymes; loss of brush-border enzymes in the small intestine can prevent oligosaccharides and disaccharides from being further hydrolyzed; and pancreatic insufficiency will reduce intraluminal enzyme activity in the small gut, causing maldigestion of fats and proteins. Inherited disorders of the exocrine pancreas can cause maldigestion by causing pancreatic insufficiency secondary to chronic pancreatic inflammation. In contrast, malabsorption is strictly a dysfunction of the absorptive process in the small gut resulting from a reduction in the size of the absorptive surface caused by responses to factors including gluten, inflammation, infection, surgical resection, and infiltration. Various transport defects also lead to malabsorption of specific substances (e.g., glucose-galactose malabsorption, and zinc deficiency in the congenital disorder acrodermatitis enteropathica). In clinical practice, however, the term *malabsorption* is often used to encompass all aspects of impaired digestion and absorption.

In the following sections, the digestion and absorption of carbohydrates, fats, and proteins is discussed separately. It must be remembered, however, that a complex interplay takes place among nutrients, regulatory peptides, enzymes, gallbladder and pancreatic function, and bowel motility, leading to an integrated absorptive process that commences with the ingestion of food and culminates in the excretion of feces.

Digestion and Absorption of Carbohydrates

After the salivary and pancreatic α-amylases have acted on dietary starch and glycogen, the carbohydrate content of the small intestine will consist of newly formed maltose; ingested monosaccharides; dietary disaccharides such as lactose, sucrose, maltose, and trehalose; oligosaccharides such as dextrins and maltotriose; and indigestible oligosaccharides and polysaccharides such as cellulose, agar, and other dietary fibers.

The brush-border enzymes with disaccharidase and oligosaccharidase activity are listed in Table 38.1. The sucrase-isomaltase complex comprises most (80%) of the sucrase, isomaltase, and maltase activity of the small intestine. It hydrolyzes sucrose to its constituent monosaccharides, cleaves glucose from α-limit dextrins with 1,6 bonds, and hydrolyzes maltose. The activity of the complex is four- to fivefold greater in the jejunum than the ileum. Changes in diet have a marked effect on the expression of the complex; starvation leads to a rapid decline in activity that is quickly restored on refeeding. Secretion of all small intestinal saccharidases may decrease with infection or inflammation of the small bowel to the extent that carbohydrate malabsorption occurs, leading to **diarrhea**, flatulence, and weight loss.

TABLE 38.1 Brush-Border Oligosaccharidases

Enzyme	Principal Substrate	Products
Lactase (EC 3.2.1.23)	Lactose	Glucose + galactose
Sucrase (EC 3.2.1.48)	Maltose/sucrose	Glucose or fructose + glucose
Isomaltase (EC 3.2.1.10)	1,6-α-linkages in Isomaltose and α-dextrins	Glucose
	Maltose	Glucose
Trehalase (EC 3.2.1.28)	Trehalose	Glucose
Glucoamylase complex (EC 3.2.1.20)		
Glucoamylase (maltase)-1 and glucoamylase (maltase)-2 (have similar substrate specificities)	1,4-α-linkages at nonreducing ends of amylose, amylopectin, glycogen, and straight-chain 1,4-α-glucopyranosyl oligomers, including maltose	Glucose

From Semenza, G., Auricchio, S., & Mantei, N. (2001). Small-intestinal disaccharidases. In C. R. Scriver, A. L. Beaudet, W. S. Sly, & D. Valle (Eds.), *The metabolic and molecular bases of inherited disease* (8th ed., pp. 1623–1650). New York: McGraw-Hill.

The lactase–phlorizin hydrolase complex is the only brush-border enzyme able to hydrolyze lactose and is therefore essential for the survival of mammals early in life. Infectious and inflammatory diseases greatly reduce lactase–phlorizin hydrolase activity, leading to symptomatic intolerance to milk (bloating, abdominal pain, diarrhea, and flatulence). The developmental regulation of lactase is discussed later, in the section on disaccharidase deficiencies. Also present in the brush-border is the α-glucosidase maltase-glucoamylase, which removes individual glucose molecules from the nonreducing end of α(1,4)-oligosaccharides and disaccharides. This enzyme accounts for approximately 20% of the total maltase activity of the small intestine. Trehalase is also found in the brush-border of the small intestine and hydrolyzes trehalose, an α(1,1)-disaccharide of glucose found in yeast and mushrooms. The developmental pattern of trehalase appears to follow that of sucrase-isomaltase.

In addition to their actions on disaccharides, the brush-border enzymes further hydrolyze the products of amylase action, including maltose, maltotriose, and α-limit dextrins. The brush-border enzymes appear to act in an integrated manner in that a flow of substrate occurs from glucoamylase and isomaltase to sucrose, producing the monosaccharides glucose, galactose, and fructose. These monosaccharides are transported into enterocytes by the sodium-dependent glucose (and galactose) transporter (SGLT1) and the GLUT5 transporter, which transports fructose across the apical membrane of the enterocyte. Absorbed glucose and fructose are transported across the basolateral membrane, out of the enterocyte, and into the portal system by the GLUT2 transporter.

Carbohydrate digestion in the small intestine is not always complete. Up to 10% of ingested starch and sucrose can pass undigested and unabsorbed into the colon. It has been estimated that colonic bacteria require 70 g of carbohydrate per day. Much of this is derived from endogenous sources, such as glycoproteins from GI secretions, with the remainder coming from unabsorbed dietary carbohydrate and dietary fiber. Bacterial action creates short-chain fatty acids that are rapidly absorbed by the colonic mucosa and are thought to provide fuel for the colonocytes. Starch and oligosaccharides are osmotically active; they draw water into the gut and retain luminal fluidity. The colon, however, can absorb up to four times the normal colonic water load; for this reason diarrhea is not always present in oligosaccharide malabsorption.

Digestion and Absorption of Lipids

The recommended daily fat intake in Europe and North America is 70 to 85 g. Less than 5 g/24 h is recoverable in feces, indicating the overall efficiency of the normal processes of fat digestion and absorption. Most dietary fat is in the form of long-chain triacylglycerols (triglycerides). Pancreatic lipase is quantitatively the most important hydrolytic enzyme, but the contribution of gastric lipase to overall hydrolysis should not be underestimated. Gastric lipase is secreted by the gastric mucosa and normally accounts for up to 17.5% of fatty acids released from triglycerides following a meal. Lipase has a wide pH optimum and is active in both the stomach and duodenum. This nonpancreatic lipase may have a significant role in lipid digestion when pancreatic function is impaired and in the neonatal period before pancreatic lipase activity is fully developed. A lingual lipase is also produced but is thought to be of little significance in humans.

Fats are first emulsified in the stomach by its churning action and are stabilized by interaction with luminal lecithin and protein fragments. The lingual and gastric lipases do not require bile salts or cofactors for their action; they have a pH optimum of 3 to 6, and their action produces 1,2-diacylglycerols and fatty acids. These products further stabilize the surface of the triglyceride emulsion and in the duodenum promote the binding of pancreatic colipase. In addition, the liberated fatty acids stimulate release of cholecystokinin (CCK) from the duodenal mucosa.

In the presence of bile salts and colipase, pancreatic lipase acts at the oil-water interface of the triglyceride emulsion to produce fatty acids and 2-monoacylglycerols. Colipase is secreted by the pancreas as an inactive proenzyme and is then activated by trypsin. Other significant enzymes involved in the breakdown of fats are cholesterol ester hydrolase, phospholipase A2, and a nonspecific bile salt–activated lipase.

Only a small proportion of ingested triacylglycerol is completely hydrolyzed to glycerol and fatty acids. These products form micelles with bile salts and lysophosphoglycerides; the micelles convey the nonpolar lipid molecules from the lumen

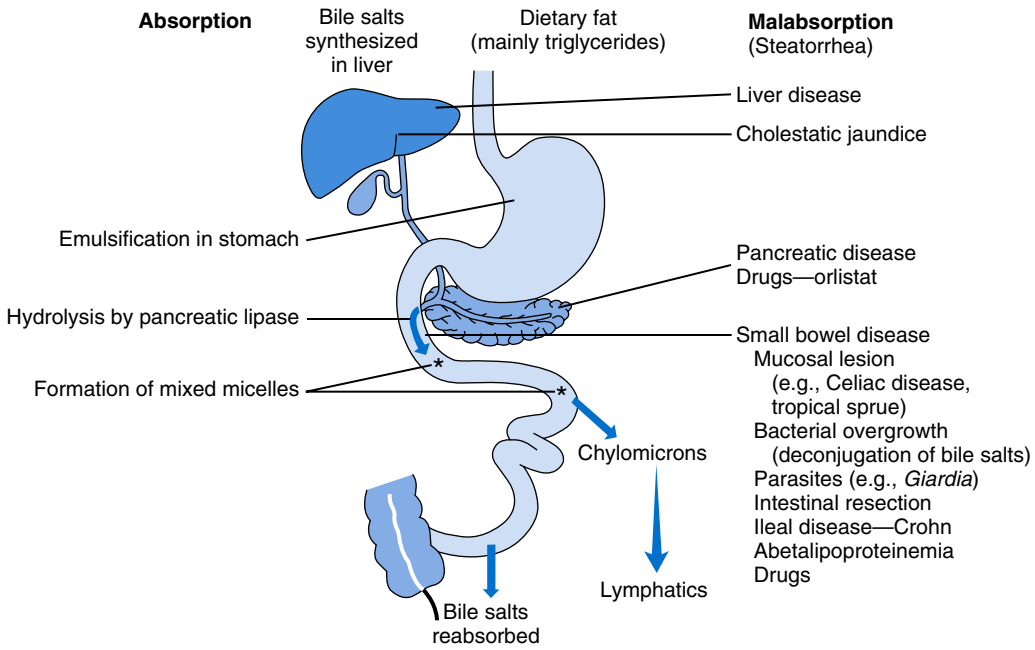


Fig. 38.3 Summary of the processes involved in fat absorption and malabsorption. (From Clark, M. L., & Silk, D. B. [2005]. Gastrointestinal disease. In P. Kumar, & M. Clark [Eds.], *Clinical medicine* [6th ed., pp. 265–345]. Edinburgh: Saunders.)

to the epithelial cell surface and dissociate there to produce a high concentration of monoglycerols, lysophosphoglycerides, and fatty acids, which partition into the mucosal cell. Absorption involves both passive and active transport processes and is facilitated by a fatty acid–binding protein in the cytosol of the cell that has a high affinity for fatty acids. Within the cell, triacylglycerols are resynthesized from the absorbed 2-monoacylglycerols and fatty acids. The triacylglycerols— together with phospholipids, cholesterol and its ester, fat-soluble vitamins, and apolipoprotein B-48—are formed into chylomicrons that are then released by exocytosis into the lymphatic system of the small bowel. The absorption of long-chain fatty acids is facilitated by transmembrane fatty acid transport proteins.

From the lymphatic system, chylomicrons enter the bloodstream via the thoracic duct and are distributed to the liver, adipose tissue, and other organs. Medium- and short-chain fatty acids (chain length <12 carbon atoms) in mixed triglycerides are preferentially split by lipases and pass into the aqueous phase, from which they are rapidly absorbed. Medium-chain triglycerides can be absorbed without complete lipolysis and in the absence of bile. They do not require micellar solubilization and are transported from the intestinal epithelial cells predominantly via the hepatic portal vein. Fig. 38.3 summarizes the processes involved in fat absorption and conditions that compromise the efficiency of one or more stages in these processes that can result in fat malabsorption.

Digestion and Absorption of Proteins

The average daily intake of protein in developed countries is approximately 100 g, compared with an estimated requirement for adults of 50 to 70 g. Another 50 to 60 g of protein enters the intestinal tract daily in GI secretions and from desquamated mucosal cells. Normal daily fecal loss of protein is about 10 g.

Protein digestion is initiated in the stomach by the action of pepsin in a highly acidic medium. The acidity also helps denature dietary proteins, unfolding the polypeptide chains for better access by the gastric, pancreatic, and intestinal proteolytic enzymes. The polypeptides and amino acids produced in the stomach by the action of pepsin are potent secretagogues for hormones that stimulate the pancreas and intestine. Stimulated pancreatic secretion contains proenzyme forms of the proteolytic enzymes trypsin, chymotrypsin, elastase, exopeptidases, and carboxypeptidases. Proteolytic enzymes may be endopeptidases (e.g., pepsin, trypsin, chymotrypsin, elastase), which hydrolyze peptide bonds within the polypeptide chain, or exopeptidases, which hydrolyze peptide bonds of the terminal amino acids (e.g., carboxypeptidase, aminopeptidase). Stimulation of the intestine by GI hormones liberates several proteolytic enzymes from the brush-border. One of them, enterokinase, selectively cleaves a hexapeptide from the N-terminus of trypsinogen to form trypsin. Trypsin then activates more trypsin (autocatalysis) and also converts other pancreatic proenzymes into their active forms. The action of the pancreatic enzymes on partially digested proteins within the lumen produces peptides that are 2 to 6 amino acid residues in length as well as single amino acids. The peptides are largely hydrolyzed to single amino acids by the aminopeptidases and dipeptidases of the brush-border before absorption, although some dipeptides and tripeptides are absorbed and hydrolyzed to amino acids by cytosolic peptidases within the enterocytes. Multiple carrier systems with overlapping specificities for the 20 essential amino acids are involved in the transport of amino acids into cells. Absorption of amino acids

by these transport systems is faster in the jejunum than the ileum. The amino acids pass across the enterocyte basolateral membrane by passive diffusion as well as by active transport systems, which are distinct from those at the brush-border membrane. The rich vascular plexus is drained by the portal circulation, and it is by this route that absorbed amino acids reach the liver and then the systemic circulation.

Mucosal diseases may affect protein assimilation through a number of mechanisms. Reduction in the number of mucosal cells decreases peptidase activity in the intestine and the absorptive capacity for amino acids. Disease may increase the turnover of intestinal cells and their rate of desquamation. This cell loss, together with increased losses of plasma proteins from the damaged intestinal surface, can cause a negative nitrogen balance. Surgical resection of the intestine not only reduces the total absorptive surface area but may also remove a segment of the gut that is specialized for the absorption of certain nutrients (e.g., resection of the distal ileum removes the active transport system for vitamin B_{12}–intrinsic factor complex). Resection may also alter intestinal motility, leading to stasis and bacterial overgrowth, which can intensify a negative nitrogen balance. Also, rare hereditary defects in amino acid transporters (e.g., Hartnup disease) may produce distinct syndromes.

STOMACH: DISEASES AND LABORATORY INVESTIGATIONS

Growth in endoscopic procedures, with direct visualization of the interior of the stomach, has largely removed the need for the laboratory to carry out investigation of gastric contents. Situations remain, however, in which the laboratory continues to play a role in diagnosing gastric diseases and in monitoring the effectiveness of treatment.

Helicobacter pylori

In 1985, an association was made between the presence of a spiral-shaped bacterium, *Helicobacter pylori,* and peptic ulcer diseases. *H. pylori* is now accepted to be the predominant cause of gastric and duodenal ulcers, the remainder being associated with the long-term use of nonsteroidal antiinflammatory drugs (NSAIDs) and, rarely, gastrinomas. Estimates suggest that *H. pylori* is present in the mucus layer of the stomach in half the world's population. Some 30% to 50% of adults in Europe and at least 20% of adults in the United States are infected with the bacterium. Chronic infection produces an inflammatory response (**gastritis**) and carries an increased risk for developing a peptic ulcer (3- to 10-fold) and/or adenocarcinoma (2- to 10-fold). Up to 90% of gastric cancer patients are infected with *H. pylori*, compared with 40% to 60% of age-matched controls. *H. pylori* may cause dyspepsia in the absence of an ulcer, and current recommendations suggest a low threshold for testing for *H. pylori*; some advocate treatment without testing.

The mode of transmission of *H. pylori* is unclear. In many cases the infection appears to originate in childhood, presumably by the fecal-oral route, because the prevalence is higher in developing countries and is inversely related to food hygiene. Almost all individuals infected with *H. pylori* develop chronic gastritis, but only 10% of cases manifest as peptic ulcers. *H. pylori* infection predominantly affects the gastric mucosa, with the antrum usually the most densely colonized area. At least 95% of patients with duodenal ulcers have *H. pylori* infection, and eradication of the organism results in healing of the ulcer and a reduction in relapse rates.

H. pylori produces urease, and hydrolysis of this endogenously produced urea to bicarbonate and ammonia may create a more hospitable environment for the survival of the organism in the stomach. This ability of *H. pylori* to hydrolyze urea forms the basis of urea **breath tests** and direct urease tests on gastric biopsy samples.

Diagnostic Tests for *Helicobacter pylori*

In theory it would be possible to carry out endoscopy on all patients who have symptoms that could be associated with *H. pylori* infection, but in the real world this is not practical, nor would it be acceptable to the patient population. Numerous invasive and noninvasive diagnostic tests for *H. pylori* have been described (Box 38.1) and many have been reviewed.

At gastroduodenoscopy, biopsies can be taken from the gastric mucosa in the antrum or body of the stomach from which the organism can be visually detected or cultured. False-negative results may also occur when biopsies are taken during treatment with proton pump inhibitors (PPIs) or within 2 weeks of stopping PPI therapy, because these drugs alter the intragastric distribution of *H. pylori* and suppress its activity. PPIs can also lead to false-negative urea breath test results. Histamine (H_2)-receptor antagonists should be stopped at least 24 hours before a breath test. Antacids do not affect the test results. Commercially available kits can be used to identify *H. pylori* in gastric biopsy samples. These are based on a gel that incorporates urea and an indicator that changes color at an alkaline pH. The action of urease present in *H. pylori* cleaves the urea to bicarbonate and ammonia, raising the pH and inducing a color change in the indicator.

BOX 38.1 Diagnostic Tests for *Helicobacter pylori*

Invasive Tests: Using Gastric Mucosal Biopsy Samples
- Histology: microscopy after Giemsa or silver staining
- Histology: microscopy after immunohistochemical staining
- Direct urease test: biopsy included in urea/indicator solution—visual end point
- Culture: incubation in suitable media for 4 to 10 days
- Polymerase chain reaction: amplification of specific DNA sequences

Noninvasive Tests: Using Breath, Blood, Saliva, or Feces
- Breath tests: rise in $^{14}CO_2$ or $^{13}CO_2$ after ingestion of ^{14}C- or ^{13}C-labeled urea
- Serum, saliva, or feces tests: detection of IgG antibody or *H. pylori* antigen

Tests for *H. pylori* are required for the diagnosis of infection and in some situations to ascertain whether eradication therapy has been successful. High sensitivity is required to ensure that positive findings are not missed; similarly, high specificity is essential to prevent inappropriate use of eradication therapy. The Maastricht III Consensus Guidelines recommend a "test and treat" strategy, using a breath test or stool antigen test, in adults with appropriate dyspeptic symptoms who are younger than 45 years of age. The age limit may vary depending on local prevalence and the age distribution of gastric cancer. Successful eradication of *H. pylori* should be confirmed with the urea breath test or by a monoclonal antibody–based stool antigen test if urea breath tests are not available. Testing to confirm eradication should be done at least 4 weeks after completion of the course of treatment.

The urea breath test for *H. pylori* is currently the most widely used method for the noninvasive diagnosis of infection. Urea labeled with carbon-13 is given orally as a drink or a capsule. *H. pylori* rapidly hydrolyzes the urea to produce labeled bicarbonate, which is absorbed into the bloodstream and broken down to be exhaled as $^{13}CO_2$. In the absence of urease, the urea is absorbed intact and excreted by the kidney. Breath samples are collected before and 45 to 60 minutes after drinking the labeled urea. The detection of labeled carbon, which is not radioactive, is usually carried out by mass spectrometry. The breath test has sensitivity and specificity for *H. pylori* in excess of 95% and can be used both for diagnosis and to assess the success of eradication therapy.

Serologic methods are available to detect *H. pylori*–specific antibodies (immunoglobulin G [IgG] or IgA), but they have some drawbacks compared with the urea breath test. The systemic antibody response is variable, with equivocal results often occurring in subjects above 50 years of age. The sensitivity (92%) and specificity (83%) are also lower than those for the breath test. Serologic tests cannot be used to confirm eradication of the bacterium because of the persistence of the antibodies for variable periods after completion of treatment. Point-of-care devices exist to detect *H. pylori* antibodies, but current evidence is that these perform poorly in terms of both sensitivity and specificity and cannot be recommended. These tests may be useful in specific situations, such as when PPI therapy cannot be withheld because of ulcers that are bleeding.

H. pylori is shed in feces, and several tests have been described that can detect the organism. Polyclonal or monoclonal antibodies to *H. pylori* can be configured into various immunoassay formats, although sensitivity and specificity are lower than for breath tests. Commercial kits are available that use the polymerase chain reaction to amplify nuclear sequences specific for *H. pylori* in feces (or saliva); they have a sensitivity and specificity of 95% and 94%, respectively. In time, stool tests for *H. pylori* may replace the urea breath test, particularly in assessing whether eradication has been successful.

POINTS TO REMEMBER

Helicobacter pylori

- *H. pylori* is the predominant cause of gastric and duodenal ulcers.
- There is an increased risk for gastric cancer with *H. pylori* infection.
- The urea breath test is useful in the diagnosis of *H. pylori* infection.
- Tests based on serologic/molecular techniques are available to detect *H. pylori* in blood or feces.

Gastric Acid Secretion and Gastrinomas

Before the discovery of the role of *H. pylori,* patients with gastric or duodenal ulcers were often tested for hyperacidity of the stomach, either in the basal state or after stimulation. Patients with duodenal ulcers are typically hypersecretors of acid, whereas those with peptic ulcers are more often normal or low secretors, but there was significant overlap between the two groups of patients. Collection of gastric juice and analysis of acid output was at one time extensively carried out in the investigation of possible gastrinomas. This invasive technique has now been replaced by the greater availability of plasma gastrin measurements, endoscopy, and imaging modalities, including computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), and octreotide scanning.

Gastrin

Three molecular forms of gastrin occur in blood and tissues: G-34, G-17, and G-14. Gastrins originate from the cleavage of a single precursor, preprogastrin—a peptide consisting of 101 amino acids. The smallest peptide sequence of gastrin possessing biologic activity is the carboxy-terminal tetrapeptide (G-4), but it is only 10% to 20% as potent as G-17. A synthetic pentapeptide (pentagastrin) was used in the past to stimulate acid secretion for collection and analysis, but it is rarely used now.

Gastrin is produced and stored mainly by endocrine cells (G cells) of the antral mucosa and to a lesser extent by G cells of the proximal duodenum and Δ cells of the pancreatic islets. After secretion, gastrin is transported by the blood through the liver to the parietal cells of the fundus of the stomach, where it stimulates the secretion of gastric acid. Gastrin also stimulates secretion of gastric pepsinogens and intrinsic factor by the gastric mucosa, release of secretin from the small intestinal mucosa, and secretion of pancreatic bicarbonate and enzymes and hepatic bile; it increases gastric and intestinal motility, mucosal growth, and blood flow to the stomach. It is secreted in response to antral distention from food and by the presence of amino acids, peptides, and polypeptides in the stomach from partially digested proteins. Other stimuli of gastrin include alcohol, caffeine, insulin-induced hypoglycemia, and calcium.

Maximal secretion of gastrin occurs at an antral pH of 5 to 7. At pH 2.5, the secretion of gastrin is reduced by approximately 80%, with maximal suppression occurring at pH 1.

Secretion is inhibited by the direct action of acid on the G cells. This negative feedback prevents excess acid production regardless of the stimulant.

Gastrinoma and the Zollinger-Ellison Syndrome

In 1955, Zollinger and Ellison described a syndrome consisting of multiple peptic ulcers, gastric hypersecretion, and non–β islet cell tumors of the pancreas secreting gastrin. Gastrinomas are a rare cause (<0.5%) of gastroduodenal ulcers. The persistently high circulating gastrin concentrations lead to the hypersecretion of gastric acid and increased parietal cell mass, as the hormone is trophic to parietal cells. Most cases occur between the ages of 30 and 50 years, although the condition has been reported in patients as young as 7 and as old as 90 years of age. It is more common in men (60% of cases) than in women.

Symptoms may be those of *H. pylori* peptic ulcer disease, but they can also include diarrhea (secondary to inactivation of pancreatic enzymes in the acidic environment). Duodenal ulceration and ulcers resistant to standard therapies must be considered suspicious. Gastrinomas are most often sporadic, but they may also be associated with multiple endocrine neoplasia type 1 (MEN1, Wermer syndrome—see Chapter 20), a syndrome characterized by the presence of two or more tumors sited in the pituitary, parathyroid glands, or pancreas.

Gastrinomas are commonly located in the pancreas, but they can arise from the stomach, duodenum, or other tissues. They are more often (60%) malignant than benign, with metastases frequently present at the time of diagnosis. Measurement of plasma gastrin in a fasting sample is the initial step in aiding the differential diagnosis. A mildly elevated plasma gastrin concentration may be observed in long-term PPI therapy, hypochlorhydria, pernicious anemia, and G-cell hyperplasia. *H. pylori* infection can also lead to increased plasma gastrin and in some cases atrophic gastritis. Increases in plasma gastrin in chronic renal failure appear to be related to the severity of renal impairment.

Fasting plasma gastrin concentrations in the Zollinger-Ellison syndrome are usually markedly increased, and a concentration of more than 10 times the upper reference limit in the presence of gastric acid hypersecretion is virtually diagnostic of a gastrinoma. No correlation has been observed between the severity of symptoms and the extent of plasma gastrin elevation. However, the fasting plasma gastrin concentration at presentation in sporadic gastrinomas is related to the extent of tumor burden and the presence of hepatic metastases; it is therefore of prognostic value.

Measurement of Plasma Gastrin

In plasma from healthy subjects, the predominant forms of gastrin are amidated G-34 and G-17. In patients with gastrinomas, the gastrins found in the circulation display unpredictable heterogeneity with a shift toward larger peptides. For the detection of gastrinomas, the assay should detect all secreted forms of gastrin to prevent false-negative findings. Gastrin is unstable in serum or plasma, and samples may lose up to 50% of their immunoreactivity over 48 hours at 2°C to 8°C because of the action of proteolytic enzymes. Blood samples should be collected in tubes containing an anticoagulant and a protease inhibitor (e.g., aprotinin) to prevent degradation. Samples should be processed rapidly and the plasma stored at −20°C until assayed.

INTESTINAL DISORDERS AND THEIR LABORATORY INVESTIGATION

This section includes discussion of celiac disease, disaccharidase deficiency, bacterial overgrowth, bile salt malabsorption, inflammatory bowel disease (IBD), and protein-losing enteropathy and the main laboratory investigations associated with the diagnosis and monitoring of these disorders.

Celiac Disease

Pathophysiology of Celiac Disease

Celiac disease occurs in genetically predisposed subjects as a consequence of an inappropriate T cell–mediated immune response to ingestion of gluten from wheat and to similar proteins in barley and rye. The external trigger to the development of celiac disease in genetically susceptible individuals is found in gluten, which is the complex group of proteins found in wheat and other grains.

The identification in 1997 of small bowel tissue transglutaminase as the autoantigen of celiac disease has led to a greater understanding of the pathogenesis of the condition. The tissue transglutaminases are a family of calcium-dependent enzymes released from cells during wound healing. They catalyze the cross-linking or deamidation of proteins leading to stabilization of the wound area. Deamidation of gliadin peptides enhances their binding to HLA DQ2/DQ8 and increases recognition of these peptides by gut-derived T cells from subjects with celiac disease. The characteristic enteropathy is then induced by the release of interferon-γ and other proinflammatory cytokines.

A 33–amino acid peptide of gluten appears to be the primary initiator of the inflammatory response. It is resistant to breakdown by all gastric, pancreatic, and brush-border membrane proteases, thus allowing it to reach the small intestine intact. The peptide can be detoxified by exposure to a bacterial prolyl endopeptidase, suggesting a therapeutic strategy for celiac disease.

Increased intestinal permeability in untreated celiac disease that is reversible on withdrawal of gluten from the diet has been recognized since the early 1980s. Evidence suggests that this increase in paracellular permeability may be mediated by increased expression of zonulin, a protein that opens small intestinal tight junctions, or by decreased expression of intercellular epithelial cell adhesion molecules such as Z0-1, catenin, and cadherin. The zonulin pathway is now thought to play a significant role in the entry of allergens into cells and hence in the autoimmune response.

Clinical Considerations

Celiac disease is a common disorder in Caucasian populations with a prevalence of approximately 1%, but it also occurs

in northern Indian and North African populations. It is rare among people of Chinese, Japanese, and African Caribbean origin.

A wide spectrum of clinical manifestations of celiac disease is seen, with most diagnoses being made in adult life. Classic celiac disease—manifesting in infancy with failure to thrive, abdominal distention, and diarrhea—is now uncommon in developed countries. Most adults present with nonspecific symptoms; the condition may be picked up during the investigation of patients suspected of having irritable bowel syndrome (IBS). Mild iron deficiency is common. The initial presentation may be to a wide range of clinical specialties (Table 38.2). A strong association with other autoimmune diseases, especially type 1 diabetes and autoimmune thyroid disease, has been reported. In type 1 diabetes, the prevalence of celiac disease is approximately 5% and serologic screening to detect these cases has been advocated.

Tests for Celiac Disease

Appropriately standardized serologic tests have high clinical sensitivity and specificity for diagnosis and monitoring of compliance with a gluten-free diet. Table 38.3 compares the sensitivity and specificity of the serologic tests commonly used for the diagnosis of celiac disease.

Historically serologic testing for celiac disease involved testing for antigliadin antibodies and antiendomysial antibodies by indirect immunofluorescence. The sensitivity and specificity of these tests was poor; they have largely been superseded by measurement of antibodies against tissue transglutaminase (TGA) and deaminated gliadin peptide (DGP). Both IgG and IgA tests are available for these antibodies as quantitative enzyme-linked immunosorbent assays (ELISAs). The diagnostic accuracy of DGP-IgA is similar to that of TGA-IgA, with sensitivities of 85% to 90% and a specificity of approximately 95%. Sensitivity for DGP-IgG is not as good as that for DGP-IgA but has a specificity approaching 99%. Subjects with selective IgA deficiency (plasma IgA concentration <0.05 g/L, incidence ~1 in 600) are at greater risk for celiac disease. The tests directed at the IgA antibodies to TGA and DGP may produce false-negative results in patients with IgA deficiency; total IgA measurement should be considered in requesting serologic tests for celiac disease. When IgA deficiency is identified, the IgG-based tests should be used to confirm or exclude a diagnosis of celiac disease. No current recommendations suggest that DGP-based tests should replace tests based on TGA antibodies, but their use is increasingly supported in both adult and pediatric populations. The response to gluten withdrawal can be assessed clinically, by serologic tests, or by endoscopic or wireless capsule techniques. Endoscopic biopsies are considered the gold standard but are invasive and expensive, and they are now less frequently used for this purpose.

Strict adherence to a gluten-free diet leads to mucosal healing in celiac disease and reduces the risk for bowel malignancy (small bowel lymphoma). TGA can be used as a marker for monitoring dietary adherence in addition to its diagnostic role. Failure of symptoms to respond to a prescribed gluten-free diet may indicate nonadherence to the diet, other coexisting conditions such as small bowel overgrowth, **lactose intolerance**, microscopic colitis, or the presence of refractory celiac disease. This last is characterized by persistent villous atrophy with an increase in intraepithelial lymphocytes in the small bowel while the patient is on a strict long-term gluten-free diet. In both responsive and refractory celiac disease, antibody titers are typically decreased with dietary therapy and remain within reference intervals unless individuals are reexposed to gluten.

TABLE 38.2 Sample of Clinical Specialties to Which a Child or Adult With Celiac Disease May Present

Clinical Specialty	Symptoms/Manifestations
General medicine	Tired all the time
Gastroenterology	Diarrhea, flatulence, weight loss
Hematology	Anemia
Obstetrics/gynecology	Infertility
Orthopedics	Fracture, osteopenia
Dermatology	Dermatitis herpetiformis, hyperkeratosis
Neurology	Peripheral neuropathy
Rheumatology	Arthropathy
Endocrinology	Short stature, thyroid disease
Diabetes	Diarrhea, anemia

Disaccharidase Deficiencies

The presence of the brush-border disaccharidases is essential for carbohydrate absorption; a reduction in their activity leads to carbohydrate malabsorption and intolerance. Carbohydrate malabsorption can result in osmotic diarrhea with abdominal pain and distention and flatulence resembling the symptoms of IBS. Inherited disorders of sucrase-isomaltase and of transport proteins causing glucose-galactose malabsorption and fructose malabsorption have been described but are rare. Adult lactase deficiency is common in humans; in many Asian populations it is the normal state.

TABLE 38.3 Comparison of Serologic Tests for Celiac Disease

Antibody	Method	Sensitivity (%)	Specificity (%)
IgA-endomysial antibody	Immunofluorescence on monkey esophagus or human umbilical cord	80–100	>99
IgA-antigliadin antibody	Quantitative ELISA	75–95	95
IgA-antitissue transglutaminase antibody	Quantitative ELISA	>90	>99
IgA-deamidated gliadin peptide antibody	Quantitative ELISA	90	90

ELISA, Enzyme-linked immunosorbent assay; *IgA*, immunoglobulin A.

TABLE 38.4 Prevalence of Hypolactasia in Adults

Racial Group	Prevalence (%)
Chinese	>90
African Americans	54–81
Asians	60–90
Greeks	60–78
North Europeans	5–30

BOX 38.2 Methods for Detecting Lactase Deficiency

- Lactase in mucosal biopsy
- Oral lactose tolerance
  - Measure increase in plasma glucose or galactose
  - Measure increase in breath H_2 or $^{13}CO_2$
  - Measure urinary excretion of lactose after oral load

Lactase Deficiency

Lactase activity has its highest prevalence in the earliest months of life and declines after weaning. The prevalence of lactase deficiency varies widely with ethnic origin; although only 5% to 15% of northern Europeans are lactase-deficient, hypolactasia or alactasia can be found in more than 70% of individuals of African and Asian origin (Table 38.4). Congenital lactase deficiency is a very rare disorder that becomes apparent once the infant has been exposed to milk feeds. Stools have a low pH and contain large amounts of lactose and glucose, the latter being produced by bacterial degradation of undigested lactose. Secondary lactase deficiency is common in celiac disease owing to villous atrophy, but it may also occur in tropical sprue, small bowel **Crohn disease** (CD), radiation enteritis, giardiasis and other bowel infections, chronic alcoholism with malnutrition, and the enteropathy associated with acquired immunodeficiency syndrome (AIDS).

Tests for Lactase Deficiency

Many methods have been proposed for the diagnosis of lactase deficiency (Box 38.2). Disaccharidase activities can be measured in jejunal biopsies, although this is rare in routine practice. The xylose absorption test was first introduced as a test for carbohydrate malabsorption in the 1930s. In humans, D-xylose, a pentose sugar of plant origin, is absorbed by a passive mechanism in the jejunum. After an oral dose of D-xylose, it can be measured in blood or urine samples. It is now accepted that the amount of xylose excreted is affected not just by reduced absorption in the gut but also by renal, hepatic, and cardiac disease, and the test has largely been superseded by the hydrogen breath test. Oral lactose tolerance tests, measuring the increase in plasma glucose after ingestion of lactose, have been used to diagnose lactase deficiency. The usual dose of lactose is 50 g, but lower doses should be used in children (2 g/kg up to a maximum of 50 g). Blood samples are taken over a 2-hour period and the peak increment in glucose is recorded; a rise greater than 1.1 mmol/L (20 mg/dL) for capillary samples or greater than 1.4 mmol/L (>25 mg/dL) for venous plasma excludes lactase deficiency.

Hydrogen Breath Tests

Hydrogen is not a normal end product of mammalian metabolism, and hydrogen in breath is derived from bacterial metabolism in the intestinal tract. After lactose ingestion, the disaccharide is normally broken down into its constituent monosaccharides, which are absorbed. In the presence of lactase deficiency, unabsorbed disaccharide passes into the large bowel, where bacteria act, producing hydrogen that is absorbed into the systemic circulation and subsequently exhaled. Breath hydrogen can be measured by a variety of techniques based on the electrochemical detection of hydrogen; both laboratory and handheld point-of-care devices are available. Fasting breath hydrogen is normally less than 5 ppm (5 μL/L), and concentrations greater than 20 ppm (20 μL/L) at 2 hours after oral administration of 50 g lactose in adults (2 g/kg up to a maximum of 50 g in children) are suggestive of malabsorption or bacterial overgrowth. A lower cutoff of 10 ppm (10 μL/L) has been suggested that can improve diagnostic sensitivity without altering specificity. In some individuals (estimated at 2% to 13%), the large bowel bacteria do not produce hydrogen and fasting breath hydrogen will be low; in such individuals a normal result cannot exclude lactase deficiency. A positive breath hydrogen result after ingestion of lactose also may occur in glucose-galactose malabsorption; this can be confirmed or excluded by repeating the test substituting 25 g each of glucose and galactose for the lactose. An increase in breath hydrogen confirms the diagnosis. Lactose intolerance or malabsorption has become a very topical issue with clinicians as well as dieticians and nutritionists. Many of these have abandoned any measurements to confirm malabsorption-maldigestion and simply rely on exacerbation of symptoms with the ingestion of a pint of milk. Alternatively, the diagnosis is based on symptomatic improvement when patients are placed on a dairy-free diet.

Bacterial Overgrowth

The proximal small intestine (duodenum and jejunum) normally contains few bacteria. Most ingested bacteria do not survive the acidic environment of the stomach; therefore few live organisms enter the small bowel. The motility of the jejunum prevents fecal-type organisms from progressing up into the jejunum from the cecum. The ileum normally contains some fecal-type bacteria. Colonization of the upper small bowel is described as bacterial overgrowth and usually occurs as a consequence of other abnormalities (structural or motility disorders) of the small intestine. Gastric hypochlorhydria (associated with the use of PPIs) may contribute to bacterial overgrowth. Abnormalities in the systemic immune system (e.g., isolated IgA deficiency, hypogammaglobulinemia, combined immunodeficiency, infection with the human immunodeficiency virus [HIV]) can result in bacterial overgrowth. Impaired motility, pancreatic insufficiency, intrinsic small bowel disease, or surgically induced blind loops are often associated with bacterial overgrowth.

Small bowel bacterial overgrowth can be asymptomatic or associated with nonspecific symptoms that include abdominal distention, flatulence, and diarrhea resembling IBS. Bacteria colonizing the small bowel (e.g., *Escherichia coli* and *Bacteroides* species) deconjugate and dehydroxylate bile salts, resulting in conjugated bile salt deficiency and mild fat malabsorption. Bacterial metabolism of vitamin B_{12} can occur, leading to deficiency, whereas plasma folate concentration may be increased owing to production by the bacteria.

The gold standard diagnostic procedure is microbiologic examination of small bowel contents, but this is seldom practical. Noninvasive procedures based on oral administration of substances metabolized by bacteria to yield products that can be detected in breath are now the most common tests for bacterial overgrowth, but there is controversy as to which test to use and their sensitivity/specificity.

The ^{14}C-glycocholate breath test was one of the first tests to be used for noninvasive diagnosis of bacterial overgrowth, but it has been superseded by tests that avoid the use of radioactive isotopes. Small bowel bacterial breath tests in routine use involve test substances labeled with ^{13}C and the breath hydrogen tests. Lactulose is not normally absorbed from the small bowel and is therefore available for metabolism by bacteria throughout the gut; alternatively, labeled glucose can be used. In a normal subject, breath hydrogen does not increase until the substrate reaches the large intestine; the time from ingestion to a rise in breath hydrogen is an indicator of small bowel transit time. In bacterial overgrowth, an early rise in breath hydrogen of at least 20 ppm (20 μL/L) is observed within 30 minutes of ingestion. The early increase is diagnostic when it can be distinguished clearly from the later colonic phase. Frequent measurements are essential (e.g., at 5-minute intervals) for the first 30 minutes. The finding of an increased breath hydrogen output has high specificity for bacterial overgrowth but poor sensitivity. Variations in gastric emptying rate and small bowel transit times limit the diagnostic accuracy of the breath tests.

Bile Acid Malabsorption

Bile acids are synthesized in the liver and pass into the lumen of the small bowel via the gallbladder. Bile acids are present in bile as taurine or glycine conjugates; because the pH of bile is slightly alkaline and contains significant amounts of sodium and potassium, most of the bile acids and their conjugates exist as salts (i.e., bile salts). In practice, the terms *bile acids* and *bile salts* are frequently used synonymously. Their major function is to act as surface-active agents, forming micelles and facilitating the digestion of triglycerides and the absorption of cholesterol and fat-soluble vitamins. Little reabsorption of bile acids occurs in the proximal small bowel, but normally more than 90% is reabsorbed in the terminal ileum. The bile acids return to the liver via the portal circulation and can be resecreted into bile (enterohepatic circulation). Less than 10% of secreted bile acid is lost in the feces (~0.2 to 0.6 g/24 h).

Bile acid malabsorption leading to chronic diarrhea occurs when ileal disease is present or after resection of the terminal ileum; it may also occur after cholecystectomy and is present in some patients with IBS. Malabsorption of bile salts produces diarrhea by two different mechanisms. When significant bile salt depletion occurs, the deficiency of intraluminal bile salts leads to fat malabsorption and **steatorrhea**. More commonly, malabsorption of bile salts in the ileum leads to increased concentrations of bile salts in the colon, where they alter water and electrolyte absorption. This leads to net secretion of water into the lumen and diarrhea.

The most widely used test for bile acid malabsorption involves oral administration of the synthetic radiolabeled bile acid selenohomocholyltaurine (SeHCAT). Whole-body gamma counting is used to determine the basal level at 1 hour and then repeated at 7 days when in normal subjects more than 15% of the dose is retained. Retention of less than 10% of SeHCAT at 7 days can predict a good response to bile acid sequestrants.

Two blood tests have been proposed as alternatives to SeHCAT to avoid administration of a radioactive substance: fibroblast growth factor-19 (FGF-19) and 7α-hydroxy-4-cholesten-3-one (C4). C4 is the intermediary step between cholesterol and taurocholic acid and reflects the activity of hepatic cholesterol 7α-hydroxylase and therefore, because this is the rate-limiting enzyme, the rate of bile acid synthesis. Bile acid malabsorption is associated with increased plasma concentrations of C4 as hepatic synthesis is upregulated to maintain the pool of circulating bile acids. The presence of bile acids in the ileum has been shown to stimulate the production of FGF-19. Good correlation has been demonstrated between both serum C4 and FGF-19 and SeHCAT.

With growing awareness of the role of bile salt malabsorption in chronic diarrhea in a proportion of patients with IBS or IBD and the therapeutic effectiveness of cholestyramine, measurement of C4 and/or FGF-19 may become routine.

POINTS TO REMEMBER

Malabsorption

- Carbohydrate malabsorption can occur in disaccharidase deficiency.
- Celiac disease causes malabsorption by reduction in the villous surface area.
- Fat malabsorption is associated with exocrine pancreatic insufficiency.
- Fecal elastase is a useful test for exocrine pancreatic insufficiency.

Inflammatory Bowel Disease and Irritable Bowel Syndrome

Intestinal symptoms including abdominal pain or discomfort with diarrhea or constipation are common to both IBD and IBS. IBD includes **ulcerative colitis** (UC) and CD. IBS is a functional bowel disorder for which there is no identifiable pathologic process or known cause. Although IBS may produce symptoms that are sufficient to interfere with daily life, it is seldom associated with serious morbidity. IBS may affect up to 10% to 20% of the adult population in the developed world, with a typical age at presentation of 20 to 30 years and

a female preponderance. The prevalence of UC is approximately 100 to 200 per 100,000 people and of CD approximately 50 to 100 per 100,000 people, with no significant difference between genders. There is, however, a difference in prevalence with ethnic origin, with IBD being more common among Caucasians than individuals of Asian or African origin. Up to 20% of cases of IBD manifest in childhood, although the most common age at presentation is between 15 and 30 years of age. Both UC and CD are chronic conditions that tend to follow a remitting and relapsing course. The prognosis of patients with CD tends to be worse than that of those with UC, although in the latter it is estimated that up to 10% of patients will require surgical intervention within 10 years of diagnosis.

The cause of IBD is not fully understood, but both genetic and environmental factors have been implicated. Genome-wide association studies have identified over 160 susceptibility loci/genes that are significantly associated with IBD. Studies in monozygotic twins, however, have shown a lower than expected concordance rate (15% to 19% for UC, 27% to 50% for CD) if IBD were simply an inherited disorder. It is thought that individuals with a genetic predisposition to IBD need to be exposed to a range of environmental triggers that affect the host's immunity. Tissue damage is a consequence of neutrophil activation and the production of proinflammatory cytokines/chemokines. The success of treatments based on immunosuppressive agents or tumor necrosis factor-alpha (TNF-α) inhibitors appears to support an immune basis for the development of IBD. Environmental factors implicated in triggering the development of active disease include a protective effect of breast-feeding and an increased risk in patients who have recurrent GI infections or exposure to drug therapy (antibiotics, oral contraceptives, NSAIDs) along with socioeconomic factors, stress, smoking, and diet.

The conventional diagnostic pathway includes a full blood count to detect/exclude anemia, markers of generalized inflammation (erythrocyte sedimentation rate [ESR] and C-reactive protein [CRP]), and serologic testing for celiac disease (see earlier section on tests for celiac disease). In the past 20 years, there have been significant advances in biomarkers that can aid in differentiating IBD from IBS and in monitoring disease activity or response to treatment. Most of these are fecal markers.

Calprotectin

Calprotectin is a member of the S100 family of zinc- and calcium-binding proteins and is a heterodimer of S100A8/9. First discovered in 1980, it accounts for approximately 60% of cytosolic protein in neutrophils. It is also present in monocytes and macrophages at lower concentrations than in neutrophils and may have antimicrobial properties. Any disruption to the mucosal architecture of the GI tract allows neutrophils that accumulate at sites of active inflammation to be released into the lumen and subsequently shed in feces. The fecal concentration of calprotectin has been shown to correlate well with the gold standard indium-111–labeled granulocyte test and is directly related to the extent of inflammation. Calprotectin appears to be evenly distributed in fecal samples and has reasonable stability at room temperature.

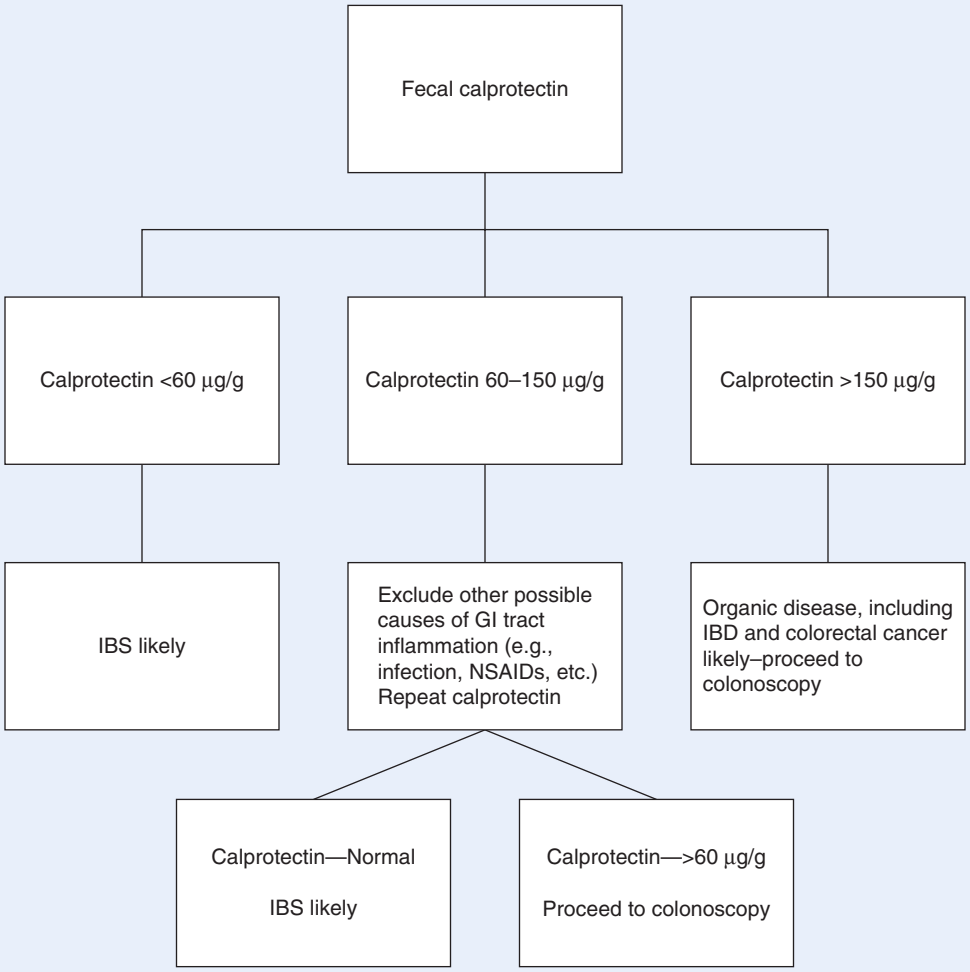
The initial studies of fecal calprotectin were on its ability to differentiate between IBS and IBD. In a study of 602 consecutive referrals from primary care to gastroenterology clinics at King's College Hospital (London, UK) patients had fecal calprotectin measured and were assessed for IBS using the Rome questionnaire. The sensitivity and specificity for calprotectin for the detection of organic disease were 89% and 79%, respectively. Combining calprotectin with the Rome questionnaire had a predictive value approaching 100%.

Several meta-analyses have investigated the value of fecal calprotectin for distinguishing IBD from IBS. The studies included in these meta-analyses may have been biased by preselection of patients in primary care before referral to secondary or tertiary care. A retrospective analysis of data from 946 patients from 48 primary care practices found a sensitivity of 82% and a specificity of 77%. Nevertheless, the (UK) National Institute for Health and Care Excellence (NICE) has recommended the widespread use of calprotectin in general practice to select patients at high risk for having IBD for referral to gastroenterologists, with the vast majority of the rest with IBS (normal calprotectin) remaining in primary care and treated appropriately. The cost savings of this approach are substantial.

Calprotectin has been found to differentiate between active and inactive IBD. The aim of treatment of IBD is currently to secure and maintain remission, with the ultimate aim of achieving mucosal healing. Assessment of treatment efficacy until recently has been based on clinical scores, including the Crohn Disease Activity Index (CDAI) and the Ulcerative Colitis Disease Activity Index (UCDAI), and by changes in ESR and CRP. However, many patients in clinical remission have increased fecal calprotectin concentrations, and it is perhaps these who would benefit from escalation of treatment to achieve mucosal healing. Several studies have demonstrated that a fall in the fecal calprotectin concentration indicates a good response to treatment and that this can precede any clinical response. It has been recommended that fecal calprotectin should be included in the range of tests before consideration of initiating or changing IBD therapy. Various studies have provided data on the use of fecal calprotectin in predicting relapse or recurrence of IBD, and these have been reviewed. Treating the inflammatory component of IBD regardless of symptoms promises to alter the natural history of the disease in a way similar to that seen in rheumatoid arthritis.

Fecal calprotectin is not, however, specific to IBD and is increased in patients with colorectal carcinoma, the chronic use of NSAIDs, bacterial infections, and diverticular disease. Other fecal markers that are undergoing active research include M2-PK, lactoferrin, S100A12, and cationic eosinophilic protein.

ALGORITHM FOR THE USE OF FECAL CALPROTECTIN IN DIFFERENTIATING IRRITABLE BOWEL SYNDROME AND INFLAMMATORY BOWEL DISEASE



Therapeutic Drug Monitoring in Patients With Inflammatory Bowel Disease

The increasing use of thiopurine analogs (azathioprine [AZA] and mercaptopurine [MP]) and the introduction of the anti–tissue necrosis factor (anti-TNF) drugs infliximab (IFX) and adalimumab (ADA) have led to the introduction of the concept of therapeutic drug monitoring (TDM) to gastroenterology. The metabolism of AZA (a prodrug) is complex, involving a number of enzymatic pathways that result in active, inactive, and potentially toxic metabolites. Measurement of red cell thiopurine methyltransferase (TPMT) activity has become a prerequisite before starting patients on AZA therapy because the presence of the low metabolic activity phenotype is a relative contraindication to the use of thiopurine analogs, and those who are heterozygous require a reduced dose. Measurement of the plasma concentration of 6-thioguanine (6-TG) can be used to assess adherence or the adequacy of dosage.

The anti-TNF drugs, particularly IFX and ADA, have proved effective in inducing and maintaining remission in patients with IBD. However, because these drugs are themselves humanized antibodies against TNF, many patients develop antibodies against them. Simultaneous measurement of the plasma drug concentrations and the antibodies against these drugs has meant that a failure to sustain a beneficial response to anti-TNF therapy can be correctly ascribed to either subtherapeutic drug concentrations or the presence of antidrug antibodies and alternative therapies can be devised and initiated.

Assessment of Functional Small Bowel Length

The amino acid citrulline, which is involved in the urea cycle, is predominantly synthesized in the enterocytes of the small intestine. The short bowel syndrome can be present after major surgical resection of the gut in the treatment of Crohn disease or necrosis. The remaining small bowel length can be estimated at surgery, but this is unreliable at lengths over 50 to 80 cm. Plasma citrulline concentrations have been shown to correlate well with small bowel length and absorption of xylose ($P < .001$), but not with CRP, CDAI, or ESR. Plasma citrulline measurements therefore appear to be a noninvasive

means of assessing small bowel absorptive capacity that is not influenced by intestinal inflammation.

Protein-Losing Enteropathy

Loss of significant quantities of plasma proteins into the bowel lumen and their subsequent excretion in feces is associated with a range of GI disorders. These include IBD, diseases in which the intestinal lymphatics are obstructed (e.g., lymphoma, Whipple disease), and disorders of immune status such as systemic lupus erythematosus and some food allergies.

In the healthy bowel, fecal protein is largely derived from enterocytes shed from the mucosal surface and from intestinal secretions. The normal GI loss of albumin is less than 10% of albumin catabolism, representing a daily loss of less than 1% to 2% of the circulating protein pool. In protein-losing enteropathy, this may increase to 40%, resulting in hypoalbuminemia and edema.

The diagnosis of protein-losing enteropathy should be considered in the investigation of patients with hypoalbuminemia in whom renal loss, liver disease, and malnutrition have been excluded. Methods using the administration of radioactive albumin or dextran are the gold standards but are seldom used now.

Fecal excretion of α_1-antiprotease inhibitor (AAT) can be used as a marker of GI protein loss. AAT is a 54-kDa glycoprotein (formerly called α_1-antitrypsin) present in plasma at a concentration of approximately 1.0 to 2.0 g/L that is resistant to degradation in the GI tract. Increased AAT excretion can be found in both small and large bowel disease both in adults and children. Interpretation must be made with knowledge of the plasma AAT concentration because the test is invalid in the presence of low plasma AAT concentrations (e.g., in AAT deficiency or impaired hepatic synthesis).

POINTS TO REMEMBER

Inflammatory Bowel Disease

- Crohn disease and ulcerative colitis are the most common IBDs.
- Fecal calprotectin can help distinguish IBD from IBS.
- Fecal calprotectin is related to disease activity and can predict relapse in IBD.
- Therapeutic drug monitoring is valuable in treating IBD with anti–TNF-α agents

PANCREAS: DISEASE AND ASSESSMENT OF EXOCRINE PANCREATIC FUNCTION

The pancreas plays a central role in the absorptive process for carbohydrates, fats, and proteins. Disorders of the exocrine pancreas therefore are frequently associated with symptoms of malabsorption such as diarrhea or steatorrhea. In this section, pediatric and adult exocrine pancreatic disorders are briefly discussed and tests that can be used to assess exocrine pancreatic function are described. Information on neuroendocrine tumors of the pancreas is also provided. Textbooks on gastroenterology or medicine provide much greater detail on the clinical aspects of exocrine pancreatic disorder.

Pediatric Disorders of the Exocrine Pancreas

Pancreatic disorders of childhood have been reviewed and are summarized in Box 38.3.

Cystic fibrosis (CF) is the most common severe autosomal recessive disease, with an estimated gene frequency in Western Europe and the United States of between 1:25 and 1:35 and a disease incidence of approximately 1 in 2500 to 1 in 3200. The pathogenesis and diagnosis of CF are described in Chapter 24. Pancreatic insufficiency is present in the neonatal period in 65% of infants with CF, and a further 15% develop it during infancy and early childhood. The 20% who do not develop pancreatic insufficiency have a better prognosis and develop fewer complications.

Measurement of pancreatic elastase-1 in feces (see the following section on noninvasive tests of exocrine function) is considered to be a reliable test for pancreatic insufficiency in infants over the age of 2 weeks with CF and in older children at the time of diagnosis. This test can also be used to detect the onset of pancreatic insufficiency in those previously sufficient.

Recurrent bouts of acute pancreatitis in childhood should arouse a suspicion of a hereditary cause. Two gene defects have been shown to be associated with pancreatitis manifesting in childhood: the cationic trypsinogen gene *(PRSS1)* and the serine protease inhibitor Kazal type 1 *(SPINK1)* gene. The majority of patients with *PRSS1*-related hereditary pancreatitis have a pathogenic mutation in exons 2 or 3 of the *PRSS1* gene. Duplication or triplication of the *PRSS1* gene has been reported

BOX 38.3 Spectrum of Pancreatic Disease in Childhood

Disorders of Morphogenesis

- Annular pancreas, pancreas divisum, pancreatic hypoplasia and agenesis, heterotopic pancreas

Inherited Syndromes Affecting the Pancreas

- Cystic fibrosis
- Schwachman-Diamond syndrome, Johnson-Blizzard syndrome, Pearson bone marrow pancreas syndrome

Gene Mutations Leading to Pancreatic Disease

- Hereditary pancreatitis, cationic trypsinogen gene mutations, trypsin inhibitor gene mutations

Pancreatic Insufficiency Syndrome

- Isolated enzyme deficiencies: lipase, colipase, enterokinase
- Pancreatic insufficiency secondary to other disorders
- Celiac disease

Acquired Pancreatitis in Childhood

- Idiopathic, traumatic, drugs, viral, metabolic, collagen vascular disease, autoimmune, fibrosing, nutritional (trophic)

to cause hereditary pancreatitis. The mechanism is assumed to be inappropriate activation of trypsin with consequent proteolytic action causing cell damage. Approximately 22% of patients with idiopathic pancreatitis have been reported to have at least one copy of the c.101A>G (p.Asn34Ser) mutation in the *SPINK1* gene, compared with approximately 2.5% of the general population. This protease inhibitor is believed to be the principal inhibitor of trypsin and chymotrypsin within the pancreas, and the mutation results in a loss of its inhibitory capacity.

Adult Disorders of the Exocrine Pancreas

The major exocrine pancreatic disorders manifesting in adult life are acute pancreatitis, chronic pancreatitis, and carcinoma of the pancreas. The use of enzyme tests in the diagnosis of acute pancreatitis is discussed in Chapter 19. The causes of pancreatitis are given in Box 38.4.

Chronic pancreatitis is defined by irreversible pancreatic damage, with histologic evidence of inflammation and fibrosis leading to destruction of cells and loss of endocrine and exocrine function. In developed nations, the most common cause is alcohol (60% to 90% of all cases of chronic pancreatitis), although because only 5% to 15% of heavy drinkers develop the disease, other predisposing factors must be present (e.g., smoking, diet high in fat and protein).

BOX 38.4 Causes of Pancreatitis in Adults

Acute
- Gallstones
- Alcohol
- Infection (e.g., mumps, Coxsackie B)
- Pancreatic tumors
- Drugs (e.g., azathioprine, estrogens, corticosteroids, didanosine)
- Iatrogenic (e.g., postsurgical, endoscopic retrograde cholangiopancreatography)
- Hyperlipidemias
- Miscellaneous (e.g., trauma, scorpion bite, cardiac surgery)
- Idiopathic

Chronic
- Alcohol
- Tropical (nutritional)
- Hereditary (trypsinogen and inhibitory gene defects, cystic fibrosis)
- Idiopathic
- Trauma
- Hypercalcemia

From Burroughs, A. K., & Westaby, D. (2009). Liver, biliary tract disease and pancreatic disease. In P. Kumar, & M. Clark (Eds.), *Clinical medicine* (7th ed., pp. 319–385). Edinburgh: Saunders.

Tests of Exocrine Function of the Pancreas

The predominant exocrine function of the pancreas is the production and secretion of pancreatic juice, which is rich in enzymes and bicarbonate. Normal pancreatic juice is colorless and odorless; it has a pH of 8.0 to 8.3 and a specific gravity of 1.007 to 1.042. The total 24-hour volume secreted may be as high as 3000 mL.

Laboratory tests used to assess exocrine pancreatic function can be divided into invasive and noninvasive categories. Invasive tests require GI tract intubation to allow the collection of pancreatic secretions. Noninvasive tests were developed to avoid the need for intubation, which is uncomfortable for the patient, time-consuming, and therefore expensive. Noninvasive ("tubeless") tests are simpler to perform but in general lack the sensitivity and specificity of invasive tests, particularly for the diagnosis of mild pancreatic insufficiency.

Invasive Tests of Exocrine Pancreatic Function

In these tests the total volume of pancreatic juice, the concentration of bicarbonate, and activities of the pancreatic enzymes amylase, lipase, and trypsin are measured in duodenal fluid.

The secretin-cholecystokinin test is based on the principle that secretion of pancreatic juice and bicarbonate output are related to the functional mass of the pancreas. After an overnight fast, a tube is sited in the duodenum to permit the collection of pancreatic juices. Secretin (1 U/kg body weight) is given intravenously, and the duodenal fluid is collected at 15-minute intervals for at least 1 hour. Secretin stimulates the secretion of pancreatic juice and bicarbonate. CCK (or the synthetic equivalent, ceruletide, the C-terminal octapeptide sequence of the intact hormone, where the functional activity resides, Fig. 38.4) can then be given to stimulate the secretion of pancreatic enzymes, allowing a more complete assessment of pancreatic reserve than can be obtained with secretin alone. The secretin-cholecystokinin and secretin-ceruletide tests are considered the gold standard for assessing pancreatic exocrine function. Although seldom used in adults, this test is still used in infants with pancreatic insufficiency to distinguish between CF and Schwachman-Diamond syndrome (pancreatic α-cell aplasia). In the former, mucous plugs reduce pancreatic secretion, but the enzyme concentrations increase in a normal physiologic response to cholecystokinin, whereas in the latter enzyme secretion is absent or minimal.

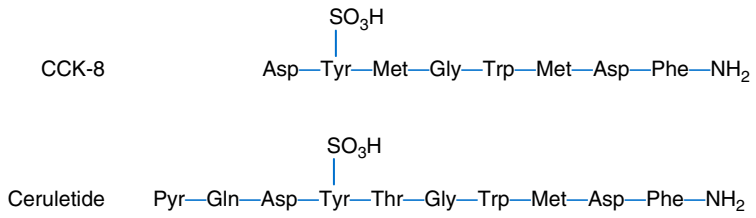


Fig. 38.4 Comparison of the amino acid sequences of cholecystokinin *(CCK)*-8 and ceruletide.

Noninvasive Tests of Exocrine Pancreatic Function

A range of tubeless tests has been proposed, but none has adequate sensitivity for reliably detecting early pancreatic disease. When malabsorption is present, such tests are of value in confirming or excluding pancreatic disease. Considerable overlap often occurs between results observed in normal individuals and those found in patients with pancreatic disease; this is due mainly to the large functional reserve of the pancreas. It has been estimated that pancreatic insufficiency cannot clearly be demonstrated until at least 50% of the acinar cells have been destroyed. Clinical signs of pancreatic insufficiency often do not appear until 90% of acinar tissue has been destroyed.

The noninvasive tests are based on the reduction in secretion of pancreatic enzymes with measurement of the enzymes in feces (chymotrypsin or elastase) or detection of products of their catalytic reactions after oral administration of synthetic substrates in urine (*N*-benzoyl-L-tyrosine-*p*-aminobenzoic acid [NBT-PABA] or pancreolauryl test) or in breath (^{13}C mixed-chain triglyceride breath test). The NBT-PABA and pancreolauryl substrates are no longer commercially available.

Pancreatic chymotrypsin is almost completely digested in its passage through the gut in adults, but the residual activity of the enzyme in feces is stable for several days at room temperature. Its output in stool correlates poorly with chymotrypsin secretion in duodenal contents when both are measured after stimulation with secretin-CCK. In patients without pancreatic disease, the incidence of falsely low results is approximately 10% to 15% and may be due to voluminous stools (>300 g/24 h), thus less enzyme per gram of feces; inadequate food intake; partial gastrectomy or mucosal disease (e.g., celiac disease), which causes inadequate stimulation of pancreatic secretion; or obstruction of the common bile duct. Falsely normal results in patients with mild pancreatic insufficiency may be as high as 50%. In a collaborative study in children, both fecal chymotrypsin and elastase showed 100% sensitivity for detecting pancreatic insufficiency in CF; but the specificity of chymotrypsin was lower than that of elastase in a control group of children with small intestinal disease.

Pancreatic elastase-1 is a pancreas-specific protease present in pancreatic juice. It is not degraded during passage through the gut, and concentrations in feces are five- to sixfold greater than those in pancreatic juice. Elastase-1 can be measured by ELISA with two monoclonal antibodies specific to the human enzyme. Treatment of patients with pancreatic enzyme supplements therefore does not interfere with the test. Fecal elastase-1 has been evaluated extensively in both CF and adult pancreatic insufficiency, and its use is recommended in both groups. Fecal elastase is often undetectable (<15 μg/g) in children with CF, and values below 200 μg/g after 4 weeks of age are indicative of pancreatic insufficiency.

Using random fecal samples, specificities of 98% and 100% have been reported in healthy controls and specificities of 90% to 97% in patients with nonpancreatic GI disease. Positive results (i.e., <200 μg/g) have been reported in patients with clinical or laboratory evidence of malnutrition who also have IBD or chronic diarrhea (nonpancreatic).

Measurement of pancreatic elastase-1 in feces has high sensitivity for the detection of severe and moderate chronic pancreatitis in adults. It has better sensitivity than other tests for detecting mild chronic pancreatitis and high sensitivity and high negative predictive value for discriminating between diarrhea of pancreatic and nonpancreatic origin. The test is not specific for pancreatitis and detects moderate to severe impairment of pancreatic exocrine secretion from any cause. It is considered to be the most suitable test to confirm pancreatic insufficiency in CF infants older than 2 weeks. A negative test does not exclude mild disease and false-positive results in some nonpancreatic diseases and in very watery samples limit its diagnostic accuracy.

Gastrointestinal Regulatory Peptides

The gut is the largest endocrine organ in the body but is also a major target for many hormones released locally or from other organs. GI regulatory peptides are released from the pancreatic islets (e.g., somatostatin) or from endocrine cells spread throughout the gut mucosa and collectively known as the diffuse endocrine system. Many of these peptides (such as vasoactive inhibitory peptide [VIP] and somatostatin) are present in the enteric nerves. They are also found in the central nervous system (CNS) and have important roles in the neuroendocrine control of the gut. Although many of them fulfill the classic criteria for a hormone by acting on distant cells, others function as neurotransmitters or have local (paracrine) effects on adjacent cells. Collectively they influence motility, secretion, digestion, and absorption in the gut. They regulate bile flow and the secretion of pancreatic hormones and affect the tonicity of the vascular walls, blood pressure, and cardiac output.

Table 38.5 summarizes the basic chemical characteristics of four of the major GI regulatory peptides and indicates their site of origin and main functions. More detailed descriptions of these peptides are given in the following paragraphs. A list of other regulatory peptides of the GI tract is presented in Table 38.6.

Gastrointestinal Neuroendocrine Tumors and Tumor Markers

GI neuroendocrine tumors may be endocrine pancreatic tumors or carcinoid tumors arising from enterochromaffin cells that occur throughout the GI tract. Carcinoid tumors are described in Chapter 26. Approximately two-thirds of patients with tumors arising from pancreatic islet cells present with excessive hormone production. This group of tumors includes insulinomas, gastrinomas, VIPomas, glucagonomas, and somatostatinomas. Insulinomas and glucagonomas are not usually associated with GI symptoms. Gastrinomas were discussed earlier. The somatostatinoma syndrome is associated with steatorrhea, gallstones, and hyperglycemia.

The watery diarrhea hypokalemic achlorhydria syndrome (WDHA syndrome, Werner-Morrison syndrome) is a consequence of a tumor producing excessive amounts of VIP. The WDHA syndrome may be suspected in a patient who

TABLE 38.5 Characteristics of Prominent Forms of Principal Gut Regulatory Peptides

Hormone/ Peptide	Molecular Weight (Da)	No. of Amino Acids	Main Gut Localization	Principal Physiologic Action
Gastrin Family				
Cholecystokinin	3918	33	Duodenum and jejunum	Stimulates gallbladder contraction and intestinal motility; stimulates secretion of pancreatic enzymes, insulin, glucagon, and pancreatic polypeptides; has a role in indicating satiety; the C-terminal 8–amino acid peptide (CCK-8) retains full activity
Gastrin-17	2098	17	Gastric antrum and duodenum	Stimulates the secretion of gastric acid pepsinogen, intrinsic factor, and secretin; stimulates intestinal mucosal growth; increase gastric and intestinal motility
Gastrin	3839	34		
Secretin-Glucagon Family				
Secretin	3056	27	Duodenum	Stimulates pancreatic and jejunal secretion of bicarbonate, enzymes, and insulin; reduces gastric and intestinal motility, inhibits gastrin release and gastric acid secretion
VIP	3326	28	Enteric nerves	Relaxes smooth muscle of gut, blood, and genitourinary system; increases water and electrolyte secretion from the pancreas and gut; releases hormones from the pancreas, gut, and hypothalamus
GIP	4976	42	Duodenum and jejunum	Stimulates insulin release; reduces gastric and intestinal motility; increases fluid and electrolyte secretion from the small intestine

GIP, Glucose-dependent insulinotropic peptide; *VIP*, vasoactive intestinal polypeptide.

TABLE 38.6 Brief Description of Other Gastrointestinal Regulatory Peptides

Hormone/Peptide	Major Tissue Location	Principal Known Actions
Bombesin	Throughout the gut and pancreas	Stimulates release of CCK and gastrin
Calcitonin gene–related peptide	Enteric nerves	Unclear
Chromogranin A	Neuroendocrine cells	Secretory protein
Enkephalins	Stomach, duodenum	Opiate-like actions
Enteroglucagon	Small intestine, pancreas	Inhibits insulin secretion
Ghrelin	Stomach	Stimulates appetite, increases gastric emptying
GLP-1	Pancreas, ileum	Increases insulin secretion
GLP-2	Ileum, colon	Enterocyte-specific growth hormone
Leptin	Stomach	Appetite control
Motilin	Throughout the gut	Increases gastric emptying and small bowel motility
Neuropeptide Y	Enteric nerves	Regulation of intestinal blood flow
Neurotensin	Ileum	Affects gut motility; increases jejunal and ileal fluid secretion
Pancreastatin	Pancreas	Inhibits pancreatic endocrine and exocrine secretion
Pancreatic polypeptide	Pancreas	Inhibits pancreatic and biliary secretion
PYY	Colon	Inhibits food intake
Somatostatin	Stomach, pancreas	Inhibits secretion and action of many hormones

CCK, Cholecystokinin; *GLP*, glucagon-like peptide; *PYY*, peptide tyrosine tyrosine.

produces large volumes (>1 L/24 h) of secretory diarrhea with dehydration and hypokalemia. The diagnosis is confirmed by the finding of high plasma VIP concentrations and demonstration of somatostatin-receptor uptake on imaging.

The remaining one-third of patients with neuroendocrine tumors have no specific clinical symptoms associated with the tumors, which are described as nonfunctional.

The pattern of hormonal and precursor production by neuroendocrine tumors is complex. Most secrete several tumor markers. Chromogranin A, a member of a family of secretory proteins, has a diagnostic sensitivity of more than 90% for neuroendocrine tumors. Plasma chromogranin A concentration is elevated in most tumors and is an alternative to more specific markers in monitoring the effectiveness of therapy. Although chromogranin A has high sensitivity, false-positive findings have been observed in several nonendocrine tumors, including prostate cancer.

Chromogranin A has been reported to be slightly increased in IBD and also in other malignancies, including lung, prostate, breast, and uterine. It is increased by the use of PPIs and hydrogen receptor blockers, which can be problematic because many patients suspected of having a neuroendocrine tumor may have started on these in primary care. Chromogranin A is renally excreted, so significant renal dysfunction increases the concentration. For these reasons, chromogranin A is usually considered to be a marker of prognosis and for monitoring response to therapy rather than a diagnostic tool.

TABLE 38.7 Summary of Disorders Leading to Malabsorption

Disorders of Intraluminal Digestion	
1. Altered gastric function	Postgastrectomy syndrome, Zollinger-Ellison syndrome
2. Pancreatic insufficiency	Chronic pancreatitis, cystic fibrosis, pancreatic cancer
3. Bile acid deficiency	Disease/resection of terminal ileum, small bowel overgrowth
Disorders of Transport Into the Mucosal Cell	
1. Generalized disorders due to reduction in absorptive surface area	Celiac disease, tropical sprue
2. Specific disorders	Hypolactasia, vitamin B_{12} in pernicious anemia, zinc in acrodermatitis enteropathica
Disorders of Transport Out of the Mucosal Cell	
1. Blockage of the lymphatics	Primary lymphangiectasia Abdominal lymphoma
2. Inherited disorders	Abetalipoproteinemia

Investigation of Maldigestion and Malabsorption

This section summarizes causes of malabsorption and suggests a general laboratory approach to these disorders. Table 38.7 summarizes the main causes of malabsorption under the three categories of intraluminal disorders and malabsorption secondary to disorders of either transport into the mucosal cells or transport out of the mucosal cells.

The clinical presentation of a patient with malabsorption or maldigestion classically includes the following features:

Evidence of general ill health. Anorexia, weight loss, fatigue after minor effort, and dyspnea may be seen. Edema (due to hypoalbuminemia), tetany (low plasma calcium concentration), and dehydration secondary to electrolyte imbalance and water loss may be present. In exocrine pancreatic insufficiency, however, hyperphagia is the rule: patients may often report a very high (5000 kcal/day, 21,000 kJ/day) food intake without weight gain.

Isolated nutrient deficiencies. Iron, folate, or vitamin B_{12} deficiency may manifest as anemia, which may be mild; vitamin K deficiency as a bleeding tendency; and vitamin D deficiency as bone disease (see Chapter 27 for more details on vitamins). They are reflected by a variety of signs and symptoms (glossitis, pallor, dermatitis, petechiae, bruising, hematuria, muscle or bone pain, or neurologic abnormalities).

Abdominal symptoms. These include discomfort, distention, flatulence, and borborygmi (rumbling and gurgling sounds resulting from movement of gas in the intestine).

Watery diarrhea and possible steatorrhea. In severe cases of steatorrhea the stool is typically loose, bulky, offensive, greasy, light-colored, and difficult to flush away. Alternatively, the stools may appear normal but may be more bulky or be passed more frequently.

Early manifestation of malabsorption will, however, be more subtle than this list might indicate. The alteration in volume or consistency of the stool may be slight, and only mild symptoms may be attributable to the GI tract. The patient may report only anorexia, fatigue, and lack of interest in daily activities. It is in these cases that the astute physician who suspects malabsorption on clinical grounds will rely on the laboratory to assist in the diagnosis. Initial laboratory investigations consist of routine tests, with abnormalities that may indicate the possibility of malabsorption (e.g., blood hemoglobin concentration, mean red cell volume, serum concentrations of folate, ferritin, calcium, albumin, liver enzyme activities, and tests for antibodies in celiac disease).

Evaluation of Fat Absorption

The evaluation of fat absorption or malabsorption is required in only a small minority of patients undergoing investigation for GI disorders, because a firm diagnosis can often be made on clinical grounds alone.

Historically the assessment of fat malabsorption was carried out by measurement of fecal fat excretion in a timed (2 to 3 days) fecal collection. This had many limitations and is seldom performed today. A range of alternative tests have been proposed, but most had similar limitations to those for fecal fat measurement in either their sensitivity and specificity or analytic constraints. Currently, investigation of fat malabsorption is limited to the measurement of fecal elastase and the ^{13}C mixed-chain triglyceride breath test to identify exocrine pancreatic insufficiency with consequent reduction in secretion of lipase.

The ^{13}C mixed-chain triglyceride breath test uses ^{13}C-labeled medium- and long-chain fatty acids as a means of assessing intraluminal pancreatic lipase activity. The labeled substrate is administered orally with a standard meal of toast and butter. Breath samples are collected over a 5-hour period and exhaled $^{13}CO_2$ is expressed as a percentage of the administered dose, with normal excretion in the range of 25% to 40%. The sensitivity and specificity of the test for identifying pancreatic insufficiency has been reported as 89% and 81%, respectively. Similar results have been reported for the ^{14}C-triolein test using a variety of fat loads and procedures. The disadvantage of this test is the need to measure the ^{14}C-labeled CO_2 immediately, which requires the availability of a scintillation counter and facilities for the administration of radioisotopes.

Investigation of Chronic Diarrhea

Although diarrhea is a common problem, no clear definition has existed to distinguish it from the range of stool weight, frequency, consistency, or volume that occurs in the normal population. A 2003 proposal that sought to encompass these different elements suggests that for a Western diet, diarrhea may be defined as "the abnormal passage of loose or liquid stools more than three times daily and/or a volume of stool greater than 200 g/day." Guidelines suggest that diarrhea may

BOX 38.5 Causes of Chronic Diarrhea

Colonic
- Ulcerative colitis
- Crohn disease
- Microscopic colitis

Small Bowel
- Celiac disease
- Small bowel enteropathies (e.g., Crohn disease, Whipple disease, tropical sprue, amyloidosis, intestinal lymphangiectasia)
- Bile salt malabsorption
- Disaccharidase deficiency
- Small bowel bacterial overgrowth
- Lymphoma
- Giardiasis (and other chronic infections)
- Irritable bowel syndrome

Endocrine
- Hyperthyroidism
- Diabetes
- Hypoparathyroidism
- Addison disease
- Neuroendocrine tumors (e.g., VIPoma, carcinoid tumors)

Other
- Chronic laxative abuse (e.g., bulimia nervosa)
- Surgical causes (e.g., small bowel resection, internal fistulas)
- Drugs and alcohol
- Autonomic neuropathy

Modified from Thomas, P. D., Forbes, A., Green, J., Howdle, P., Long, R., Playford, R., et al. (2003). Guidelines for the investigation of chronic diarrhoea. 2nd edition. *Gut, 52*(suppl V), 1–15.

be defined as chronic when it has continued for 4 weeks or more; such persistence indicates the likelihood of a noninfectious cause requiring further investigation.

Several quite different mechanisms can lead to diarrhea. In carbohydrate malabsorption, the presence of unabsorbed solutes in the bowel causes osmotic diarrhea as water enters the bowel from tissue. By contrast, the diarrhea of most cases of laxative abuse or VIPomas is due to active secretion of water and electrolytes into the bowel, which is described as secondary diarrhea. IBD causes diarrhea as a consequence of the inflammatory process.

Many diseases more commonly thought to cause "diarrhea" in fact lead to more frequent passage of stools but not usually to an increased stool weight (or volume). Such disorders (e.g., IBS) generally fall outside the scope of the definition of chronic diarrhea. Guidelines for the management of IBS are available, and the best results are achieved with an integrated approach to the problem, involving clinicians with a special interest, dietitians, a psychiatric-psychologic approach, etc.

Box 38.5 describes the many causes of chronic diarrhea; most are due to disease of the colon, in which laboratory diagnostic tests are of little value with the exception of fecal calprotectin for the differentiation of IBS from IBD.

BOX 38.6 Causes of the Acute Abdomen

Common Causes
- Acute appendicitis
- Acute cholecystitis, ascending cholangitis
- Small and large bowel obstruction
- Intussusception, volvulus, and strangulated hernias
- Renal/ureteric colic
- Perforated peptic/duodenal ulcer
- Acute pancreatitis
- Acute diverticulitis
- Gynecologic conditions (e.g., ectopic pregnancy, salpingitis)
- Ruptured aortic aneurysm
- Mesenteric lymphadenitis

Less Common Causes
- Gastroenteritis including *Salmonella, Shigella, Yersinia,* measles virus, etc.
- Crohn disease
- Pyelonephritis
- Meckel diverticulitis
- Acute intermittent porphyria

Conditions That May Simulate the Acute Abdomen
- Myocardial infarction or myocarditis

Laxative Abuse

Surreptitious laxative abuse is an important cause of diarrhea that is often overlooked. It may be as common as 20% of cases in secondary or tertiary referral centers. In Munchausen syndrome by proxy, adults have administered laxatives surreptitiously to young children. A clinical diagnosis rarely can be made; no single clinical feature reliably predicts a positive test. When there is a high index of clinical suspicion, urine screens for laxatives may be helpful.

The Acute Abdomen

The acute abdomen is defined as undiagnosed intense abdominal pain lasting less than 1 week. Many of the diseases that present in this way have a high mortality if not treated surgically, but others may have high morbidity and mortality if surgical intervention is carried out. The clinical diagnosis is based primarily on the nature of the pain. Common causes of the acute abdomen are detailed in Box 38.6. Nonabdominal conditions may simulate an acute abdomen. History and examination often yield a likely diagnosis, but laboratory investigations may help in establishing this (e.g., pregnancy test in suspected ectopic pregnancy, serum amylase, or lipase in acute pancreatitis). A full blood count may reveal a raised neutrophil count suggestive of infection or anemia secondary to blood loss from the GI tract. Knowledge of renal function and acid-base status is important if surgery is being considered.

Acute Pancreatitis

Acute pancreatitis is an inflammatory condition of the pancreas that may have a self-limiting short course but can be fatal. Almost all patients with acute pancreatitis have severe

epigastric pain, usually of sudden onset and often radiating to the back. In more severe cases, there is nausea and vomiting, fever, hypotension, shock, and multiorgan failure that may lead to death. Biochemical features of acute pancreatitis include uremia, hypoalbuminemia, hypokalemia, hyperglycemia, metabolic acidosis, hypoxemia, and abnormal liver tests. None of these are invariably present or diagnostic for pancreatitis. In acute necrotizing pancreatitis, methemalbuminemia may be detectable. The use of enzyme measurements in the diagnosis of acute pancreatitis is explained in Chapter 19.

AT A GLANCE

Causes of Acute Pancreatitis

- Gallstones
- Alcohol excess
- Drugs
- Hypertriglyceridemia
- Hypercalcemia
- Trauma
- Infections
- Rare causes including tumors

REVIEW QUESTIONS

1. Gastrin
   a. is secreted by the liver and stomach.
   b. is secreted when stomach pH is low.
   c. stimulates gastric acid secretion.
   d. inhibits secretion of intrinsic factor.
2. The hydrogen breath test using glucose or lactulose as a substrate assesses
   a. intestinal bacterial overgrowth.
   b. celiac disease.
   c. the presence of *H. pylori.*
   d. bile acid malabsorption.
3. A peptide neurotransmitter that relaxes smooth muscle in the gut and increases water and electrolyte secretion is
   a. cholecystokinin (CCK).
   b. secretin.
   c. vasoactive intestinal polypeptide (VIP).
   d. gastric inhibitory polypeptide.
4. The three phases of the digestive process include the neurogenic, gastric, and the intestinal phases. The intestinal phase is initiated
   a. by distention of the stomach.
   b. by the sight, smell, and taste of food.
   c. upon stimulation of the cerebral cortex in the brain.
   d. when weakly acidic digestive products of proteins and lipids enter the duodenum.
5. An example of an invasive test for detection of *H. pylori* in peptic ulcer disease would be
   a. breath testing following ingestion of ^{14}C-labeled urea.
   b. microbiologic culture of a gastric biopsy sample.
   c. lab measurement of a specific IgG antibody.
   d. detection of a specific fecal antigen.
6. Which one of the following is considered to be a disease of the intestine?
   a. Zollinger-Ellison syndrome
   b. Peptic ulcer disease
   c. Celiac disease
   d. Cystic fibrosis (CF)
7. In Western countries, the most common cause of chronic pancreatitis in adults is
   a. gallstones.
   b. infections.
   c. pancreatic tumors.
   d. alcohol consumption.
8. Which one of the following peptides stimulates increased pancreatic secretion of bicarbonate?
   a. Glucose-dependent insulinotropic peptide (GIP)
   b. Gluten
   c. Secretin
   d. Vasoactive intestinal polypeptide (VIP)
9. A tumor of the pancreatic islet cells that produces in part massive gastric hypersecretion, diarrhea, and steatorrhea causes
   a. the Zollinger-Ellison syndrome.
   b. celiac disease.
   c. lactose intolerance.
   d. cystic fibrosis (CF).

SUGGESTED READINGS

Lindsay, J., Langmead, L., & Preston, S. (2012). Gastrointestinal disease. In P. Kumar, & M. Clark (Eds.), *Clinical medicine* (8th ed). Edinburgh: Saunders.

Malfertheiner, P., Megraud, F., O'Morain, C., et al. (2007). Current concepts in the management of *Helicobacter pylori* infection: the Maastricht III Consensus Report. *Gut*, *56*, 772–781.

Barakat, M. T., Meeran, K., & Bloom, S. R. (2004). Neuroendocrine tumours. *Endocrine Related Cancer*, *11*, 1–18.

van Heel, D. A., & West, J. (2006). Recent advances in celiac disease. *Gut*, 1037–1046.

Armstrong, D., Don-Wauchope, A. C., & Verdu, E. F. (2011). Testing for gluten-related disorders in clinical practice: the role of serology in managing the spectrum of gluten sensitivity. *Canadian Journal of Gastroenterology*, *5*, 193–197.

Kumar, P. J., & Clark, M. L. (2002). Malabsorption and weight loss. In S. Bloom (Ed.), *Practical gastroenterology* (pp. 371–382). London: Martin Dunitz.

Smith, M. J., Cherian, P., Raju, G. S., et al. (2000). Bile acid malabsorption in persistent diarrhoea. *Journal of the Royal College of Physicians London*, *34*, 448–451.

Thomas, P. D., Forbes, A., Green, J., et al. (2003). Guidelines for the investigation of chronic diarrhoea. *Gut*, *52*(Suppl. V), v1–15.

Ponder, A., & Long, M. D. (2013). A clinical review of recent findings in the epidemiology of inflammatory bowel disease. *Clinical Epidemiology*, *5*, 237–247.

Walsham, N. E., & Sherwood, R. A. (2016). Fecal calprotectin in inflammatory bowel disease. *Clinical and Experimental Gastroenterology*, *9*, 1–9.

Sherwood, R. A. (2012). Faecal markers. *Journal of Clinical Pathology*, *65*, 981–985.

Yarur, A. J., Abreu, M. T., Deshpande, A. R., et al. (2014). Therapeutic drug monitoring in patients with inflammatory bowel disease. *World Journal of Gastroenterology*, *20*, 3475–3484.

Papadia, C., Sherwood, R. A., Kalantzis, C., et al. (2007). Plasma citrulline concentration: a reliable marker of small bowel absorptive capacity independent of intestinal inflammation. *American Journal Gastroenterology*, *102*, 1–9.

Leus, J., Van Biervliet, S., & Robberecht, E. (2000). Detection and follow up of exocrine pancreatic insufficiency in cystic fibrosis: a review. *European Journal of Pediatrics*, *159*, 563–568.

Masamune, A. (2014). Genetics of pancreatitis: the 2014 update. *Tohoku Journal Experimental Medicine*, *232*, 69–77.

Masoero, G., Zaffino, C., Laudi, C., et al. (2000). Fecal pancreatic elastase 1 in the work up of patients with chronic diarrhea. *International Journal of Pancreatology*, *28*, 175–179.

Ardill, J. E. S. (2008). Circulating markers for endocrine tumors of the gastrointestinal tract. *Annals of Clinical Biochemistry*, *45*, 539–559.

Thomas, P. D., Forbes, A., Green, J., et al. (2003). Guidelines for the investigation of chronic diarrhoea. *Gut*, *52*(Suppl. V), v1–15.

Duncan, A. (2000). Screening for surreptitious laxative abuse. *Annals of Clinical Biochemistry*, *37*, 1–8.

39

Bone and Mineral Metabolism

William D. Fraser

OBJECTIVES

1. Define the following:
   - Bone marker
   - Matrix
   - Osteoblast
   - Osteoclast
   - Osteomalacia
   - Osteopenia
   - Osteoporosis
   - Rickets
   - Type I collagen
2. Describe the structure and function of bone.
3. Detail the physiology and regulation of calcium and phosphate.
4. List causes, symptoms, and laboratory analyses used in the differential diagnosis of the following:
   - Hypercalcemia
   - Hypermagnesemia
   - Hyperphosphatemia
   - Hypocalcemia
   - Hypomagnesemia
   - Hypophosphatemia
5. List and describe commonly used methods in the measurement of the following analytes:
   - Bone alkaline phosphatase (Bone ALP)
   - Parathyroid-related protein
     Phosphate
   - Magnesium
     PTH
   - Total/free calcium
     Vitamin D and metabolites
6. State the effect each of the following disorders has on total and free calcium, phosphate, albumin, and PTH:
   - Primary and secondary hyperparathyroidism
   - Primary and secondary hypoparathyroidism
   - Pseudohypoparathyroidism
7. Discuss 1,25-dihydroxyvitamin D, including its metabolism, physiology, and roles in bone health, mineral metabolism, and disease.
8. List and describe five markers of bone formation and resorption.
9. For the following disorders, state the cause(s), symptoms, laboratory analyses, and lab results used to diagnose:
   - Osteoporosis
   - Paget disease of bone
   - Rickets

KEY WORDS AND DEFINITIONS

Adynamic bone disease (ABD) A type of renal osteodystrophy characterized by reduced osteoblasts and osteoclasts, and low bone turnover.

Bone alkaline phosphatase An isoenzyme of alkaline phosphatase and a biochemical marker of bone formation.

Calcitonin A 32-amino-acid polypeptide hormone elaborated by the parafollicular cells of the thyroid gland in response to hypercalcemia.

CTX An antigen produced when type I collagen is digested by the proteinase cathepsin K yielding crosslinked carboxy-terminal telopeptide of type I collagen; a plasma marker for bone resorption.

Deoxypyridinoline (DPD) A crosslink of type I collagen present in bone that is excreted, free or protein-bound, in serum or urine and serves as a marker of bone resorption.

Humoral hypercalcemia of malignancy (HHM) A malignancy caused by bone resorption mediated by circulating factors released from distant tumor cells.

Hypercalcemia Increased concentration of calcium in plasma; manifestations include fatigability, muscle weakness, depression, anorexia, nausea, and constipation; most commonly caused by primary hyperparathyroidism or malignancy.

Hypercalcemia-associated malignancy (HCM) Hypercalcemia caused by the presence of malignancy.

Hypermagnesemia A condition characterized by abnormally high concentrations of magnesium in plasma.

Hyperphosphatemia A condition characterized by abnormally high concentrations of phosphates in plasma.

Hypocalcemia A condition characterized by a low concentration of calcium in plasma.

Hypomagnesemia A condition characterized by a low concentration of magnesium in plasma.

Hypoparathyroidism The condition produced by greatly reduced function of the parathyroid glands.

Hypophosphatemia A condition characterized by a low concentration of phosphate in plasma.

Mineralization The process by which the body uses minerals to build bone structure.

N-telopeptide (NTX) A biochemical marker of bone resorption.

Osteitis fibrosa A complication of hyperparathyroidism in which the bones are soft and often deformed; also called osteitis fibrosa cystica.

Osteoblasts Cells responsible for formation of bone, including synthesis of type I collagen and noncollagenous proteins and mineralization of osteoid.

Osteocalcin (OC) A protein found in the extracellular matrix of bone and dentin and involved in regulating mineralization in the bones and teeth.

Osteoclasts Large, multinuclear cells responsible for resorption of bone.

Osteomalacia Inadequate or delayed mineralization of osteoid; the adult equivalent of rickets (interruption in the development and mineralization of the growth plate in children).

Osteopenia A condition characterized by decreased bone mineral density. Often a predecessor to osteoporosis. It is diagnosed with a bone mineral density test.

Osteoporosis A condition characterized by reduction in bone mass, leading to fractures with minimal trauma; postmenopausal osteoporosis occurs in women after menopause; senile osteoporosis occurs in both men and women later in life.

Paget disease of bone A localized, bone disease characterized by osteoclastic bone resorption followed by replacement of bone in a chaotic fashion. Prevalence varies markedly between countries.

Parathyroid hormone (PTH) A peptide hormone secreted by parathyroid glands in response to decreased free calcium that increases calcium in blood by increasing bone resorption, increasing renal reabsorption of calcium, and increasing the synthesis of 1,25-hydroxyvitamin D [25(OH)D]; the latter increases intestinal absorption of calcium and phosphate.

Parathyroid hormone-related protein (PTHrP) A protein that mimics many actions of PTH, but is a product of a different gene that is expressed in many normal tissues and overexpressed by some tumors in cases of humoral hypercalcemia of malignancy (HHM).

Procollagen type I carboxy-terminal propeptide (PICP) A biochemical marker of bone formation.

Procollagen type I N-terminal propeptide (PINP) A biochemical marker of bone formation.

Pyridium crosslinks A family of molecules that links collagen molecules to each other; the breakdown products of collagen are excreted into blood then urine with attached crosslinks including pyridinoline (PYD) and deoxypyridinoline (DPD).

Renal osteodystrophy Bone diseases associated with chronic renal failure, including high turnover (osteitis fibrosa or secondary hyperparathyroidism) and low turnover (osteomalacia and adynamic bone) diseases.

Rickets A disorder in children caused by a lack of vitamin D, calcium, or phosphate that leads to softening and weakening of the bones. In adults it is known as osteomalacia.

Tartrate-resistant acid phosphatase 5b (TRACP5b) An enzyme derived from osteoclasts; it is a marker of bone resorption.

Vitamin D Fat-soluble vitamin produced by skin upon exposure to sunlight (vitamin D3, also called cholecalciferol) or adsorbed from foods that contain it (vitamin D2 or ergocalciferol); deficiency causes rickets in children and osteomalacia in adults. Related tests with clinical utility include assays for 25(OH)D and 1,25-dihydroxyvitamin D (calcitriol).

This chapter provides an overview of skeletal metabolism then details the clinical chemistry of calcium and other ions, including phosphate and magnesium.

OVERVIEW OF SKELETAL METABOLISM

Bone is composed primarily of an extracellular mineralized matrix with a smaller cellular fraction. It is a dynamic tissue under continuous turnover or *remodeling.* **Osteoclasts** and **osteoblasts** are located on bone surfaces and are responsible for bone resorption and formation, *Osteocytes,* the most abundant cells in mature bone, are located in lacunae within the bone matrix. Osteoclasts resorb bone, osteoblasts lay down new bone at a site of previous bone resorption, and osteocytes nourish the skeleton and regulate bone cell activity (Fig. 39.1). The remodeling cycle can be divided into activation, resorption, reversal, formation, and termination/resting phases.

Giant multinucleate osteoclasts resorb bone by producing hydrogen ions to mobilize minerals and lysosomal enzymes to digest the organic matrix. Osteoblasts form bone by synthesizing the organic matrix, including type I collagen, and participating in the **mineralization** of newly synthesized matrix. The development of the osteoblast phenotype has been divided into three phases (Fig. 39.2).

The organic matrix of bone is 90% type I collagen combined with a large number of noncollagenous proteins. The organic matrix is mineralized by the deposition of inorganic calcium and phosphate in small crystals with carbonate, magnesium, sodium, and potassium. Bone contains 99% of the body's calcium, most of the phosphate (85%), and much of the magnesium (55%).

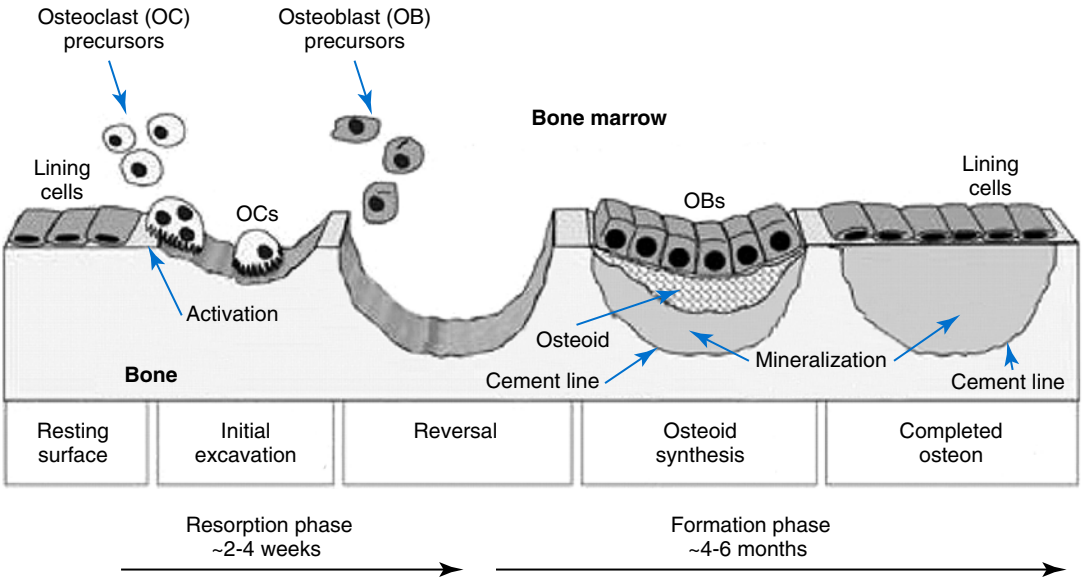


Fig. 39.1 Bone remodeling is shown in two dimensions, in vivo. It occurs in three dimensions, with osteoclasts continuing to enlarge the cavity at one end and osteoblasts beginning to fill in at the other end. (From Riggs, B. L., & Parfitt, A. M. [2005]. Drugs used to treat osteoporosis: the critical need for a uniform nomenclature based on their action on bone remodeling. *J Bone Miner Res, 20*, 177–184.)

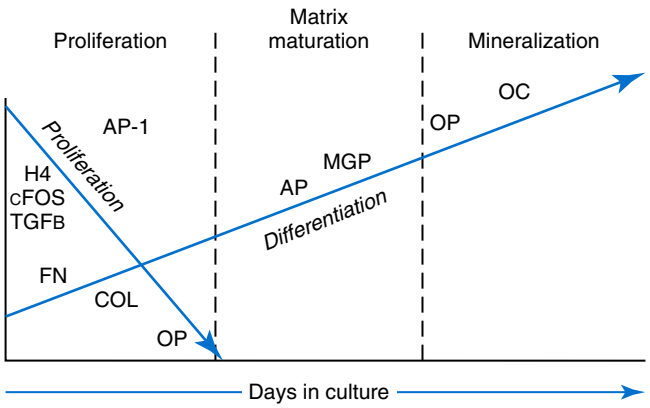


Fig. 39.2 Development of the osteoblast phenotype. The reciprocal relationship between cell growth and differentiation is shown, *arrows* depict expression of cell cycle and cell growth–related genes *(proliferation arrow)* and that of genes related to maturation of the osteoblast phenotype and production of the extracellular matrix *(differentiation arrow)*. The three principal periods in the developmental sequence are designated by *vertical broken lines. AP,* Alkaline phosphate; *COL,* type I collagen; *MGP,* matrix Gla protein; *OC,* osteocalcin; *OP,* osteopontin. (Modified from Lian, J. B., & Stein, G. S. [1995]. Development of the osteoblast phenotype: molecular mechanisms mediating osteoblast growth and differentiation. *Iowa Orthop J, 15*, 118–140.)

Bone Remodeling

From 10% to 30% of the skeleton is remodeled yearly. Bone growth and turnover are influenced by the metabolism of calcium, phosphate, magnesium, PTH, and 1,25-dihydroxyvitamin D (1,25$(OH)_2$D). Bone formation and resorption are affected by thyroid hormones, estrogens, androgens, cortisol, insulin, growth hormone, insulin-like growth factors (IGF-I and IGF-II), transforming growth factor-β (TGF-β), fibroblast growth factor 23 (FGF23), and platelet-derived growth factor (PDGF). Numerous cytokines alter bone remodeling by stimulating resorption; these include interleukins (IL)-1, -4, -6, and -11; macrophage and granulocyte/macrophage colony-stimulating factors (MCSF, GCSF), and TNF-α.

During childhood and adolescence, bone formation markedly exceeds bone resorption (bone modeling)—a process that ends with epiphyseal closure at the end of puberty and is followed, in women, by a period of consolidation for the next 5 to 10 years, when the bone becomes fully mineralized. In healthy individuals, resorption and formation remain in balance for several decades. After the menopause, the decline in estrogen in women triggers an increase in bone resorption that exceeds the capacity of formation and results in a decrease in bone mass. Men do not experience this phase of rapid bone loss unless they develop altered physiology, such as hypogonadism, but they do experience longer-term age-related loss at a rate similar to that in women. **Osteoporosis** results from remodeling imbalance.

BIOCHEMICAL MEASUREMENTS IN METABOLIC BONE DISEASE

CALCIUM

Calcium is the fifth most common element in the body and the most prevalent cation. An average human (70 kg) contains 1 kg, or 25 mol, of calcium predominantly as extracellular crystals in hydroxyapatite ($Ca_{10}(PO_4)_6(OH)_2$).

Biochemistry and Physiology

In blood virtually all calcium is found in plasma, which has a mean calcium concentration of 9.5 mg/dL (2.38 mmol/L). Calcium exists in three physicochemical states in plasma (Table 39.1).

The free (ionized) fraction is the biologically active form. Its concentration is tightly regulated by PTH and 1,25(OH)$_2$D. Synthesis and secretion of PTH by the parathyroid glands is controlled via a calcium-sensing receptor (CaSR). A decrease in circulating free calcium is detected at the CaSR, resulting in the chief cells releasing PTH to act via the kidneys (reabsorption), the gut (absorption), and the skeleton (bone resorption) to increase free calcium (Fig. 39.3) In a classic feedback loop an increase in free calcium interacts with the CaSR, switching off PTH synthesis and release.

There is 80% of protein-bound calcium associated with albumin, with the remaining 20% associated with globulins. Because calcium binds to negatively charged sites on proteins, its binding is pH dependent. For each 0.1 unit change in pH, 0.2 mg/dL (0.05 mmol/L) of inverse change occurs in free calcium.

Calcium can be redistributed among the three plasma pools.

Calcium is classified as intracellular or extracellular. Intracellular calcium has key roles including muscle contraction, hormone secretion, glycogen metabolism, and cell division. The intracellular concentration of calcium in the cytosol of unstimulated cells is around 0.1 μmol/L, less than 1/20,000 of that in extracellular fluid.

Extracellular calcium provides calcium ion for the maintenance of intracellular calcium, bone mineralization, blood coagulation, and plasma membrane potential. Calcium stabilizes the plasma membranes and influences permeability and excitability. A decrease in the plasma-free calcium concentration causes increased neuromuscular excitability and can lead to tetany.

TABLE 39.1 Physicochemical States of Calcium, Phosphate, and Magnesium in Human Plasma

	APPROXIMATE PERCENT OF TOTAL		
State	**Calcium**	**Phosphate**	**Magnesium**
Free (ionized)	50	55	55
Protein-bound	40	10	30
Complexed	10	35	15
Total, mg/dL	8.6–10.3	2.5–4.5	1.7–2.4
Total mmol/L	2.15–2.57	0.81–1.45	0.70–0.99

Modified from Marshall, R. W. (1976). Plasma fractions. In B. E. C. Nordin (Ed.), *Calcium, phosphate and magnesium metabolism* (pp. 162–185). London: Churchill Livingstone.

Clinical Significance

Hypocalcemia

Low total plasma calcium (**hypocalcemia**) may be due to a reduction in albumin-bound calcium, the free fraction of calcium, or both. Hypoalbuminemia is the most common cause of apparent hypocalcemia, particularly in hospitalized patients. Low plasma albumin is seen in chronic liver disease, nephrotic syndrome, congestive heart failure, malignancy, malnutrition, and post-surgical volume replacement with saline or colloidal. Causes of hypocalcemia include preanalytical interference such as presence of ethylenediaminetetraacetic acid (EDTA), chronic kidney disease (CKD), malignancies, drugs, and hypoparathyroidism.

Hypoparathyroidism is mainly caused by parathyroid gland destruction or removal during neck surgery (90%). In pseudohypoparathyroidism resistance to PTH results in increased circulating PTH.

Symptoms and signs of hypocalcemia include paresthesia, increased neuromuscular irritability, central nervous system abnormalities, and fatigue. The severity of symptoms is related to the rate of decrease in free calcium. This can occur during transfusion when large quantities of citrated blood are infused; citrate complexes calcium, leading to lower free calcium concentration without decreasing the total calcium. It is also observed in hysterical over-breathing where an increase in respiratory rate results in respiratory alkalosis, with the increased pH causing protein binding of free calcium. Treatment of symptomatic hypocalcemia may require intravenous calcium, if secondary to hypoparathyroidism or pseudohypoparathyroidism, **vitamin D** or vitamin D analogs plus oral calcium are administered. PTH therapy by intramuscular injection is available to treat hypoparathyroidism. In hypomagnesemic hypoparathyroidism, intravenous magnesium will normalize magnesium, which restores control

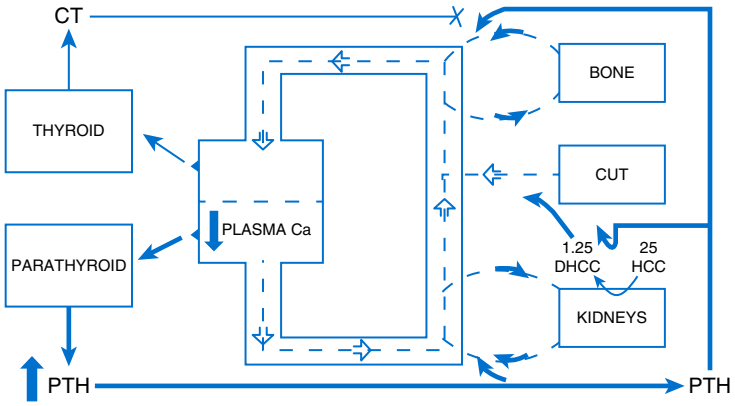


Fig. 39.3 Regulation of calcium via parathyroid hormone *(PTH)* secretion and its direct and indirect action at the kidneys, intestine, and bone. (Modified from Fraser, W. D. [2009]. Hyperparathyroidism. *Lancet, 374*, 145–158.)

over PTH synthesis and secretion, and results in normocalcemia with calcium supplementation. An important exception to symptomatic hypocalcemia occurs in CKD when marked acidosis is present. Hypocalcemia triggers muscle contraction, but acidosis impairs muscle contraction.

Hypercalcemia

The major causes of **hypercalcemia** in clinical practice include primary hyperparathyroidism (PHPT), malignancies, renal failure, drugs, and vitamins D or A overdose. Hypercalcemia results when the influx of calcium into the extracellular fluid compartment from the skeleton, intestine, or kidney is greater than the efflux. When the kidney's capacity to excrete filtered calcium is exceeded, hypercalcemia develops; where renal failure is present and calcium excretion is decreased, hypocalciuria may paradoxically be present.

PHPT is the most common pathologic cause in outpatients, whereas malignancy is common in hospitalized patients. Vitamin D or analog therapy has become a common cause of hypercalcemia. These account for 90% to 95% of all cases.

In developed countries 80% of PHPT patients are free of overt symptoms on presentation because of detection through chemistry panels, including calcium. The most common signs and symptoms of hypercalcemia are nonspecific and include nausea, vomiting, constipation, weakness, kidney stone formation, depression, polyuria, and tissue calcification.

Measurement of intact PTH with contemporaneous calcium is the most sensitive and specific test for parathyroid function and is central to the differential diagnosis of hypercalcemia. If intact PTH is within the upper half of the reference interval, it is inappropriate for the prevailing calcium concentration. Serum $1,25(OH)_2D$ is often in the upper half of the reference interval or increased in PHPT, PTH stimulates production from 25(OH)D. Nonparathyroid causes of hypercalcemia are associated with very low or suppressed PTH. $1,25(OH)_2D$ is low-normal or suppressed in nonparathyroid hypercalcemia, except in sarcoidosis or other granulomatous diseases and certain lymphomas, in which pathologic tissues contain the 25-hydroxyvitamin D-1α-hydroxylase which produces $1,25(OH)_2D$.

Hypercalcemia occurs in 5% to 30% of individuals with cancer. Solid tissue malignancies, particularly squamous cell carcinomas, commonly produce **parathyroid hormone-related peptide (PTHrP)**, which stimulates bone resorption. Skeletal metastases from cancer can also produce hypercalcemia.

Samples must be obtained prior to initiating treatment for hypercalcemia as lowering the calcium in **hypercalcemia-associated malignancy (HCM)** will stimulate PTH secretion and can cause diagnostic confusion (hypercalcemia with detectable PTH).

Therapies are directed toward treating the malignancy, decreasing the plasma calcium by saline diuresis, and decreasing osteoclastic resorption (bisphosphonates, denosumab, calcitonin). Corticosteroids can be useful in reducing intestinal absorption of calcium in $1,25(OH)_2D$-mediated hypercalcemia.

Familial benign hypocalciuric hypercalcemia (FBHH) is characterized by the presence of hypercalcemia and hypocalciuria (low fractional calcium excretion).

Measurement of Calcium

Methods for quantifying calcium in blood can measure free Ca^{2+} or the total concentration of calcium and the subsequent calculation of albumin adjusted calcium. The term *ionized calcium,* although widely used, is a misnomer; all calcium in plasma or serum is ionized.

Free calcium is the biologically active fraction of blood calcium; it is tightly regulated by PTH and $1,25(OH)_2D$, and thus is the best indicator of calcium status. Methods for total, albumin adjusted, and free calcium are currently in use and have their own sources of error. Adjustment of total calcium for the prevailing albumin concentration is often performed to correlate the total calcium concentration to the free calcium. Free calcium measurements and adjusted calcium calculations have been recommended because of the consequences of delayed treatment and the cost of working-up patients with misleading total calcium results.

Measurement of Total Calcium

Only photometric, ion selective electrode (ISE), and atomic absorption spectrophotometry are used in clinical laboratories for measuring serum and urine total calcium.

The specimen is acidified to convert protein-bound and complexed calcium to free calcium before measurement by ISE (for additional information see Chapter 10).

Photometric methods. Total calcium is most frequently measured by spectrophotometry using metallochromic indicators or dyes. Of the metallochromic indicators that change color on selectively binding calcium, *o*-cresolphthalein complexone (CPC) [3′,3″-bis({bis-[carboxymethyl]amino}-methyl)-5′,5″-dimethylphenolphthalein] and arsenazo III are most widely used.

***o*-Cresolphthalein complexone.** In alkaline solution, the metal-complexing dye CPC forms a red chromophore with calcium; the color is measured between 570 and 580 nm. The sample is diluted with acid to release protein-bound and complexed calcium. Organic base, most often diethylamine, 2-amino-2-methyl-1-propanol, or 2-ethylaminoethanol, is added to buffer the reaction and to produce an alkaline pH. Interference by magnesium is reduced: by adding 8-hydroxyquinoline, by buffering the reaction mixture to near pH 12, and by measuring the absorbance near 580 nm. Urea reduces the turbidity of lipemic specimens and enhances complex formation. Adding ethanol or other organic solvents reduces blank absorbance. Calcium forms 1:1 and 2:1 complexes with CPC, the former predominating at lower concentrations. The calibration curves are nonlinear at low concentrations. Multipoint calibration has been recommended. Sodium acetate may improve linearity.

Arsenazo III. Arsenazo III [1,8-dihydroxynaphthalene-3,6-disulfonic acid-2,7-bis(azo-2)-phenylarsonic acid], at mildly acidic pH, has higher affinity for calcium than magnesium and produces an intense, purple complex. A reaction pH of 6 is used; imidazole buffers the reaction. The

spectral properties of arsenazo III are dependent on pH. Interference from most biological pigments is reduced by measuring the calcium–dye complex near 650 nm. Citrate can cause negative interference, particularly with dry-slide techniques. Clinically significant interference may be noted in patients receiving citrated blood or blood products. Unlike CPC, which has limited stability, the arsenazo III reagent is stable.

Atomic absorption spectrometry methods. The Clinical Laboratory and Standards Institute (CLSI) has approved an atomic absorption spectrophotometry (AAS) method as a reference method for measuring total serum calcium. The reference method is reported to have an accuracy of 100 ± 2%, compared with 100 ± 0.2% for isotope dilution–mass spectrometry (ID–MS) the definitive method developed by the National Institute of Standards and Technology.

Total calcium adjusted for albumin (adjusted or corrected calcium). Wide variation in the concentrations of compounds that bind calcium in blood may be noted; this variation will affect the measured total calcium concentration without changing the free calcium. Several types of calculation have been suggested to adjust the total calcium concentration. The goal is to produce an adjusted result that would have been found if the concentrations of all compounds that bind calcium had been within their respective reference intervals. Only adjustments based on albumin are commonly used. The term *adjusted* is preferable to *corrected calcium.*

Adjusted calcium is calculated from total calcium and albumin by calculating a correction factor by multiplying the deviation of plasma albumin from the mean of its reference interval by the slope of the regression of total calcium against albumin. Two equations are frequently used with results expressed as mg/dL and mmol/L:

$$\text{Adjusted total calcium (mg/dL)} = \text{Total calcium (mg/dL)} + 0.8\,[4 - \text{Albumin (g/dL)}]$$

$$\text{Adjusted total calcium (mmol/L)} = \text{Total calcium (mmol/L)} + 0.02\,[40 - \text{Albumin (g/dL)}]$$

These are simplistic calculations and in practice it is recommended that laboratories should establish equations for their reference populations that incorporate linear regression for the different methods used to measure the total calcium and albumin concentrations. Many factors affect the distribution of calcium among the fractions (Box 39.1). The reliability of adjustment for serum albumin deteriorates in patients with very low or high albumin concentrations and in patients with severe disease and multiple organ failure as seen in intensive care units. Direct determination of free calcium by ISE is preferable to adjusted calcium but it should be remembered that ISE methods are also subject to effects due to prevailing albumin concentrations.

Specimen requirements. Serum and heparinized plasma are the preferred specimens. Citrate, oxalate, and EDTA anticoagulants should not be used for spectrophotometric methods as they form complexes with calcium. Acidification of urine is recommended to ensure complete dissociation of calcium from complexes that can be formed at high pH seen in some samples.

BOX 39.1 Factors Altering the Distribution Between Protein-Bound, Complexed, and Free Calcium and Compounding the Interpretation of Total Calcium

Factors Altering Protein Binding of Calcium
- Altered concentration of albumin or globulins
- Abnormal proteins
- Heparin
- pH
- Free fatty acids
- Bilirubin
- Drugs
- Temperature

Factors Altering Complex Formation
- Citrate
- Bicarbonate
- Lactate
- Phosphate
- Pyruvate and α-hydroxybutyrate
- Sulfate
- Anion gap

Interferences. Hemolysis, icterus, lipemia, paraproteins, and magnesium have been reported to interfere with photometric methods. Many methods use bichromatic analysis, multiwavelength corrections, or blanking to reduce interference. Lipemic specimens should be ultracentrifuged before analysis, or treated to remove the lipid fraction.

Although hemolysis can cause a negative error because red cells contain lower concentrations of calcium than plasma, more significant errors may be caused by the spectral interference of hemoglobin. If hemolyzed specimens must be analyzed, blanking with ethylene glycol-*O,O′*-bis (2-aminoethyl)-*N,N,N′,N′*-tetraacetic acid (EGTA) is recommended.

Individual instruments and methods should be evaluated for their susceptibility to effects of interferents. Care should be taken to prevent sample contamination with calcium. How the patient is prepared and how the specimen is obtained can have a significant effect on both free and total calcium measurements (see Chapter 3 and Box 39.2).

Measurement of Free (Ionized) Calcium

See also Chapters 10 and 24.

Blood gas analyzers, using ISEs, provide rapid whole blood determinations of plasma free calcium and electrolytes, as well as blood gases. Sensitive potentiometers measure the voltage difference between the calcium or pH and reference electrodes for calibrating solutions or samples. A microprocessor

BOX 39.2 Preanalytical Factors Affecting Measurement of Serum Total or Free Calcium

In Vivo
Tourniquet use and venous occlusion
Changes in posture: 10%–12% increase in total calcium and 5%–6% increase in free calcium on standing
Exercise
Hyperventilation
Fist clenching
Alimentary status
Alterations in protein binding
Alterations in complex formation

In Vitro
Inappropriate anticoagulants
Dilution with liquid heparin
Interfering levels of heparin
Contamination with calcium
Bungs, glassware, tubes
Specimen handling
Alterations in pH (free calcium)
Adsorption or precipitation of calcium
Spectrophotometric interference
Hemolysis, icterus, lipemia

calibrates the system and calculates calcium concentration and pH. Most instruments simultaneously measure free calcium and pH at 37°C.

Calcium ISEs contain a calcium-selective membrane, which encloses an inner reference solution of calcium chloride often containing saturated AgCl and physiologic concentrations of NaCl, KCl, and an internal reference electrode. The reference electrode, usually Ag/AgCl, is immersed in this inner reference solution. Modern calcium ISEs use liquid membranes containing the ion-selective calcium sensor dissolved in an organic liquid trapped in a polymeric matrix. The Ca^{2+} ionophores may be based on a polyvinyl chloride matrix and contain ETH1001 or ETH129 as a carrier. Instead of these two neutral carriers, ion exchangers, such as organophosphate sensors have been used. Neutral carrier membranes contain an uncharged calcium-selective organic molecule, such as ETH1001, dissolved in a plasticizer and trapped in a polyvinyl chloride membrane. These molecules have a favorable steric and electrostatic pocket or site for selectively binding calcium. Ion exchangers or negatively charged carrier membranes are calcium salts, such as calcium bis(di-*n*-octylphenyl) phosphate, dissolved in di-*n*-octylphenyl phosphonate, and trapped in a polyvinyl chloride membrane. The electrochemical cell is completed by the external reference electrode, an Ag/AgCl or calomel electrode, which is in contact with the specimen through a liquid/liquid junction or a salt bridge of potassium chloride or sodium formate. The potential difference across the cell is logarithmically related to the activity of free calcium ions in the sample by the Nernst equation. By convention, free calcium is converted from activity to concentration using its activity coefficient, which is itself dependent on ionic strength.

Most free calcium analyzers adjust and maintain samples at 37°C, thereby ensuring that results are physiologically relevant.

Interferences. Because ISEs measure ion activity, they are affected by the ionic strength of a specimen (see Chapter 10). Free calcium analyzers (and the associated calibrators) are optimized for serum, plasma, or whole blood. Modern electrodes have high selectivity for calcium over Na^+, K^+, Mg^{2+}, H^+, and Li^+ ions. At normal concentrations, these cations have little effect on the accuracy of free calcium measurements. Wide variations in the concentration of Na^+ and high concentrations of Mg^{2+} and Li^+ may influence the concentration of free calcium. Electrodes are insensitive to H^+, with insignificant interference noted between pH 5 and 9.

Many anions including protein, phosphate, citrate, lactate, sulfate, and oxalate form complexes with calcium ions. Although these anions reduce the concentration of free calcium by complex formation, they do not directly interfere with measurement of the calcium that is free. Protein deposits on the electrode may act as a divalent cation exchanger, resulting in positive interference with high concentrations of Mg^{2+} ions. Newer electrodes use a dialysis membrane or a neutral carrier to reduce or eliminate this protein effect, which typically is less than +0.02 mmol/L (0.08 mg/dL) for 1 g/dL (10 g/L) of protein. The type of dialysis technology or protein exclusion technology can result in different effects of protein concentration on free calcium measurement, especially at extremes of albumin concentration.

Effect of pH. Albumin, with up to 30 binding sites for calcium, accounts for 80% of the protein-bound calcium. Increasing the pH of a specimen in vitro increases the ionization and negative charge on albumin leading to an increase in protein-bound calcium and decrease in free calcium. Decreasing pH decreases ionization and negative charge, decreasing protein-bound calcium and increasing free calcium. Free calcium changes by 5% for each 0.1 unit change in pH (Fig. 39.4).

Because of this inverse relationship between free calcium and pH, specimens must be analyzed at the patient's pH.

Specimen requirements. Specimens for free calcium must be collected and handled to minimize alterations in pH and free calcium caused by loss of CO_2 and blood cell metabolism. Free calcium may be measured in heparinized whole blood, heparinized plasma, or serum. Where specimens are analyzed rapidly, heparinized whole blood is preferred because it reduces processing time, uses smaller specimen volume, and avoids alteration in pH associated with centrifugation. All syringes and evacuated tubes should be filled completely, kept tightly sealed, and handled anaerobically to prevent loss of CO_2 and the increase in pH that may occur when specimens are exposed to air. It is best that the sample for free calcium is collected in a separate container to minimize the likelihood that the specimen may be analyzed aerobically.

Whole blood specimens should be analyzed within 15 to 30 minutes of sampling. Free calcium is reported to be stable in

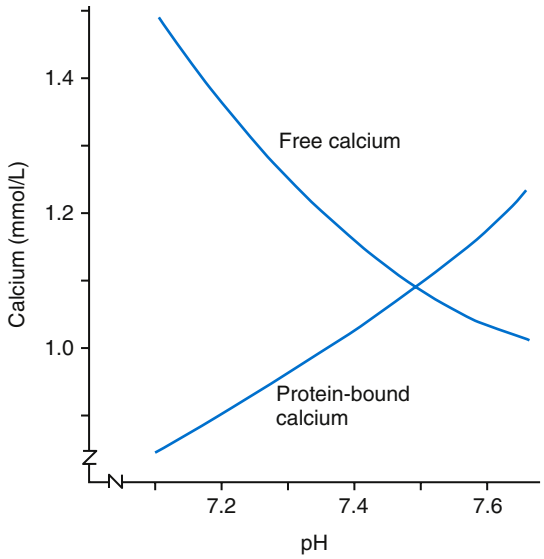


Fig. 39.4 Effect of pH on free and protein-bound calcium.

whole blood for at least 1 hour at room temperature and for 4 hours at 4°C. If specimens cannot be analyzed promptly, they can be collected in an ice-water slurry to minimize metabolism. For delayed analysis, serum may be the optimal sample type because of elimination of the anticoagulant and reduction in the occurrence of microclots. Once centrifuged, specimens are stable for hours at 25°C and for days at 4°C in a sealed tube.

The use of aerobic specimens for the measurement of free calcium with correction to pH 7.4 has been criticized and is not recommended. Free calcium at pH 7.4 may be misleading in patients with respiratory and metabolic alkalosis or acidosis.

Effects of anticoagulants. Heparin is the only acceptable anticoagulant for free calcium determinations. Heparin, a polyanion, significantly lowers free calcium at the concentrations (≥30 to 100 U/mL) found in many conventional blood gas syringes. Several commercial syringes containing lyophilized heparin are suitable for free calcium determinations. Most evacuated collection tubes, when filled completely, contain concentrations of heparin (15 U/mL) that only slightly decrease free calcium. It is important that all syringes and tubes be filled completely to minimize dilution and/or heparin effects.

Calibrators and quality controls. Various calibration solutions are used for free calcium analyzers. The buffers in which these calibrators are prepared may affect the liquid junction potential and calcium binding; this is usually corrected by the instrument software.

Aqueous quality control materials are available for free calcium. Because simple aqueous controls may not reliably detect changes in performance with patient specimens, the use of serum-based quality control materials is recommended.

Patient Preparation and Sources of Pre-analytical Error for Total and Free Calcium Measurements

Patient preparation and the manner of specimen collection can significantly affect the results of total and free calcium determinations (see Box 39.2).

A common source of pre-analytical error is the increase in total, but not free, calcium concentration associated with tourniquet use and venous occlusion during sampling. Errors of 0.5 to 1.0 mg/dL (0.12 to 0.25 mmol/L) in total calcium may result from the increase in protein-bound calcium caused by the vascular efflux of water during stasis.

Fist clenching or other forearm exercise should be avoided before phlebotomy, this causes a decrease in pH and an increase in free calcium.

Changes in posture cause fluid shifts within 10 minutes and thus alter the concentrations of cells and large molecules, including albumin and total calcium. Standing decreases intravascular water and increases the total calcium concentration by 0.2 to 0.8 mg/dL (0.05 to 0.20 mmol/L) while recumbency has the opposite effect.

Reference Intervals for Total and Free Calcium in Serum and Plasma

Total Calcium

The reference interval for serum/plasma calcium is usually defined by an upper limit of 10.1 to 10.5 mg/dL (2.52 to 2.62 mmol/L) and a lower limit between 8.5 and 8.8 mg/dL (2.12 to 2.20 mmol/L). Pediatric reference intervals have been established as part of the CALIPER study (see Appendix).

An analytical goal, for between-day imprecision, the coefficient of variation (CV), is 0.9% or less based on within-person biological variation. Current methods achieve between day CVs of 1.5% or less within laboratories.

Free Calcium

The reference interval of free calcium (in adults) is 4.6 to 5.3 mg/dL (1.15 to 1.33 mmol/L). Whole blood specimens develop a liquid junction potential different from that of serum or plasma because of the presence of erythrocytes. A positive bias that is directly proportional to the hematocrit has been reported. Free calcium values differ among capillary blood, venous blood, and serum samples because of pH

differences. Reference intervals should be determined using the local instrument, specimen type, and collection protocol, and reference subjects representative of the population served.

Physiologic Variation in Calcium

Plasma calcium concentration varies with age, gender, season, during pregnancy, and throughout 24 hours. Total and free calcium decrease in the elderly. During pregnancy, total calcium and adjusted calcium decline in parallel with plasma albumin, free calcium is unchanged. The fetal circulation is relatively hypercalcemic, with higher total and free calcium in cord blood than in maternal plasma. Calcium concentrations decline after birth in healthy term neonates during the first few days, but soon increase to concentrations slightly greater than in adults.

Interpretation of Total and Free Calcium Results

Calcium status is best determined by measuring free calcium. Interpretation of total serum calcium is complicated by its association with protein and inorganic and organic ions. Disagreement between free and total calcium occurs in a high percentage of specimens. One study reported disagreement between free and total or adjusted calcium values in 18% and 31% of patients, respectively.

Free calcium is more useful than total calcium determination in hospital patients who have received citrated blood or platelets, heparin, bicarbonate, intravenous solutions, or calcium, and in intensive care because of abnormal protein concentrations and circulating factors that alter calcium binding to albumin. Abnormally low free calcium is frequently found in critically ill patients.

The calcium metabolism of patients with renal disease is best evaluated by the determination of free calcium because of alterations in protein, pH, protein binding of calcium, and calcium complexes with organic and inorganic anions.

Free calcium may be useful in the diagnosis of hypercalcemia. Patients with surgically proven PHPT often have increases in free calcium greater than total calcium. Free calcium is more sensitive than total calcium in detecting HCM, as patients frequently have decreased plasma albumin concentrations. Less commonly, paraproteins produced in myeloma may bind calcium, complicating the interpretation of total or adjusted calcium measurements.

In neonates, free rather than total calcium is recommended because of its speed of analysis and ability to use small skin-puncture specimens, its greater validity in the presence of hyperphosphatemia, alterations in pH, and persistence of α-fetoprotein after birth.

Urinary Calcium

Urinary calcium excretion reflects calcium intake, intestinal absorption, skeletal resorption, renal tubular filtration, and reabsorption. Healthy men and women excrete up to 300 mg (7.5 mmol) of calcium per day on a diet with unrestricted calcium content, and up to 200 mg (5 mmol) per day on a calcium-restricted diet (500 mg or 12.5 mmol calcium per day or less).

The reference interval for urinary calcium for spot fasting or timed specimens collected after an overnight fast is less than 0.16 mg/100 mL (<0.04 mmol/L) of glomerular filtrate (GF), as calculated by the equation:

$$\text{UCa (mg/100 mLGF)} = \frac{[\text{UCa (mg/dL)}] \times [\text{Serum creatinine (mg/dL)}]}{\text{Urinary creatinine (mg/dL)}}$$

An important, but under-recognized contributor to urine calcium excretion, is urinary sodium. Patients with hypercalciuria and normal sodium excretion have renal leak hypercalciuria, which can result in increased secretion of PTH and can be corrected by a thiazide. Hypercalciuria with increased urinary sodium excretion responds rapidly to a reduced sodium intake.

Under fasting conditions, the intestinal and renal components are relatively fixed, and calcium excretion (mg/100 mL [mmol/L] of GF) in the fasting state is used to assess the skeletal component. A value greater than 0.16 mg/100 mL (>0.04 mmol/L) of GF implies an increase in osteoclastic bone resorption.

Specimens may be collected in a container containing acid to prevent calcium salt precipitation. HCl (6 mol/L) is commonly used. Specimens collected without acid should be acidified and allowed to stand for 1 hour before thorough remixing and aliquoting. The ability of postcollection acidification to redissolve all calcium salts has been questioned.

POINTS TO REMEMBER

Calcium

- The skeleton is the major store of calcium.
- The free calcium is the physiologically important ion and is tightly regulated.
- PTH and vitamin D metabolites combine to regulate calcium via the kidney, intestine, and bone.
- Calcium adjusted for albumin can supply important information regarding an individual's calcium status.
- A calcium sensing receptor is present on the chief cells of the parathyroid gland that is responsive to changes in circulating free calcium.

PHOSPHATE

An adult has 600 g or approximately 20 mol of phosphorus in inorganic and organic phosphates, 85% is in the skeleton, and the rest is in soft tissue.

Biochemistry and Physiology

Plasma contains both inorganic and organic phosphate, but only inorganic phosphate is measured. Inorganic phosphate exists as both monovalent ($H_2PO_4^-$) and divalent (HPO_42^-) phosphate anions. The ratio of $H_2PO_4^-$ to HPO_42^- is pH dependent and varies from 1:1 in acidosis to 1:4 at pH 7.4 and 1:9 in alkalosis. Only 10% of phosphate in serum is protein-bound; 35% is complexed with sodium, calcium, and magnesium; the remainder, 55%, is free.

Inorganic phosphate is a major component of hydroxyapatite in bone; thus it plays an important role in the structural support of the body and provides phosphate for the extracellular and intracellular pool.

In soft tissue, phosphate is mainly intracellular. In cells, most is organic and is incorporated into nucleic acids, phospholipids, phosphoproteins, and high-energy compounds. Adenosine triphosphate (ATP) and other phosphates, such as creatine phosphate, are involved in many energy-intensive processes, such as muscle contractility, neurologic function, and electrolyte transport. Phosphate is an essential element of cyclic nucleotides and cofactors such as nicotinamide adenine dinucleotide phosphate (NADP). It is important for the activity of several enzymes, including adenylate cyclase, 25-hydroxyvitamin D-1α-hydroxylase, and in the production of 2,3-diphosphoglycerate, the key compound regulating oxygen affinity of hemoglobin. Intracellular phosphate is involved in the regulation of intermediary metabolism of proteins, fats, carbohydrates, and in gene transcription and cell growth.

The regulation of phosphate involves the kidneys, intestine, and skeleton. Plasma phosphate in adults is maintained between 0.7 and 1.4 mmol/L. Plasma phosphate concentration demonstrates a diurnal variation with a significant increase after the evening meal to a peak in the early hours of the morning, decreasing to a nadir early morning. This circadian rhythm is blunted by fasting. $1,25(OH)_2D$ can promote phosphate and calcium absorption via the intestine and can increase phosphate mobilization through osteoclastic resorption of bone. PTH can have effects on phosphate via the kidneys by increasing expression of the type IIa sodium–phosphate co-transporter (NPT2a). PTH promotes phosphaturia and lowers circulating phosphate. PTH can stimulate both bone resorption and bone formation depending on the fluctuation in PTH and so can either release or deposit phosphate in bone. Variations observed in PTH and $1,25(OH)_2D$ in health and disease do not fully explain the changes in phosphate homeostasis and so the existence of "phosphatonin(s)" such as fibroblast growth factor (FGF)23 had long been suspected.

FGF23 is a mid to long-term modulator of phosphate homeostasis. FGF23 increases kidney fractional excretion of phosphate. It decreases production of $1,25(OH)_2D$, by decreasing 25-hydroxyvitamin D 1-hydroxylase activity. These key effects lead to decreased plasma phosphate. FGF23 is secreted by bone cells (osteocytes, osteoblasts, and osteoclasts) in response to sustained increased plasma phosphate or increased plasma $1,25(OH)_2D$. In CKD, FGF23 increases in a compensatory mechanism to counter increasing plasma phosphate. As kidney function deteriorates, responsiveness to FGF23 declines reducing FGF23's ability to lower plasma phosphate.

Clinical Significance

Hypophosphatemia

Hypophosphatemia, a concentration of inorganic phosphate in serum below the reference interval (<2.5 mg/dL [<0.81 mmol/L]), is common in hospitalized patients (2%).

Hypophosphatemia or phosphate depletion in blood may be caused by a shift of phosphate from extracellular to intracellular spaces; renal phosphate wasting; decreased intestinal absorption, intracellular phosphate loss. Major causes of hypophosphatemia and phosphate depletion include intracellular shift, such as that in alcoholism; carbohydrate induced-insulin secretion; vomiting; diarrhea; inadequate intestinal phosphate absorption, such as in patients taking aluminum- or magnesium-containing antacids; renal phosphate wasting; ketoacidosis; genetic causes; and certain drugs such as anticonvulsants and diuretics.

Treatment of hypophosphatemia depends on phosphate concentrations and on symptoms. Patients with moderate hypophosphatemia often require only treatment of the underlying disorder and, possibly, oral phosphate. In patients with marked symptoms of hypophosphatemia, particularly if respiratory muscle weakness is present, parenteral administration of phosphate is indicated.

Hyperphosphatemia

The most common cause is inability of the kidneys to excrete phosphate. Hyperphosphatemia is a major clinical problem in CKD (see Chapter 35). In acute kidney injury and CKD, a decrease in glomerular filtration rate (GFR) reduces the renal excretion of phosphate, resulting in hyperphosphatemia. Moderate increases in plasma phosphate occur in individuals with low PTH, PTH resistance, or acromegaly.

The clinical manifestations of hyperphosphatemia depend on its rate of onset. A rapid increase in plasma phosphate may be associated with hypocalcemia. Symptoms may include tetany, seizures, and hypotension.

Therapy for hyperphosphatemia is directed toward correcting the cause. In renal failure and in hypoparathyroidism, dietary restriction of phosphate and agents that bind phosphate in the intestine are useful in lowering phosphate.

Measurement of Phosphate

Most methods used to measure serum inorganic phosphate are based on the reaction of phosphate ions with ammonium molybdate forming a phosphomolybdate complex measured by a spectrophotometer.

$$7H_3PO_4 + 12(NH_4)_6Mo_3O_{24} + 4H_2O \rightarrow 7(NH_4)_3\left[PO_4(MoO_3)_{13} + 51NH^{+4} + 51OH^- + 33H_2O\right]$$

The colorless phosphomolybdate complex may be measured directly by ultraviolet (UV) absorption (340 nm) or reduced to molybdenum blue and measured at 600 to 700 nm. An acidic pH is necessary for the formation of complexes. A higher pH can result in spontaneous reduction of molybdate. Solubilizing agents such as Tween 80 prevent protein precipitation. Measurement of unreduced complexes has several advantages, including simplicity, speed, and reagent stability. Disadvantages include greater interference by hemolysis, icterus, and lipemia when 340 nm wavelength is used.

Specimen Requirements

Serum and heparinized plasma are preferred specimens. Inorganic phosphate concentrations are 0.2 to 0.3 mg/dL (0.06 to 0.10 mmol/L) lower in heparinized plasma than in serum. Anticoagulants such as citrate, oxalate, and EDTA interfere with formation of the phosphomolybdate complex.

Phosphate concentrations in plasma or serum are increased by prolonged storage at room temperature or 37°C. Hemolyzed specimens are unacceptable, erythrocytes contain high concentrations of organic phosphate esters. Inorganic phosphate increases by 4 to 5 mg/dL (1.29 to 1.61 mmol/L) per 24 hours in hemolyzed specimens at 4°C, and more rapidly at 37°C.

Phosphate is stable in separated serum for several days at 4°C and for months when frozen, provided evaporation and lyophilization are prevented.

Interferences

Positive or negative interference has been noted with hemolyzed, icteric, and lipemic specimens. Mannitol, fluoride, and monoclonal immunoglobulins are reported to interfere.

Reference Intervals

Phosphate is often referred to as "phosphorus," but this is inaccurate and misleading; only phosphate, not elemental phosphorus, circulates in blood.

In adults, the reference interval for serum phosphate is 2.5 to 4.5 mg/dL (0.81 to 1.45 mmol/L). In children, it is 4.0 to 7.0 mg/dL (1.29 to 2.26 mmol/L). Serum phosphate concentrations are about 50% higher in infants than in adults and decline throughout childhood. Detailed pediatric reference intervals were established in the CALIPER study (see Appendix).

Urinary phosphate excretion varies with age, muscle mass, renal function, PTH and time of day. Urinary excretion of phosphate is essentially equivalent to dietary intake. On a nonrestricted diet, the reference interval for urinary phosphate is 0.4 to 1.3 g/24 h (12.9 to 42.0 mmol/24 h).

Urine should be collected in 6 mol/L HCl, to avoid precipitation of phosphate complexes. Simultaneous measurement of phosphate and creatinine in serum and urine with fasting morning spot or 1- to 2-hour timed collections permits calculation of the renal phosphate threshold ($TmPO_4$/GFR). The clearance of phosphate divided by creatinine clearance can be plotted on a nomogram, and the $TmPO_4$/GFR determined.

POINTS TO REMEMBER

Phosphate

- Phosphate can be stored in bone combined with calcium as hydroxyapatite.
- Abnormalities of phosphate metabolism can lead to significant bone disorders including rickets.
- Phosphate metabolism is under the regulation of PTH, vitamin D metabolites, and FGF23.
- Phosphate in serum is subject to significant fluctuation throughout the day influenced by nutritional intake.
- CKD results in significant increases in serum phosphate.

MAGNESIUM

Magnesium is the fourth most abundant cation in the body and the second most prevalent intracellular cation. The total body magnesium content is 25 g (~1 mol), of which 55% resides in the skeleton. One-third of skeletal magnesium is exchangeable and is thought to serve as a reservoir for maintaining the extracellular magnesium. Forty-five percent of magnesium is intracellular.

Biochemistry and Physiology

The concentration of magnesium in cells varies from 2.4 to 7.3 mg/dL (1 to 3 mmol/L). The higher the metabolic activity of a cell, the greater is its magnesium content. Within cells, most of the magnesium is bound to proteins and negatively charged molecules; 80% to 90% of cytosolic magnesium is bound to ATP, and MgATP is the substrate for numerous enzymes. Of the total cellular magnesium 0.5% to 5.0% is free. The transport of magnesium across the cellular membrane is regulated by a magnesium transport system.

Extracellular magnesium accounts for 1% of the total body magnesium content. Of the magnesium in plasma 55% is free, 30% is associated with proteins (primarily albumin), and 15% is complexed with phosphate, citrate, and other anions (Table 39.2).

Magnesium is a cofactor for 300+ enzymes. Magnesium is important in oxidative phosphorylation, glycolysis, cell replication, nucleotide metabolism, and protein biosynthesis. A decrease in plasma magnesium lowers the threshold of axonal stimulation and increases nerve conduction velocity. Magnesium influences neurotransmitter release at neuromuscular junctions. Reducing plasma magnesium results in increased neuromuscular excitability.

The Institute of Medicine recommends a daily magnesium intake of 310 to 420 mg. The small bowel absorbs 20% to 80% of dietary magnesium; absorption is proportional to the amount in the diet.

The kidneys play a major role in regulating magnesium balance. There is 70% to 80% of magnesium that is ultrafilterable with 5% to 15% being reabsorbed passively in the proximal

TABLE 39.2 Physicochemical States tools of Calcium, Phosphate, and Magnesium in Human Plasma

	APPROXIMATE PERCENT OF TOTAL		
State	**Calcium**	**Phosphate**	**Magnesium**
Free (ionized)	50	55	55
Protein-bound	40	10	30
Complexed	10	35	15
Total, mg/dL	8.6–10.3	2.5–4.5	1.7–2.4
Total mmolfL	2.15–2.57	0.81–1.45	0.70–0.99

Modified from Marshall RW. Plasma fractions. In: Nordin BEC, ed. *Calcium, Phosphate and Magnesium Metabolism*. London: Churchill Livingstone; 1976:162–85.

tubules. In the thick ascending limb of the loop of Henle 65% to 75% is absorbed paracellularly.

Clinical Significance

Hypomagnesemia/Magnesium Deficiency

Hypomagnesemia is common in hospital patients. Ten percent of patients admitted to hospital and 65% of patients in the intensive care unit (ICU) may be hypomagnesemic. Moderate or severe magnesium deficiency is usually due to loss of magnesium from the gastrointestinal tract (GI) or kidneys.

Vomiting, diarrhea, and nasogastric suction lower, and may deplete, body magnesium. Excessive urinary losses of magnesium from the kidneys are important causes of magnesium deficiency; other causes include alcohol, diabetes mellitus, and aminoglycoside antibiotics.

Magnesium deficiency is usually secondary to another disease or a therapeutic agent. Features of the primary disease may complicate or mask magnesium deficiency. Neuromuscular hyperexcitability with tetany and seizures may be present. Cardiac arrhythmia is a more serious outcome of deficiency.

Extracellular magnesium accounts for 1% of total body magnesium, but plasma magnesium correlates poorly with total body magnesium. Determination of serum magnesium is the most widely used test to assess magnesium status. Intracellular magnesium depletion and magnesium deficiency may exist despite a normal plasma magnesium. Hypocalcemia, hypokalemia, neuromuscular hyperirritability, and cardiac arrhythmias should be an alert to the possible presence of magnesium deficiency.

Acute symptomatic magnesium deficiency is treated with parenteral magnesium; mild depletion is treated with oral magnesium.

Hypermagnesemia

Magnesium intoxication is rarely encountered, a mild to moderate increase in magnesium concentration is observed in 12% of hospital patients. Symptomatic **hypermagnesemia** is almost always caused by excessive intake, resulting from administration of antacids, enemas, and parenteral fluids containing magnesium. Many patients have concomitant renal failure, limiting the kidneys ability to excrete magnesium.

Depression of the neuromuscular system is the most common manifestation of magnesium intoxication. Very high concentrations may result in cardiac arrest. Calcium acutely antagonizes the toxic effects of magnesium and can promote kidney excretion; patients with severe magnesium intoxication may require intravenous calcium. Peritoneal dialysis or hemodialysis against a low-magnesium dialysis bath effectively lowers magnesium.

Measurement of Total Magnesium

Photometric Methods

Several metallochromic indicators or dyes change color on selectively binding magnesium. Calmagite [1-(1-hydroxy-4-methyl-2-phenylazo)-2-naphthol-4-sulfonic acid], a metallochromic indicator, forms a colored complex with magnesium in alkaline solution, measured at 530 to 550 nm. EGTA is added to reduce calcium interference. Potassium cyanide prevents the formation of heavy metal complexes, and polyvinylpyrrolidone and surfactants reduce interference from protein and lipemia. Methylthymol blue forms a blue complex with magnesium, measured around 600 nm. A formazan dye [1,5-bis(3,5-dichloro-2-hydroxyphenyl)-3-formazan carbonitrile] forms a complex with magnesium at alkaline pH, measured at 630 nm by thin-film reflectance photometry. *N,N′*-[1,2-ethanediylbis(oxy-2,1-phenylene)bis(*N*-carboxymethyl)] glycine chelates calcium. This thin-film reflectance method shows relatively little interference from icteric, lipemic, and hemolyzed specimens. Magon, or xylidyl blue [1-azo-2-hydroxy-3-(2,4-dimethylcarboxanilido)-naphthalene-1′-(2-hydroxybenzene)], binds magnesium in alkaline solution, causing a spectral shift and forming a red complex, measured around 600 nm. Protein interferences are reduced by dimethyl sulfoxide.

Atomic Absorption Spectrometry

As with calcium, AAS methods provide greater accuracy and precision for magnesium measurements than do photometric methods.

Enzymatic Methods

Enzymatic methods have been developed with hexokinase using Mg^{2+}-ATP as a substrate. The rate of the enzyme-catalyzed reaction is dependent on magnesium concentration. When hexokinase is used with glucose 6-phosphate dehydrogenase, the rate of the dehydrogenase reaction is monitored by measuring the formation of NADPH at 340 nm.

Measurement of Free (Ionized) Magnesium

Free magnesium measurements in whole blood, plasma, or serum use ISEs with neutral carrier ionophores. Current ionophores or electrodes have insufficient selectivity for magnesium over calcium. It is necessary to determine both ions simultaneously and to correct for Ca^{2+}. pH should be measured simultaneously, as protein binding of magnesium in plasma is pH dependent.

Discordance between total and free magnesium measurements has been reported in selected populations—cardiovascular disorders, diabetes mellitus, alcoholism, migraine headaches, asthma, renal transplant, head trauma, and pregnant women. Interferences in measurements of total magnesium, such as thiocyanate in smokers, may explain some discrepancies.

Specimen Requirements for Total and Free Magnesium

Serum and heparinized plasma are preferred specimens for measuring free and total magnesium. Zinc heparin, lithium-zinc heparin, and some newer heparins, developed for free calcium determination, should be avoided because they increase free magnesium. Anticoagulants, such as citrate, oxalate, and EDTA, are not acceptable because they form complexes with magnesium. Storage of serum for days at 4°C

and for months frozen does not affect the measured concentrations of total magnesium.

Serum or plasma must be separated as soon as possible to prevent an increase in magnesium due to cell leakage. Hemolyzed specimens are unacceptable. Interference by icterus or lipemia depends on the method and can be decreased by bichromatic analysis or blanking with EDTA.

Factors altering free calcium also alter free magnesium concentration. Specimens should be handled anaerobically to prevent loss of carbon dioxide and analyzed without delay to prevent changes in pH. Certain silicones or other tube additives as well as thiocyanate (smokers and diet) interfere with free magnesium.

Magnesium is primarily an intracellular ion, so depletion is not necessarily reflected in decreased plasma concentrations.

Urine specimens should be collected in acid to prevent precipitation of magnesium complexes. If acid must be added after collection, the entire specimen must be acidified, warmed, and mixed thoroughly before analysis.

Reference Intervals for Total and Free (Ionized) Magnesium

Adult reference intervals for total serum magnesium are 1.7 to 2.4 mg/dL (0.66 to 1.07 mmol/L). The reference interval is a matter of debate, as low concentrations within the interval are associated with cardiovascular risk. Conversion factors for the units used to express magnesium concentration are given below.

$$\text{mmol/L} = \text{mEq/L} \times 0.5 = \text{mg/dL} \times 0.41$$

$$\text{mEq/L} = \text{mmol/L} \times 2 = \text{mg/dL} \times 0.82$$

$$\text{mg/dL} = \text{mEq/L} \times 1.22 = \text{mmol/L} \times 2.43$$

Reference intervals for total magnesium in infants, children, and adolescents have been published; they do not differ significantly from those of adults. Pediatric reference intervals have been established as part of the CALIPER study (see Appendix).

HORMONES REGULATING BONE AND MINERAL METABOLISM

PTH and $1,25(OH)_2D$ are the primary hormones regulating bone and mineral metabolism. FGF23 is involved in phosphate homeostasis. Calcitonin has pharmacologic actions, but no obvious physiologic role. PTHrP is the principal mediator of **humoral hypercalcemia of malignancy (HHM)**, but it also has physiologic functions in fetuses, and in pregnancy and lactation. Sclerostin inhibits the Wnt signaling pathway that plays a major role in osteoblast development and function, and results in decreased bone formation and turnover.

Parathyroid Hormone

Parathyroid hormone (PTH) is synthesized and secreted by the parathyroid glands, usually two superior and two inferior, located bilaterally on or near the thyroid gland capsule. The chief cells are responsible for synthesizing, storing, and secreting PTH.

Biochemistry and Physiology

The concentration of PTH in blood is determined by the rates of synthesis and secretion by the parathyroids and its metabolism and clearance by the liver and kidneys. The primary regulators of PTH secretion are free calcium, $1,25(OH)_2D$, and phosphate. PTH acts directly on bone and the kidneys, and indirectly on intestine via $1,25(OH)_2D$, to increase plasma free calcium and ultimately decrease plasma phosphate.

Synthesis and secretion. Parathyroid cells have relatively few secretory granules for PTH storage. PTH must be synthesized as needed for secretion. Control of PTH synthesis by calcium and phosphate occurs largely at the post-transcriptional level.

PTH metabolism in the parathyroids results in the pre-sequence and the 6 amino acid long N-terminal 'pro-sequence' being enzymatically cleaved before packaging in the Golgi apparatus. Intact PTH (84 amino acids) is secreted, stored, or degraded intracellularly. Intracellular degradation is increased when plasma calcium is high and secretion of PTH is low. When PTH secretion is high (low plasma calcium), intracellular degradation is low. Together with intact PTH, several C-terminal fragments are secreted.

The classic biological activity of PTH resides in the N-terminal third of the molecule. Synthetic PTH(1–34) is at least as potent as intact PTH(1–84) in interacting with the PTH/PTHrP receptor (PTH1R, type 1 PTH receptor) and stimulating calcemic, phosphaturic, and other biological responses in kidneys and bone.

The C-terminal metabolites of PTH, have effects that antagonize those of PTH (1–84), having their own receptors in kidneys and bone, that are distinct from those mediated by PTH1R.

The concentration of free calcium in blood or extracellular fluid is the primary acute physiologic regulator of PTH synthesis, metabolism, and secretion. An inverse sigmoid relationship exists between PTH secretion and free extracellular calcium. Maximal secretion and suppression are attained with hypocalcemia and hypercalcemia. The midpoint of this relationship (or set point) is the calcium concentration at which PTH secretion is half maximal.

$1,25(OH)_2D$, phosphate, and magnesium influence the synthesis and secretion of PTH. $1,25(OH)_2D$ interacts with vitamin D receptors in the parathyroid glands to decrease PTH secretion. Hyperphosphatemia and hypophosphatemia increase and decrease PTH synthesis and secretion, respectively, and the hyperphosphatemia of CKD leads to parathyroid hyperplasia and hyperparathyroidism. Chronic severe hypomagnesemia, such as in alcoholism, has been associated with impaired PTH secretion, acute hypomagnesemia may stimulate secretion. Chronic hypomagnesemia can cause resistance to the effects of PTH.

PTH demonstrates a circadian rhythm throughout 24 hours. Variation in phosphate is important in controlling the circadian secretion of PTH.

Biological actions. PTH influences calcium and phosphate homeostasis directly through its actions on bone and kidneys and indirectly through actions on the intestine through $1,25(OH)_2D$ (see Fig. 39.3). PTH exerts its actions by interacting with type 1 PTH receptors located in the plasma membranes of target cells.

In the kidneys PTH induces 25-hydroxyvitamin D-1α-hydroxylase, increasing $1,25(OH)_2D$, which, in turn, stimulates intestinal absorption of both calcium and phosphate; it increases calcium reabsorption in the distal convoluted tubules; it decreases reabsorption of phosphate by the proximal tubules, and inhibits Na^+H^+ antiporter activity, which favors mild hyperchloremic metabolic acidosis in hyperparathyroidism.

The effects of PTH on bone are complex, as evidenced by its stimulation of bone resorption or bone formation, depending on the concentration of PTH, the duration of exposure and the PTH signaling profile. Chronic exposure to high concentrations of PTH leads to increased bone resorption that is an effect important for the maintenance of calcium homeostasis. Daily increase and decrease of PTH (seen following Teriparatide injection) results in an anabolic effect on bone that may be mediated by inhibition of sclerostin expression in osteocytes. PTH increases osteoblasts and enhances their differentiation from stromal cells.

Integration of the direct effects of PTH on bone and kidneys, and of the indirect effects on intestine through $1,25(OH)_2D$, results in alterations in calcium and phosphate concentrations in blood and urine. Urinary calcium excretion is increased because the larger filtered load of calcium, derived from bone resorption and intestinal calcium absorption, overrides the increased tubular reabsorption of calcium.

Heterogeneity of circulating parathyroid hormone. PTH circulates as the biologically active hormone and as a series of N-terminally truncated fragments containing the mid-region and C-terminal amino acids. The heterogeneity is due to the secretion of both intact hormone and C-terminal fragments by the parathyroids, peripheral metabolism of intact hormone by liver and kidneys to C-terminal fragments, and renal clearance of intact hormone and C-terminal fragments. Biologically active intact PTH (amino acids 1–84) is rapidly cleared from plasma in normal subjects (half-life <5 minutes) by the liver (60% to 70%) and the kidneys (20% to 30%).

Circulating PTH measured by most immunoassays is composed of both "inactive" fragments and intact hormone. Fragments consisting of the middle and carboxyl regions of the molecule (e.g., amino acids 34–84, 36–84) are devoid of the N-terminal region and classic PTH activity and were considered inactive degradation products.

Clinical Significance

Determination of PTH is useful in the differential diagnosis of both hypercalcemia and hypocalcemia, for assessing parathyroid function in renal failure, and for evaluating parathyroid function in bone and mineral disorders.

Measurement of Parathyroid Hormone

PTH assays have yielded different estimations due to the presence of a series of immunologic determinants along the molecule and the physiologic heterogeneity of circulating PTH-related antigens. Noncompetitive (sandwich) immunoassays are commonly used for the measurement of intact PTH (see Chapter 15 for immunoassay discussion).

PTH assays developed to measure the intact molecule of PTH(1–84) only, were called *intact assays.* These assays incorporated the use of a noncompetitive immunometric assay (IMMA) format using two antibodies: (1) a solid-phase capture antibody against the C-terminus, and (2) a labeled detection antibody against the N-terminus, or vice versa. It has become clear that they can overestimate the severity of PTH-related bone disease because they detect several molecular species of PTH. Blood contains N-terminally truncated PTH fragments that are long enough to react in the intact PTH assays but do not bind the PTH receptor. The term *intact* as applied to these immunometric PTH assays is a misnomer because C terminal PTH fragments are detected along with PTH(1–84).

In the newest PTH assays, the epitope of the N-terminally binding antibody consists of the first four to six amino acids of the PTH molecule. Such assays have been variously termed *true intact* PTH assays, *whole* PTH assays, or *cyclase-activating* PTH assays. Despite theoretical improvement, in clinical practice they have not shown advantages over earlier IMMAs. A few methods have solved the problem, using capture antibodies against the N-terminal sequence (26–32) and signal antibodies against the middle or C-terminal sequence (55–64). With the latter arrangement, a higher quantity of signal antibody is needed because of the high concentration of inactive fragments relative to intact hormone, especially in CKD.

Affinity-purified polyclonal antibodies have been widely used, particularly for the signal antibody, because of the difficulty of producing high-affinity monoclonal PTH antibodies. Capture antibodies are noncovalently bound or are covalently attached to one of a variety of solid phases. Signal antibodies most often have been radiolabeled (IRMAs) with 125I, or labeled with a chemiluminescent compound (ICMAs), such as acridinium ester or isoluminol, or an enzyme, ELISA or EIA, such as alkaline phosphatase (ALP). Other assay formats used on automated immunoassay analyzers include electrochemiluminescence detection (see Chapter 15).

N-terminal-truncated parathyroid hormone and intact parathyroid hormone methods. The N-terminal-truncated fragment(s) [non-(1–84)PTH and PTH(7–84)], containing all but a few of the N-terminal amino acids, account, on average, for 20% to 50% of intact PTH concentrations measured by early intact PTH assays in healthy subjects and CKD. The relative concentration of N-terminal-truncated PTH increases with hypercalcemia and decreases with hypocalcemia. N-terminal-truncated PTH originates from both secretion by the parathyroids and peripheral metabolism of intact hormone.

The newest intact PTH assay utilizes antibodies against PTH(1–4) and requires the presence of the most N-terminal amino acid, evidenced by its failure to recognize PTH(2–34) (Fig. 39.5).

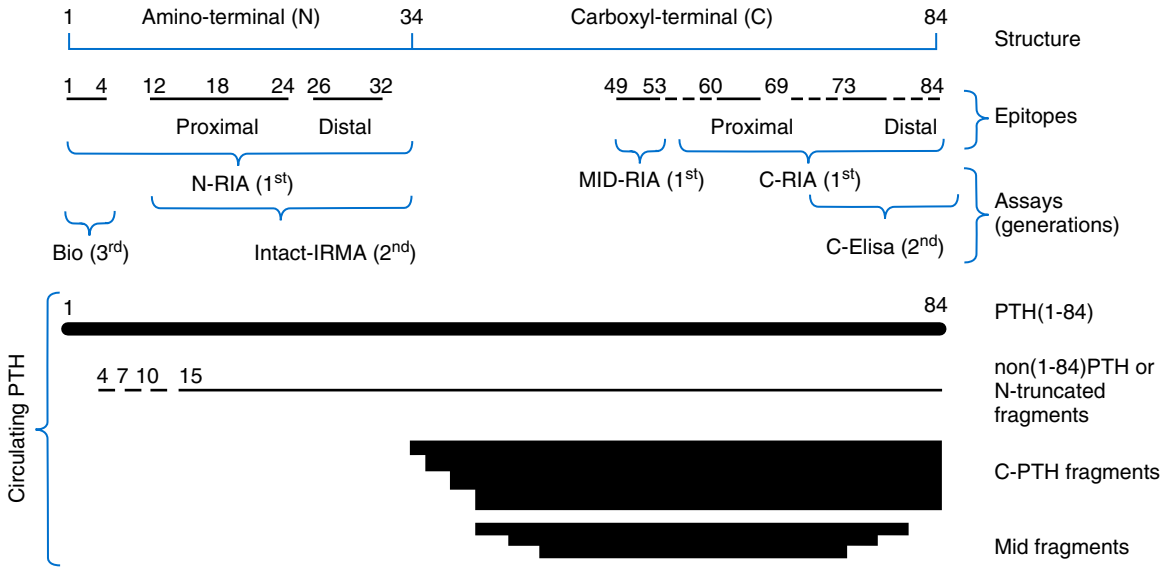


Fig. 39.5 Relationship between parathyroid hormone *(PTH)* assays, PTH assay epitopes, and PTH molecular forms detected in circulation. The *upper panel* depicts the structure of human PTH and epitopes detected by various PTH assays. Early PTH assays detect full-length PTH(1–84) in addition to PTH fragments. These assays include RIAs that use antisera specific for the amino-terminal (N-RIA), middle (MID-RIA), or carboxy-terminal (C-RIA) region of PTH. "Intact PTH" assays detect full-length PTH(1–84) and non-(1–84)PTH fragments. PTH assays (Bio) detect the full-length PTH(1–84). The *lower panel* depicts the PTH molecular forms present in the circulation. (Modified from Henrich, L. M., Rogol, A. D., D'Amour, P., Levine, M. A, Hanks, J. B., & Bruns, D. E. [2006]. Persistent hypercalcemia after parathyroidectomy in an adolescent and effect of treatment with cinacalcet HCl. *Clin Chem*, *52*, 2286–2293.)

The inability of some intact PTH methods to accurately measure biologically active intact hormone may explain why such intact PTH assays are not a reliable indicator of bone turnover in dialysis patients and fail to distinguish patients with low-, normal-, and high-turnover bone disease.

The newest intact PTH assays (whole PTH) are not totally specific for intact PTH, the presence of an amino-terminal form of PTH represents approximately 20% of the intact PTH measured with some new intact PTH methods. The clinical implications of this form have not yet been established.

Several studies highlight the lack of comparability of current PTH assays especially in patients with CKD or those receiving hemodialysis. This has raised concerns regarding the standardization of PTH assays; therefore recognized production and implementation of a commutable international standard material is required.

Mass spectrometry has been used to characterize PTH fragments and to measure intact PTH. The assay showed no detectable cross-reactivity with PTH fragments, including PTH(7–84). Using this method PTH(7–84) was not detected in the serum of CKD patients although other large C-terminal fragments were present.

Specimen requirements. EDTA plasma is often preferred, some assays require serum and results vary with sample type. Plasma should be analyzed within 72 hours if stored at 4°C. Lower concentrations of PTH are observed in serum incubated at room temperature for longer than a few hours, or after one to several days at 4°C. There is no consensus on the effects of storing samples in a freezer (−20°C or −80°C) before analysis. PTH can stick to some plastic tubes and result in decreased values obtained with storage.

Reference Intervals

Reference intervals for PTH vary with the method used and can be affected by vitamin D status of the population studied.

Typical reference intervals are:

Intact PTH: 10 to 65 pg/mL or 1.1 to 6.8 pmol/L

PTH(1–84): 6 to 40 pg/mL or 0.6 to 42 pmol/L

Interpretation of PTH concentrations must take account of the patient's circulating calcium at the time of sampling.

Lower intervals have been established by excluding individuals with vitamin D insufficiency.

Intact PTH increases with age; is low or normal during pregnancy; is lower in fetuses and umbilical cord blood; and increases during the first few days of life. Concentrations in children and adolescents are similar to adults.

Interpretation of Parathyroid Hormone Results

PTH is the most important test for the differential diagnosis of hypercalcemia. Intact PTH is increased in most patients with PHPT (Fig. 39.6) and is below normal or close to the lower limit of the reference interval in most patients with nonparathyroid hypercalcemia. In stable hypercalcemia, increased or inappropriately detectable PTH in patients with hypercalcemia and malignancy suggests coexisting PHPT and malignancy, ectopic PTH production by tumors is extremely rare. Measurements of

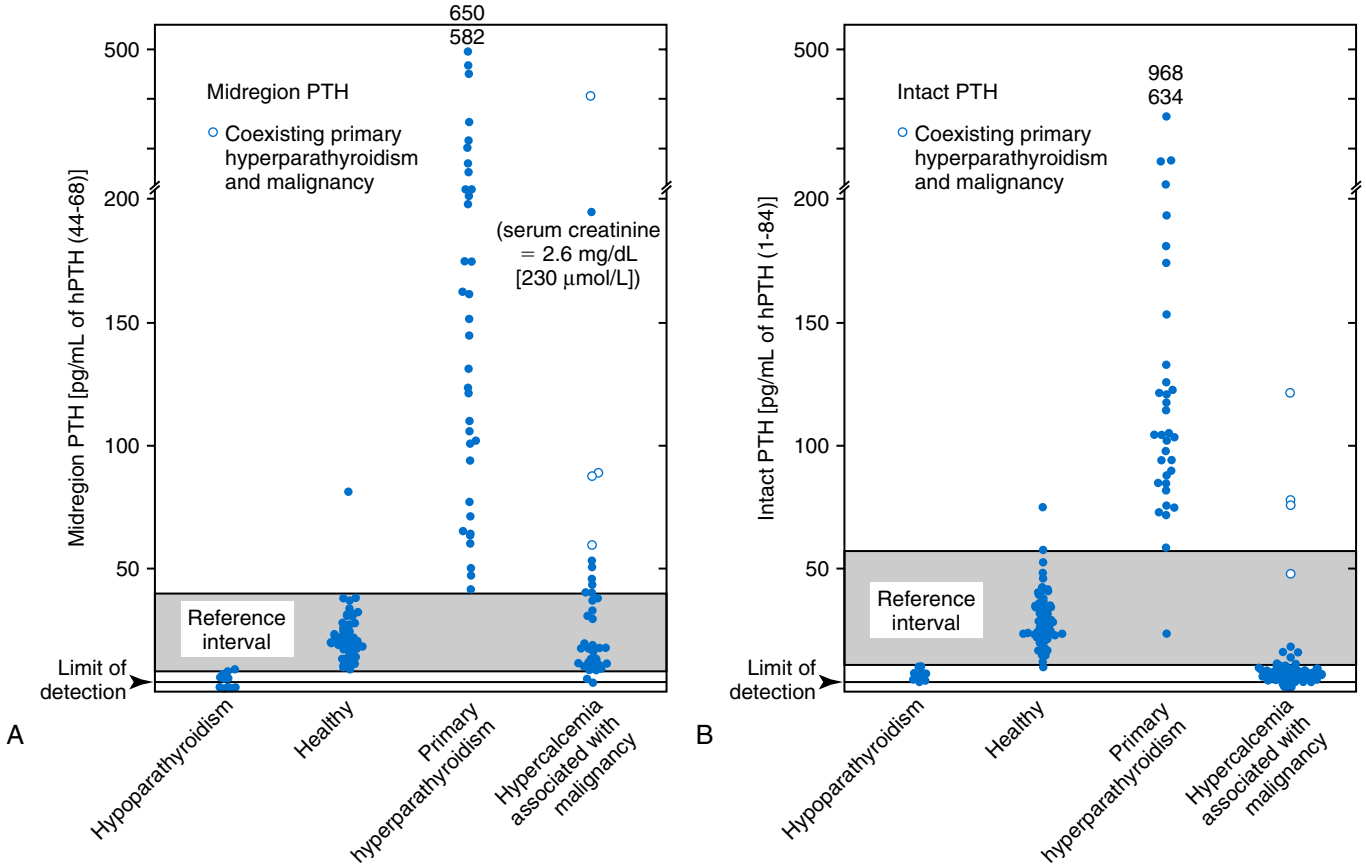


Fig. 39.6 Parathyroid hormone *(PTH)* in healthy individuals and patients with primary hyperparathyroidism, hypercalcemia-associated malignancy, and hypoparathyroidism. PTH was measured with immunoassay for its mid-region (A) and for intact PTH (B). (Modified from Endres, D. B., Villanueva, R., Sharp, C. F., Jr., & Singer, F. R. [1989]. Measurement of parathyroid hormone. *Endocrinol Metab Clin North Am, 18,* 611–629.)

urinary calcium excretion are necessary to confirm hypocalciuria in cases of suspected FBHH.

The short half-life of PTH (≤5 minutes) allows intraoperative determination of intact PTH to assess the completeness of parathyroidectomy and facilitates minimally invasive parathyroidectomy. Preoperative or intraoperative PTH may be useful for localizing hyperfunctioning parathyroid tissue by sampling multiple veins from the cervical and mediastinal regions. Postoperative PTH concentrations after thyroidectomy may be predictive of development of hypocalcemia due to a large influx of calcium into bone (hungry bone syndrome).

Subnormal or low-normal PTH is observed in most patients with hypoparathyroidism (see Fig. 39.6). The apparently detectable concentrations in patients with hypoparathyroidism or nonparathyroid hypercalcemia may be a result of imprecision of methods at low concentrations, a nonspecific serum (ligand-free matrix) effect, and/or measurement of N-terminal-truncated PTH. In 2°HPT, PTH is increased before total or free calcium becomes abnormally low—a consequence of homeostatic mechanisms for the maintenance of serum calcium.

In patients with end-stage renal disease (ESRD), the measurement of PTH is helpful in assessing parathyroid function, in estimating bone turnover, and in improving management. Parathyroid status is usually determined by measuring PTH on predialysis specimens because factors, including changes in plasma calcium and the type of dialysis membrane, affect PTH secretion and clearance.

The whole PTH(1–84) assay theoretically should more accurately assess parathyroid status and bone turnover and thus better guide therapy in patients with ESRD compared with the intact PTH assay.

POINTS TO REMEMBER

Parathyroid Hormone

- Several assays exist for the measurement of PTH. Cross reactivity with PTH metabolites and standardization remains a current problem resulting in significant variation in values obtained particularly in CKD.
- Measurement of intact PTH is pivotal in the differential diagnosis of calcium disorders.
- Surgical removal of the parathyroid glands is the commonest cause of hypoparathyroidism.
- Hypomagnesemia can result in failure of the chief cells to release PTH resulting in hypoparathyroidism. In such cases, patients will not normalize their circulating calcium until there is adequate magnesium replacement.
- PTH by daily injection is a treatment for osteoporosis.

Vitamin D and Its Metabolites

Vitamin D_3 is produced endogenously through exposure of skin to sunlight (cholecalciferol) and is absorbed from foods containing or supplemented with vitamin D_3 or D_2 (ergocalciferol). The vitamin is metabolized to its biologically active form, $1,25(OH)_2$ D_3/D_2, promoted by hormones that regulate calcium and phosphate metabolism. Deficiency of vitamin D results in impaired formation of bone, producing rickets in children and osteomalacia in adults.

Biochemistry and Physiology

Vitamin D and its metabolites may be categorized as cholecalciferols or ergocalciferols (Fig. 39.7). Cholecalciferol (vitamin D_3) is the parent compound and is produced in skin from 7-dehydrocholesterol on exposure to UV B from sunlight. Latitude, season, aging, sunscreen use, and skin pigmentation influence the production of vitamin D_3 by skin. Vitamin D_2 is manufactured by irradiation of ergosterol produced by yeasts. Total vitamin D refers to vitamin D or metabolites written without a subscript.

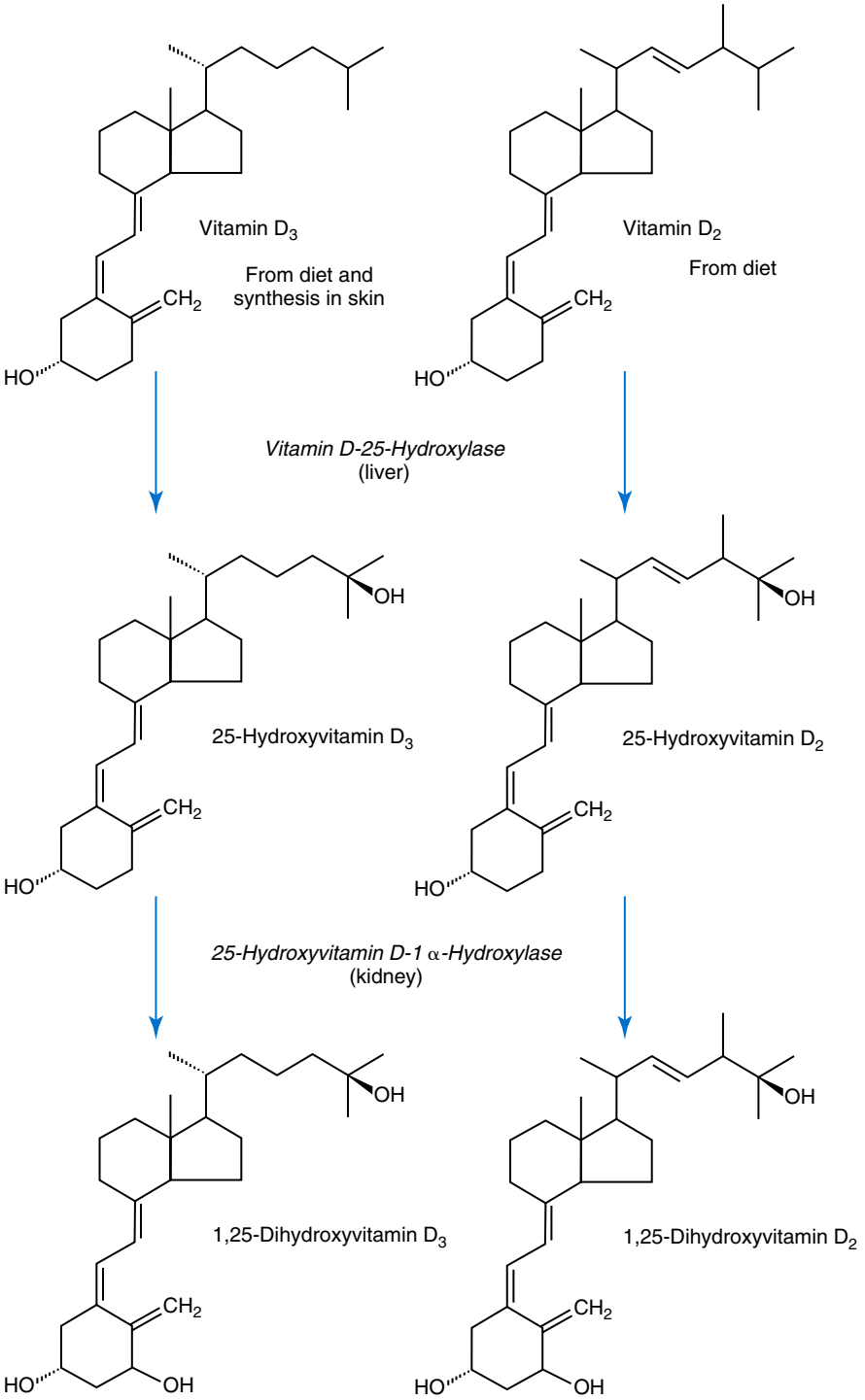


Fig. 39.7 Metabolism of vitamin D. Vitamin D2 and vitamin D3 are enzymatically hydroxylated to 25-hydroxyvitamin D in the liver and to 1,25-dihydroxyvitamin D by the kidneys.

A total of 90% of vitamin D is produced by synthesis in the skin. In North America, a significant fraction of vitamin D is acquired by ingestion of fortified foods (cereals, bread, and milk) or supplements. The recommended daily allowance is 400 IU (10 μg). In some European countries, the recommendation for people older than 60 years is 800 IU (20 μg). There is debate about the doses of vitamin D that are considered optimal for bone health and that can improve vitamin D status sufficiently to result in increases in bone mineral density (BMD) and reduce risk of fracture. Prospective randomized trials of high-dose vitamin D supplementation have increased BMD and muscle function in women, but in some cases a detrimental effect has been observed resulting in increased falls and fractures. Increasing the dose of vitamin D supplementation results in a greater proportional production of $24,25(OH)_2D$ and this may block the action of $1,25(OH)_2D$.

Metabolism, Regulation, and Transport

Vitamin D_2 and vitamin D_3 are metabolized to 25(OH)D⅔ in the liver by vitamin D 25-hydroxylase, (see Fig. 39.7). The concentration and half-life of 25(OH)D in plasma are listed in Table 39.3.

25(OH)D⅔ are metabolized to 1,25-dihydroxyvitamin D_2 or D_3, the biologically active hormone, by 25(OH)D 1α-hydroxylase, in the kidneys and placenta (see Fig. 39.7). Circulating $1,25(OH)_2D$ is tightly regulated, by PTH, phosphate, calcium, FGF23 and $1,25(OH)_2D$. PTH and hypophosphatemia increase 25(OH)D-1α-hydroxylase activity; hypocalcemia stimulates PTH secretion. Hypercalcemia, hyperphosphatemia, FGF23, and $1,25(OH)_2D$ reduce 25(OH)D 1α-hydroxylase. $1,25(OH)_2D$ also induces 25(OH)D 24-hydroxylase, producing $24,25(OH)_2D$. Normal concentrations and the half-life of $24,25(OH)_2D$ are listed in Table 39.2.

Vitamin D, 25(OH)D, and $1,25(OH)_2D$ are bound to vitamin D binding protein (DBP), a specific, high-affinity transport protein also known as *group-specific component of serum* or *Gc-globulin*. DBP is constitutively synthesized by the liver and circulates in excess (400 mg/L), with <5% of the vitamin D binding sites occupied. Only 0.03% of 25(OH)D and 0.4% of $1,25(OH)_2D$ are free in plasma (see Table 39.2). DBP concentrations are increased in pregnancy and estrogen therapy, and are decreased in nephrotic syndrome.

Biological Actions of 1,25-Dihydroxyvitamin D

$1,25(OH)_2D$ maintains calcium and phosphate concentrations in blood through its actions on intestine, bone, kidney, and the parathyroids. In the small intestine, $1,25(OH)_2D$ stimulates calcium absorption, primarily in the duodenum, and phosphate absorption by the jejunum and ileum. Three processes absorb calcium from the diet: (1) calcium entry into the brush border cytoplasm, (2) diffusion of calcium within the cell, and (3) exit of calcium from the cell across its basolateral membrane.

At high concentrations, $1,25(OH)_2D$ increases bone resorption by inducing monocytic stem cells to differentiate into osteoclasts and by stimulating osteoblasts to produce cytokine production influencing osteoclast activity. By stimulating osteoblasts, $1,25(OH)_2D$ increases ALP and **osteocalcin (OC)**. $1,25(OH)_2D$ acts directly on the parathyroids to inhibit the synthesis and secretion of PTH.

Clinical Significance

Determination of 25(OH)D may be useful in the differential diagnosis of hypocalcemia, hypercalcemia, or hypercalciuria and for evaluating vitamin D status in health and in bone and mineral disorders. Measurements of 25(OH)D, $1,25(OH)_2D$ and $24,25(OH)_2D$ have clinical value. The (UK) National Kidney Foundation guidelines have emphasized that vitamin D nutrition is assessed by measurements of 25(OH)D, not $1,25(OH)_2D$.

Vitamin D nutritional status is best determined by measurement of 25(OH)D, because 25(OH)D is the main circulating form of vitamin D (Table 39.3), it has a long half-life and is less affected by day-to-day variation, exposure to sunlight, or food intake. Groups at higher risk for developing nutritional vitamin D deficiency include breast-fed infants, strict vegetarians who abstain from eggs and milk, strict vegans, individuals with darker skin pigmentation, and the elderly.

Concentrations of 25(OH)D may be decreased by reduced availability of vitamin D, inadequate conversion of vitamin D to 25(OH)D, accelerated metabolism of 25(OH)D, and urinary loss of DBP and 25(OH)D. Reduced availability of vitamin D may occur with inadequate exposure to sunlight, dietary deficiency, malabsorption syndromes, and following gastric or small bowel resection. Severe hepatocellular disease has been associated with inadequate conversion of vitamin D to 25(OH)D. Drugs such as phenytoin and phenobarbital induce drug-metabolizing enzymes that accelerate the metabolism of vitamin D and its metabolites. Plasma 25(OH)D concentrations may be reduced in patients with nephrotic syndrome because of the urinary loss of DBP and 25(OH)D.

A significant proportion of adults in Europe and North America have poor vitamin D status. It has been difficult to harmonize reference intervals for 25(OH)D, this may be related to differences of opinion regarding the definition of subclinical vitamin D insufficiency.

TABLE 39.3 Vitamin D and Its Metabolites in Plasma

Compound	Concentration	Free (%)	Half-Life
Vitamin D	<0.2–20 ng/mL <0.5–52 nmol/L	—	1–2 days
25-Hydroxyvitamin D	10–65 ng/mL 25–162 nmol/L	0.03	2–3 weeks
1,25-Dihydroxyvitamin D	15–60 pg/mL 36–144 pmol/L	0.4	4–6 h
24,25-Dihydroxyvitamin D	0.8–2.8 ng/mL 2.0–7.2 nmol/L	Not established	NA

In hypercalcemia, 25(OH)D measurement has limited value. It can confirm intoxication after ingestion of large amounts of vitamin D or 25(OH)D. The measurement of $1,25(OH)_2D$ is diagnostic in vitamin D–dependent rickets types 1 and 2, and is helpful in disease states associated with the overproduction of $1,25(OH)_2D$ such as granuloma-forming disorders, fungal infection, Hodgkin disease, and other lymphomas. $1,25(OH)_2D$ gives confirmatory information in the evaluation of hypercalcemia, hypercalciuria, hypocalcemia, and bone and mineral disorders. $1,25(OH)_2D$ concentrations are increased in vitamin D-dependent rickets type 2, in $1,25(OH)_2D$ intoxication, and in PHPT. Reduced $1,25(OH)_2D$ can be observed in renal failure, hypercalcemia of malignancy, hyperphosphatemia, hypoparathyroidism, pseudohypoparathyroidism, type 1 vitamin D-dependent rickets, hypomagnesemia, nephrotic syndrome, and severe hepatocellular disease.

$24,25(OH)_2D$ measurements are increasingly valuable. The absence of $24,25(OH)_2D$ can result in childhood hypercalcemia and stone formation. Some reports of hypercalcemia in pregnancy have described low $24,25(OH)_2D$ contributing to the pathology. Higher 25(OH)D concentrations have a strong positive correlation with $24,25(OH)_2D$, and increasing vitamin D supplementation results in proportionally greater increases in $24,25(OH)_2D$ than $1,25(OH)_2D$.

Measurement of Vitamin D Metabolites

Specific and sensitive assays exist for measuring vitamin D, 25(OH)D, $1,25(OH)_2D$, and 24,25 $(OH)_2$ D. All assays should measure D_2 and D_3 metabolites equally (equimolar reactivity), because both D_2 and D_3 are metabolized to produce biologically active $1,25(OH)_2D$. Separate measurement of the D_2 and D_3 does not necessarily distinguish between dietary and endogenous sources of vitamin D, as food can be supplemented with D_3 or D_2.

The hydrophobic natures and small circulating concentrations of vitamin D metabolites, together with the possibility of matrix effects, have presented challenges for the development of assays that would be both specific and suitable for high-throughput clinical analysis. Two or three of the following steps are needed in any assay for 25(OH)D, $1,25(OH)_2D$ or $24,25(OH)_2D$: deproteinization or extraction, purification, and quantification. The first step frees the metabolites from DBP. Purification steps, most often using column chromatography, separate the forms of vitamin D, lipids, and interfering substances. The method of quantification depends on the metabolite being measured.

Extraction and deproteinization. Protein precipitation using zinc sulfate is currently a common step in sample preparation. The most widely used methods use acetonitrile to deproteinize the specimen and to denature DBP, freeing vitamin D metabolites. One assay uses immunoextraction for $1,25(OH)_2D$, and this process has been incorporated into an automated method.

Column chromatography. Differences in their polarities, related to their numbers of hydroxyl groups, have been used to separate vitamin D and its metabolites. $1,25(OH)_2D$ and $24,25(OH)_2D$ are more polar than 25(OH)D, which is more polar than vitamin D.

Measurement of 25-hydroxyvitamin D. Acceptance of serum 25(OH)D as an indicator of an individual's vitamin D status has resulted in the development of several assays. The metabolite is measured by competitive protein binding assay (CPBA), competitive or noncompetitive immunoassay (radioimmunoassay [RIA], enteroinsular axis enzyme assay [EIA], immunochemiluminometric assay [ICMA]), UV absorption after separation by high-performance liquid chromatography (HPLC), or by mass spectrometry after separation by chromatography (see Chapter 13). The 25(OH) D_2 and D_3 metabolites can be quantified separately by liquid chromatography-mass spectrometry (LC-MS)/MS. It is important to report the sum of the two concentrations $[25(OH)D_2 + 25(OH)D_3]$ because vitamin D status depends on the total concentration.

Determination by HPLC with UV absorption or mass spectrometry requires appropriate equipment (see Chapters 12 and 13), specialized training, and a larger sample.

Most laboratories that measure 25(OH)D have chosen competitive immunoassays. These assays are easy to perform; use widely available reagents on multi-channel analyzers; and require only a small volume of plasma.

Measurement of 1,25-dihydroxyvitamin D. $1,25(OH)_2D$ circulates at 1/1000 of the concentration of 25(OH)D and at lower concentrations than other dihydroxylated metabolites (see Table 39.3). Together with the extreme hydrophobicity and instability of the compound this greatly complicates its determination in serum. Widely used methods require extraction, chromatography, and quantification by radioreceptor assay or RIA.

Measurement of 24,25 dihydroxyvitamin D. No immunoassays have sufficient antibody specificity that enable measurement of $24,25(OH)_2D$. The methods currently in use are LC-MS/MS assays.

Specimen requirements. Serum is preferred for measuring vitamin D metabolites, although plasma is acceptable for assays using extraction and chromatography. Once separated from the clot, serum is stable at both room temperature and 4°C; specimens should be frozen if the analysis is delayed. Vitamin D metabolites in serum or plasma are not sensitive to light and do not require special handling.

Reference Intervals

Reference intervals for vitamin D metabolites are method dependent. Concentrations of less than 20 to 30 ng/mL (<50 to 75 nmol/L) can be associated with increased PTH and reduced calcium absorption. The Institute of Medicine published a recommendation that the lower limit for vitamin D sufficiency be lowered from 30 to 20 ng/mL (75 to 50 nmol/L) and that concentrations greater than 50 ng/mL (>125 nmol/L) be considered as cause for concern. Vitamin D insufficiency [25(OH)D <20 ng/mL (50 nmol/L)] was common during the winter in adults (>30 years) Caucasian men (15%) and women (30%) living in the southern United States.

25(OH)D concentrations in African Americans are lower than in non-Hispanic whites, but the clinical significance of this is uncertain. Compared with non-Hispanic whites, African Americans have higher BMD, higher PTH and $1,25(OH)_2D$; lower biochemical markers of bone remodeling; and lower rates of osteoporosis.

25(OH)D is increased by exposure to sunlight and shows seasonal variation, with the highest concentrations in summer or fall and the lowest in winter or spring. It is recommended that testing be done at the end of winter; if results are above the cutoff, the subject is likely to have adequate concentrations throughout the year. Concentrations are influenced by latitude, sunscreen use, and skin pigmentation.

Vitamin D metabolite concentrations vary with age and are increased in pregnancy. Concentrations of $1,25(OH)_2D$ are higher in children than in adults, and are highest during periods of greatest growth. Although 25(OH)D and $1,25(OH)_2D$ decrease with age, this decline may be a consequence of poor nutrition, reduced exposure to sunlight, and declining health. Concentrations of metabolites were unchanged with age in studies limited to healthy, active subjects.

$24,25(OH)_2D$ concentrations are method dependent. Ethnic differences have been reported, with black Americans having lower concentrations than Caucasians. Increasing supplementation with vitamin D results in increasing $24,25(OH)_2D$ concentrations.

POINTS TO REMEMBER

Vitamin D

- Measurement of 25 hydroxyvitamin D (25OHD) is accepted as the best reflection of vitamin D status.
- Tandem mass spectrometric methods are considered the best way of measuring 25OHD.
- A high percentage of the world's population are vitamin D deficient.
- Prolonged deficiency of vitamin D results in rickets in children and osteomalacia in adults.
- Care is required with high-dose vitamin D treatment as it can result in significant morbidity, increasing falls, and fractures in elderly women.

Fibroblast Growth Factor 23

Biochemistry and Physiology

Fibroblast growth factor 23 (FGF23) promotes phosphate excretion in urine. FGF23 consists of 251 amino acids, is present in embryonic and adult tissues, and has high expression in osteoblasts and osteocytes.

Biological Actions and Clinical Significance

FGF23 is involved in the medium- to long-term regulation of circulating phosphate. It acts directly on the kidney to decrease renal phosphate reabsorption.

FGF23 is increased in plasma in CKD and a variety of other conditions including X-linked hypophosphatemia (XLH). FGF23 may be the best measurement at predicting deterioration in CKD. It is unknown whether FGF23 is directly responsible for the decreased concentrations of $1,25(OH)_2D$ seen in CKD; FGF23 is markedly increased in CKD. The mechanism regulating FGF23 production in CKD remains unclear.

Measurement of FGF23

A single step IMMA incorporates an antibody to the C-terminal peptide fragments of FGF23 (cFGF23) that also reacts with the intact molecule (iFGF23) is available. This assay provides a combined estimate of both cFGF23 and iFGF23. Multiple other immunoassays for the measurement of the intact molecule are available. These assays can give different results in the same populations and so care is required when interpreting the results obtained. Reference intervals are considerably different for each assay. Pre-analytical factors are of importance and the sample type differs between assays. Users are advised to follow instructions provided by manufacturers.

Reference Intervals

There is significant variability in the reference intervals quoted. In healthy adults, the reference ranges are:
iFGF23: 11.7 to 48.6 pg/mL (Kainos)
cFGF23: 21.6 to 91.0 RU/mL (Immutopics)

Interpretation

Very high FGF23 concentrations can be obtained in XLH and CKD. An FGF23 concentration in the upper quartile of the reference interval may be inappropriate for a prevailing low phosphate and may require further investigation.

Calcitonin

The release of **calcitonin** from the parafollicular or C cells of the thyroid gland is stimulated by circulating calcium. Calcitonin has been used pharmacologically as an inhibitor of bone resorption, but the physiologic role of endogenous calcitonin is uncertain. Calcitonin measurements have a role in the diagnosis and follow-up of medullary thyroid carcinoma (MTC), a malignant tumor of the C cells.

POINTS TO REMEMBER

FGF23

- Osteocytes are major producers of FGF23.
- Several metabolic bone diseases result from abnormalities of FGF23 secretion and metabolism.
- FGF23 acts at the kidneys to promote phosphate excretion. Klotho is required to allow specific changes in FGF receptors in the kidney permitting FGF23 selective activity.
- Metabolism of FGF23 results in the generation of inactive C-terminal fragments that are measure in some assays.

FGF23 inhibits alpha hydroxylase in the kidney resulting in a decrease in the circulating concentration of 1,25-dihydroxyvitamin D.

Biochemistry and Physiology

Calcitonin is a 32 amino acid peptide, molecular mass 3418 Da, an N-terminal disulfide bond links the cysteine residues

1 and 7, and a C-terminal proline-amide. The hormone interacts with a specific G-protein-coupled receptor found on fully differentiated osteoclasts. The structures necessary for biological function are the C-terminal portion of the molecule.

Pharmacological doses of calcitonin decrease plasma calcium and phosphate primarily by inhibiting osteoclastic bone resorption. Higher concentrations of calcitonin are observed in young children, and during pregnancy and lactation. The hormone may protect the skeleton during periods of calcium stress.

Multiple forms of calcitonin can be observed both in healthy individuals and in patients with MTC or nonthyroidal malignancies. Much of the immunoreactive calcitonin in the circulation is larger than the monomeric hormone.

During the biosynthesis of calcitonin, the original product is a larger precursor form known as procalcitonin.

Clinical Significance

MTC occurs as a sporadic disease and as part of the multiple endocrine syndromes MEN2A, MEN2B, and familial MTC. These account for 5% to 10% of thyroid malignancies, with sporadic MTC accounting for about 80%. After thyroidectomy in MTC patients, calcitonin can serve as a tumor marker. The use of calcitonin measurements in oncology is discussed Chapter 20. Calcitonin is also increased in various nonthyroidal cancers, and other nonmalignant conditions including acute and chronic renal failure, hypercalcemia, hypergastrinemia, pulmonary disease, and severe illness.

Measurement of Calcitonin

Calcitonin is measured by immunoassay, interpretation can be complicated both by the heterogeneity of the circulating antigen and by the varying characteristics of the assays. Some assays may give results that are erroneously low at high concentrations (hook effect).

Reference Intervals

Reference intervals for calcitonin are method dependent. They should be determined for healthy and athyroidal individuals, by gender, for basal and stimulated (calcium and pentagastrin provocation tests) conditions.

An upper limit for the reference interval for basal calcitonin concentration in adults is 10 pg/mL (2.9 pmol/L) the mean concentrations are lower in women, 5.8 pg/mL (1.7 pmol/L) and higher 8.8 pg/mL (2.6 pmol/L) in men. Age, growth, pregnancy, and lactation—in addition to gender—have been reported to influence circulating calcitonin. Reference intervals are not given by assay manufacturers for groups other than healthy adults.

Parathyroid Hormone–Related Protein

PTHrP was discovered as a molecule responsible for causing HHM. It later proved to be an autocrine/paracrine factor with a multitude of functions in several organ systems. Among these are its effects on chondrocyte biology and endochondral bone formation, as well as on calcium metabolism in the fetus, and during pregnancy and lactation.

Biochemistry and Physiology

PTHrP is derived from a gene on chromosome 12 that is distinct from the *PTH* gene on chromosome 11. The N-terminal shows close homology to PTH and its PTH-like activity is contained within the N-terminal. Like PTH, PTHrP causes hypercalcemia and hypophosphatemia, and increases urinary cyclic adenosine monophosphate.

PTHrP exerts endocrine effects on skeletal development and calcium homeostasis during fetal life and lactation. Breast milk contains high PTHrP as does amniotic fluid. PTHrP regulates transepithelial calcium transport; it is a potent smooth muscle relaxant, and it regulates growth, differentiation, and development. The role of PTHrP in many tissues is unknown.

Clinical Significance

HCM is the third most frequent cause of hypercalcemia in hospitalized patients caused by HHM and local osteolysis. HHM is present in approximately 75% to 80% of patients with HCM. HHM is common in patients with squamous (lung, head and neck, esophagus, cervix, vulva, and skin), renal, bladder, and ovarian carcinomas. Hypercalcemia due to skeletal metastases and local osteolysis are common in breast cancer, multiple myeloma, lymphoma, and hematologic malignancies.

PTHrP is rarely required for diagnosis because HHM nearly always occurs in advanced disease when the diagnosis is clear. The need for PTHrP determination is greater when it becomes important in prognosis, selection of therapy, or monitoring.

Measurement of Parathyroid Hormone-Related Protein

Several competitive RIAs measure N-terminal, mid-region, and C-terminal. The N-terminal assays have been used most widely.

Specimen requirements. PTHrP is unstable in serum and plasma at 4°C and at room temperature unless collected in the presence of protease inhibitors and kept on ice. EDTA plasma should be promptly separated and frozen as red cell enzymes degrade PTHrP.

Reference Intervals

PTHrP concentrations are dependent on the assay and collection, varying from undetectable up to 5 pmol/L.

Interpretation

PTHrP is increased in 50% to 90% of patients with HCM (Fig. 39.8). In addition to squamous cell carcinoma of the lung, head and neck, esophagus, cervix, and skin, increased concentrations have been found in a wide variety of other malignancies.

Sclerostin. Sclerostin is secreted by osteocytes and articular chondrocytes, and absence favors bone formation.

Biochemistry and physiology. Sclerostin expression by osteocytes is regulated by mechanical forces. Immobilization increases sclerostin positive osteocytes and circulating sclerostin.

Clinical significance. Sclerostin is increased in hypoparathyroidism, type 2 diabetes, cancer-induced bone disease

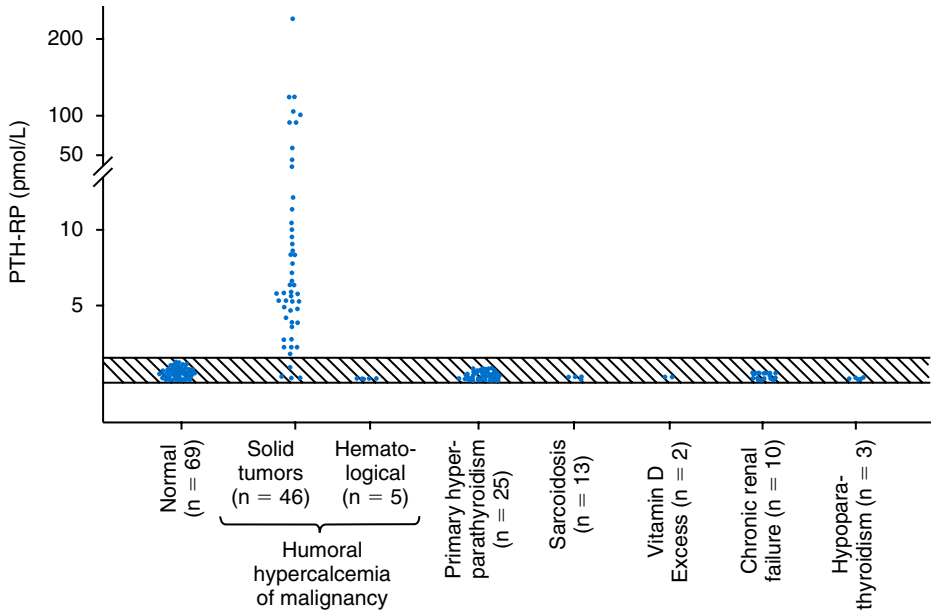


Fig. 39.8 Parathyroid hormone-related protein *(PTHrP)* in healthy individuals, patients with malignancies and other disorders. The *hatched area* indicates the reference interval. (Modified from Pandian, M. R., Morgan, C. H., Carlton, E., & Segre, G. V. [1992]. Modified immunoradiometric assay of parathyroid hormone-related protein: clinical application in the differential diagnosis of hypercalcemia. *Clin Chem*, *38*, 282–288.)

and **Paget disease of bone** (PDB). Sclerostin is decreased in 1HPT and ankylosing spondylitis.

Measurement of sclerostin. Several assays are available. EDTA–plasma and serum are the recommended sample types. Significant differences exist between commercial assays.

Reference intervals. Using the Biomedica assay in normal adults, the mean concentrations are:

Serum 0.8 pmol/L; plasma (EDTA) 1.3 pmol/L, plasma (heparin) 1.2 pmol/L, plasma (citrate 1.4 pmol/L).

BIOCHEMICAL MARKERS OF BONE TURNOVER

Markers can measure both bone formation and degradation. Markers of bone formation include products of the osteoblast, the amino- or carboxy-terminal propeptides of type I collagen, **procollagen type I N-terminal propeptide (PINP)**, and **procollagen type I carboxy-terminal propeptide (PICP)**, **bone alkaline phosphatase (bone ALP)**, and OC. Markers of bone resorption are breakdown products of type I collagen and include the amino-terminal and carboxy-terminal crosslinked telopeptide parts of collagen (**N-telopeptide [NTX]**, C-telopeptide [**CTX**], and C-telopeptide of type I collagen [ICTP]) and the **pyridinium crosslinks** (pyridinoline [PYD] and **deoxypyridinoline [DPD]**). TRAP5b is an enzyme released from resorbing osteoclasts. Osteoprotegerin (OPG) and receptor activator for nuclear factor kappa B (RANK) ligand assays are available—osteoblast proteins that influence osteoclast formation, activation, and bone resorption.

Markers of Bone Formation

Bone formation markers can be divided into three families, each reflecting a specific phase in development of the osteoblast. During each remodeling cycle, osteoblasts pass through the phases of proliferation, matrix maturation, and mineralization (see Fig. 39.2). Markers can give discordant results if bone formation is affected by a disease that targets a particular developmental phase of the osteoblast.

Propeptides of Type I Procollagen (Procollagen Type I N-Terminal Propeptide and Procollagen Type I Carboxy-Terminal Propeptide)

Type I collagen accounts for 90% of the bone organic matrix and provides it with tensile strength. The type I collagen, a triple helix of two identical α1(I) chains and an α2(I) chain, is synthesized as a precursor type I procollagen that contains large additional domains at both ends of the rod-like collagen molecule (Fig. 39.9). Cleavage of the C-terminal propeptide (PICP) is necessary before collagen molecules can be assembled into collagen fibers, whereas the N-terminal propeptide (PINP) may remain longer. PINP is released before bone mineralization, but in soft tissues, type I collagen may still contain the N-terminal propeptide. The concentrations of procollagen propeptides reflect the rate of collagen synthesis.

Plasma PINP in health is present mainly as intact, trimeric form. In certain situations (ESRD with hemodialysis, bedridden geriatric patients), smaller forms of PINP are detected in plasma. PINP assays differ with respect to whether they measure the trimeric form (intact PINP) or both forms (total PINP).

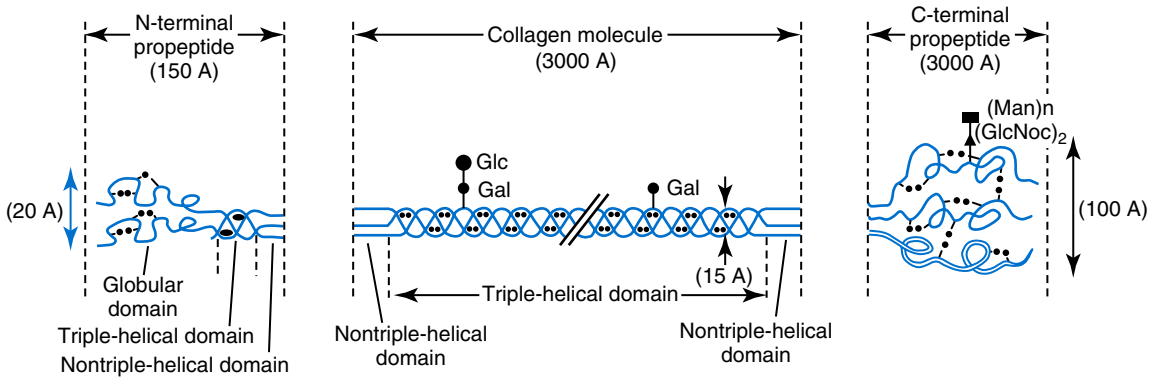


Fig. 39.9 Representation of a type I procollagen molecule. The non-triple helical domains at both ends of the collagen molecule are the telopeptides. The N-terminal propeptide consists of an N-terminal globular domain, a triple-helical domain, and a C-terminal non-triple helical domain. (From Prockop, D. J., Kivirikko, K. I., Tuderman, L., & Guzman, N. A. [1979]. The biosynthesis of collagen and its disorders [second of two parts]. *N Engl J Med*, *301*, 13–23, 77–85.)

Clearance of propeptides. Intact propeptides are degraded by the endothelial cells of the liver. Renal clearance also occurs for the smaller circulating propeptides.

Clinical significance. In principle, measurement of either propeptide should give the same information. This is the case when PICP and PINP concentrations are compared in situ. In children, circulating propeptides closely reflect the somatic growth rate, and the peak of PICP and PINP concentrations occurs earlier in girls than in boys. During puberty, the increase is larger in boys than in girls.

The measurement of response to treatment for osteoporosis therapies is a major use of bone formation markers. The concentration of PINP is increased more during active bone growth than is the concentration of PICP.

Specimen requirements. The preferred sample for measuring procollagen propeptides is serum. The antigens are very stable in serum or plasma. Storage of serum at room temperature for 1 week does not decrease concentrations.

Assays for procollagen type I carboxy-terminal propeptide and procollagen type I N-terminal propeptide. An RIA and EIA *(CICP)* are available for PICP. Assays for PINP can be divided into intact and total PINP. An RIA for intact PINP has been approved by the US Food and Drug Administration.

Reference intervals. Reference intervals for PICP are 38 to 202 μg/L in men and 50 to 170 μg/L in women. For intact PINP, the reference intervals in men are 20 to 78 μg/L, and in women 19 to 84 μg/L. The reference interval for total PINP in young healthy premenopausal women (30 to 39 years) is 16.3 to 72.2 μg/L. The reference interval for total PINP in men is 13.9 to 85.5 μg/L. Children have higher concentrations immediately after birth and later during growth spurts. In neonates (younger than 1 month of age), the reference interval of intact PINP is high at 770 to 3202 μg/L. Detailed pediatric reference intervals have been established as part of the CALIPER study (see Appendix).

Bone Alkaline Phosphatase

ALP are membrane-bound ectoenzymes that catalyze the hydrolysis of monophosphates from ester linkage under alkaline conditions (pH 8 to 10). Four different gene codes are known for the tissue: nonspecific, intestinal, placental, and germ cell (placental-like) isoenzymes. These isoforms are N-glycosylated, and only the liver ALP does not contain O-linked glycans. Bone ALP can be separated by anion-exchange HPLC or isoelectric focusing into four isoforms differing in their sialic acid content.

During mineralization of bone, the main function of bone ALP is to hydrolyze inorganic pyrophosphate (PPi) to generate phosphate (Pi). The balance between Pi and PPi is thought to be critical for mineralization because PPi inhibits the formation of hydroxyapatite.

Bone ALP is produced by osteoblasts during the matrix maturation phase (see Fig. 39.2), when the newly formed collagenous matrix is prepared for the deposition of mineral. Because ALP is firmly bound to cell membranes, release to plasma can be delayed, and increase in plasma activity takes time.

Clinical significance. Total ALP or bone ALP provides the highest clinical sensitivity and specificity in the diagnosis and monitoring of PDB. Bone ALP is more sensitive than total ALP in mild (monostotic) disease. ALP is increased in PDB and reflects increased bone formation.

Bone ALP is increased in osteoporosis, osteomalacia and rickets, hyperparathyroidism, renal osteodystrophy, thyrotoxicosis, acromegaly, and bone metastases.

ALP and bone ALP are stable in serum *in vitro* and do not require special specimen handling. Bone ALP is more useful than OC in individuals with impaired renal function because it is not cleared by GF.

Measurement of bone alkaline phosphatase. Measurement of the enzymatic activities of total and bone ALP was described in Chapter 19. Although internal standards for ALP isoenzymes have been used to increase assay precision for bone ALP activity; they tend to be technically complicated, labor intensive, imprecise, insensitive, and inaccurate.

Immunoassays are currently used for bone ALP. None of the monoclonal antibodies used is completely specific

for bone ALP. Current immunoassays exhibit 7% to 17% cross-reactivity with liver ALP.

Reference intervals. Plasma concentrations of bone ALP are influenced by age and gender. Concentrations are higher in men and increase with age in men and women.

The reference interval of bone ALP (IRMA) is 5 to 20 μg/L. With the immunoabsorption assay, it is 15.0 to 41.3 U/L for men and 11.6 to 29.6 U/L for premenopausal women. For the automated chemiluminescent immunoenzymatic method, the 95th percentile limit is 14.3 μg/L for men and 20.1 μg/L for premenopausal women. Children have much higher concentrations, especially at puberty (see Appendix).

Osteocalcin

OC is the most abundant noncollagenous protein of human bone; it is a small protein of 49 amino acids and contains three glutamyl residues at amino acid positions 17, 21, and 24, which can be converted to γ-carboxyglutamyl residues by a post-translational, vitamin K-dependent enzymatic carboxylation. These unique carboxylated amino acids bind calcium ions and are found in proteins involved in blood coagulation and in calcium transport, deposition, and homeostasis. OC binds calcium and hydroxyapatite, suggesting a physiologic role in bone mineralization.

Osteocalcin may regulate insulin secretion and sensitivity. Leptin formed in adipose tissues regulates synthesis of osteocalcin in osteoblasts. Adipose cells produce osteocalcin.

Osteocalcin is present in the plasma in carboxylated and undercarboxylated forms; it is the undercarboxylated form of osteocalcin that acts as a hormone. The undercarboxylated fraction represents 16% to 21% of total OC.

Plasma OC is cleared by the kidneys and has a half-life of 5 minutes. OC concentrations are increased in CKD. OC exhibits a diurnal variation with a nocturnal peak, decreasing by as much as 50% to a morning nadir. Circulating OC is heterogeneous; in addition to the intact molecule, several fragments also exist. Intact OC has been estimated to account for 35% of total OC, N-terminal/mid-region for 30%, and other fragments for much less in the circulation.

Clinical significance. The concentration of OC changes rapidly in situations that affect bone turnover. Plasma OC is increased in metabolic bone disease with increased bone formation-hyperparathyroidism, renal osteodystrophy, thyrotoxicosis, post fractures, acromegaly, and bone metastases. OC is decreased in hypoparathyroidism, hypothyroidism, GH deficiency, during estrogen therapy, treatment with glucocorticoids, bisphosphonates, and calcitonin.

OC concentrations may be misleading in several situations. They increase in impaired renal function without a similar increase in bone formation, because OC is cleared by the kidneys. OC may be increased during bed rest without an increase in bone formation.

About 10% to 30% of OC synthesized by the osteoblast is released into the circulation. OC fragments are released from the bone matrix during bone resorption and may contribute to circulating OC. Serum osteocalcin may be considered a marker of bone turnover rather than just bone formation.

Measurement of osteocalcin. Many immunoassays have been developed for OC since the first reported in 1980. Those assays measuring both intact OC plus the N-terminal/mid-region (1–43) fragment, are of particular interest. These assays do not measure smaller fragments of OC that are produced during bone resorption. OC concentrations are more stable when measured with assays that recognize both intact hormone (1–49) and the N-terminal/mid-region (1–43).

Specificities with respect to carboxylated and undercarboxylated forms of OC are often not known.

Circulating OC in healthy subjects varies widely between methods and laboratories.

An immunoassay for the undercarboxylated form of OC has been developed. This assay uses recombinant human undercarboxylated OC for calibration and two monoclonal antibodies, one of these specifically recognizes the 14–30 sequence, in which the three glutamic acid residues are not carboxylated.

Specimen requirements. OC is rapidly degraded in blood samples at room temperature or 4°C. Serum OC concentrations are more stable when assessed with methods measuring both intact OC and the N-terminal/mid-region fragment (1–43); concentrations are unchanged after 3 hours at room temperature and after 24 hours at 4°C.

Reference intervals. OC concentrations are influenced by age, gender, and diurnal variation. Men have higher OC than women. OC concentrations are increased during menopause. In one OC ELISA, the reference interval was 9.6 to 40.8 μg/L in men, 8.4 to 33.9 μg/L in premenopausal women, and 12.8 to 55.0 μg/L in postmenopausal women. Concentrations are higher in children—the highest of which are observed during periods of rapid growth.

Biochemical Markers of Bone Resorption

Most markers of bone resorption reflect degradation of type I collagen. These markers were initially measured in urine, but serum-based methods have been developed. An osteoclast-derived enzyme, serum TRAP5b, has been used to assess bone resorption.

Collagen Crosslinks

Extracellularly, type I collagen molecules are assembled into immature fibrils with limited tensile strength, which then are modified by formation of intramolecular and intermolecular covalent bonds or crosslinks. Intermolecular crosslinking sites are located in short non-triple-helical domains at each end of the type I collagen molecule, called *telopeptides*. The activities of two enzymes—intracellular lysyl hydroxylase 2b and extracellular lysyl oxidase—regulate which variant structures of crosslinks are formed. PYD and DPD are stable even to acid hydrolysis, whereas the pyrrole variants are quite unstable. Trivalent crosslinks are found in soft tissues and in collagen types other than type I (type II, III).

PYDs were the first mature crosslinks to be identified, but their quantity in bone is relatively low, one per three to five collagen molecules, compared with the quantity of their divalent precursors, 1 per collagen molecule. The pyrrole

crosslinks are concentrated at the N-terminal end of human bone collagen. Both telopeptides have a sequence including an aspartic acid residue, which can undergo racemization and α-isomerization. Immunoassays detect only one form of these structures.

Degradation of bone collagen. Cathepsin K is the main enzyme that degrades bone collagen. Its activity alone is sufficient to completely dissolve insoluble collagen of adult cortical bone. Osteoclasts produce other proteases, notably several matrix metalloproteinase (MMP) enzymes (MMP-9, MMP-10, MMP-12, MMP-14). Crosslinked telopeptides are released from type I collagen during bone resorption because crosslinks resist proteolysis of adjacent polypeptide chains. This may lead to the generation of a neoepitope, such as NTX, that can be detected by immunoassays.

For C-terminal telopeptides, two immunoassays, ICTP and CTX, reflect different enzymatic pathways of bone breakdown. Cathepsin K releases CTX antigen but destroys reactivity in the ICTP assay. ICTP antigen is stable after MMP-1, MMP-9, and MMP-13 digestion. The ICTP assay is also called the *CTX-MMP assay.*

NTX and CTX reflect normal osteoclast function. ICTP detects pathologic degradation of bone and soft connective tissue as seen in multiple myeloma, bone metastasis, and rheumatoid arthritis.

Comparison of urine and serum analyses. Urine was the sample of choice for measuring bone collagen degradation. Urinary hydroxyproline analysis was replaced by more sensitive and specific crosslinks (DPD, PYD) and telopeptide (CTX, NTX) assays. Urinary measurements need to be corrected for creatinine excretion to adjust for effects of urinary concentration or dilution. Including a second analyte, with its analytical variation, adds to the uncertainty of the result.

Serum and urine telopeptide assays have been compared in several studies; correlation coefficients varied from 0.6 to 0.9. Kidneys are involved in degradation of crosslinked peptides; as a result, 40% of total crosslinks in the urine are free. The mean fractional excretion of NTX (0.22) is lower than that of CTX (0.44), indicating renal degradation is greater for the former.

During antiresorptive treatment, urinary telopeptides may overestimate the decrease in bone resorption because the renal metabolism of these peptides can change. The response to antiresorptive therapy is generally greatest when measured with telopeptide assays, intermediate with total DPD, and lowest with free DPD. Serum or plasma assays have replaced urine assays in many countries.

Clinical significance of telopeptides and crosslinks. Increased urinary CTX, NTX, and DPD have been reported in osteoporosis, PDB, PHPT, and 2°HPT, hyperthyroidism, carcinoma metastases, and multiple myeloma.

Inhibition of bone resorption with pharmacologic agents (bisphosphonates, denosumab) leads to a decrease in serum/plasma or urinary CTX, NTX, and DPD; this is the most common use of these markers in clinical practice.

Measurement of telopeptides and pyridinium crosslinks. Several methods are available; the ones most often used are serum and urine assays for NTX and CTX, serum assay for 1CTP, and urine assays for DPD.

N-telopeptide. NTX in urine is measured with an ELISA. The NTX method has been adapted for the measurement of N-telopeptides in serum. The use of serum can significantly reduce within-subject day-to-day variation.

C-telopeptide CTX assay. EDTA plasma is the preferred sample. Serum can be used if rapidly separated prior to storing frozen at a minimum of −20°C. The most widely used methods recognize the α-isomer. In healthy human bone, 70% to 80% of C-telopeptides have α-isomerization, but in PDB, it accounts for only 50 to 60. α-Isomerization and racemization may vary between individuals. All CTX methods require the C-terminal amino acid to be a free arginine. Serum/plasma CTX is measured using two monoclonal antibodies.

C-telopeptide ICTP assay. An RIA for ICTP detects collagen degradation fragments in serum. The epitope measured is vulnerable to digestion by cathepsin K; this method is relatively insensitive to changes in bone resorption caused by normal bone turnover. The method is suitable for detecting the osteolytic processes taking place in multiple myeloma and metastatic bone disease, as well as the erosion process in rheumatoid arthritis.

Deoxypyridinoline and pyridinoline. DPD and PYD were originally measured by HPLC. Today, DPD is measured primarily by immunoassay using automated analyzers or manual ELISA. Unlike most HPLC methods, which can measure total DPD and PYD, the immunoassays for DPD or DPD/PYD measure primarily free, but not peptide-bound, forms.

Specimen requirements. Serum or plasma is best collected after overnight fasting. A second morning void urine, collected by 10:00 hours, is recommended. For treatment monitoring, specimens should be collected at the same time as the baseline specimen. Peak urinary excretion of PYDs occurs at 05:00 to 08:00 hours, reflecting the nocturnal peak in bone turnover. Urinary PYDs reach a nadir between 14:00 and 23:00 hours.

PYDs and telopeptides are relatively stable in urine. Exposure to UV light decreases DPD and PYD.

Reference intervals. Concentrations of collagen crosslinks are influenced by age and gender. Concentrations are markedly higher in children, with highest concentrations observed during early infancy and adolescence, which are periods of rapid bone growth. In adults, concentrations of collagen crosslinks are relatively constant between 30 and 45 years, increasing significantly after menopause. Age-related increases have been reported in men.

Tartrate-Resistant Acid Phosphatase 5b (TRACP5b)

Osteoclasts produce and secrete **tartrate-resistant acid phosphatase 5b (TRACP5b)** (see Fig. 39.2) during bone resorption. Osteoclast number can be increased in bone disorders and is decreased by antiresorptive treatment. Early methods failed to distinguish between two isoforms: TRACP5b produced by the osteoclasts, and TRACP5a produced by inflammatory macrophages and dendritic cells. Isoform 5b is a

dimer, 5a is a monomer. 10% of TRACP circulates in an enzymatically active form, the remaining 90% present as inactive fragments.

In a kinetic method, TRACP5b activity is estimated by subtracting tartrate-resistant fluoride-resistant acid phosphatase activity from total activity. Immunoassays are available for TRACP5b.

Specimen requirements. The preferred sample is serum. Diurnal variation is minor, indicating that the half-life of TRACP5b is longer than those of the other markers of bone resorption. TRACP5b is unstable on storage; it can be stored for up to 8 hours at room temperature and up to 3 days at 4°C. Long-term samples must be stored at −80°C. Six months' storage at −20°C reduces activity by 40%.

Clinical significance. In response to alendronate, TRACP5b decreased by a mean (SE) of 39% (4%), compared with 49% (4%) to 69% (5%) for urinary telopeptides (CTX and NTX), and 75% (8%) for serum/plasma CTX. Boys aged 13 to 17 years have higher concentrations than girls in the same age group.

Reference interval. An immunoassay for TRACP5b has upper limits of reference intervals of 4.2 U/L for women and 4.8 U/L for men.

Preanalytical and Analytical Variables of Bone Turnover Marker

Controllable sources of preanalytical variability include sampling time, sample preservation, and food intake. Most bone markers have a diurnal rhythm with a peak in the early morning hours (04:00 to 08:00 hours) and nadir in the afternoon/evening (13:00 to 23:00 hours). The amplitude of this variation is greatest for resorption markers, with nadir values averaging 70% of peak values. Specimens should be collected at a standardized time of day to minimize the impact of diurnal variability. For urinary markers, collection of the second morning void is recommended.

Urinary resorption markers are usually normalized by dividing by the urinary creatinine concentration. Variability (within- and between-method) in creatinine measurements, within-subject variability in urinary creatinine, and its dependence on muscle mass contribute to the variability of urinary markers.

Within-individual variability of urine markers is higher (15% to 60%) than that of plasma markers (5% to 10%). Bone ALP and TRACP5b do not demonstrate much diurnal variation because of their long half-lives.

Food intake affects CTX concentration. The diurnal variation of plasma CTX is ±40% around the 24 hours mean; fasting reduces circadian variation. Exercise can acutely increase the concentration of CTX. With other bone markers, the clinical impact of feeding versus fasting is small. The collection of samples in the fasting and resting state is standard practice.

METABOLIC BONE DISEASES

Metabolic bone disease results from a partial uncoupling or imbalance between bone resorption and formation. Decreased bone mass, or **osteopenia**, is more common than increases in bone mass. The most prevalent metabolic bone diseases are osteoporosis, osteomalacia/rickets, and **renal osteodystrophy**. Osteoporosis is characterized by loss of bone mass, microarchitectural deterioration of bone tissue and increased risk of fracture. Rickets and osteomalacia are characterized by the defective mineralization of the bone matrix. Renal osteodystrophy is a complex condition that develops in response to abnormalities of the endocrine and excretory functions of the kidneys.

Osteoporosis

Osteoporosis results in over 1.5 million fractures each year. It is associated with increased risk for vertebral, hip, and distal forearm fractures. At age 50, women have a lifetime fracture risk (at any of these sites) of 40%. Men have a lifetime fracture risk one-third that of women. Trabecular bone turns over at five to seven times the rate of cortical bone and so fractures of bones that are predominantly trabecular (vertebrae and distal forearm) occur earlier in life. One third of women older than 65 years suffer vertebral crush fracture, which can occur acutely and result in disabling pain and discomfort. Long-term complications include immobility and loss of height, protuberant abdomen, hiatus hernia, chronic constipation, urinary incontinence, and loss of self-esteem. Up to 50% of vertebral fractures may be clinically silent when they occur, only to manifest later as an incidental finding on x-ray.

Fractures of cortical bone (proximal femur or hip) occur later in life. For women, the lifetime risk of hip fracture is 15% and for men 3%. The mortality rate accompanying hip fracture may be as high as 20% in the 6 months following fracture. Mortality is higher in men than women; it increases with age, and is higher for those with coexisting illness and poor pre-fracture functional status. Twenty-five percent of survivors are confined to long-term care in nursing homes.

Peak bone mass is attained by 30 years and decreases slowly after 40 years in both men and women. Bone mass attained during growth is an important determinant of osteoporosis. Exercise and adequate nutrition play important roles in attaining and maintaining skeletal mass. During early adult life, bone formation is coupled with bone resorption so that bone mass remains stable. Aging is a major risk factor for bone loss. After 40 years, bone resorption exceeds bone formation, so that 0.5% of the skeletal mass is lost per year. In women, the decrease in sex steroids at menopause accelerates bone loss to 2% to 3% per year. Osteoporosis is commonly encountered in postmenopausal women. Advanced age, female gender, sex steroid deficiency, a family history of osteoporosis, alcohol abuse, smoking, and chronic disease are all risk factors.

After decreased bone mass is documented by bone mass measurements, the diagnostic work-up is directed at determining the cause, which could be due to endocrine or inflammatory conditions, malabsorption, and drugs, among others.

Bone markers can assess bone turnover in patients with osteoporosis, the rate of bone turnover is considered an important determinant of bone fragility in postmenopausal women. Markers for both bone resorption and bone formation

can be useful in monitoring effects of antiresorptive therapy. The International Osteoporosis Foundation and International Federation of Clinical Chemistry and Laboratory Medicine recommend one formation marker (serum PINP) and one resorption marker (plasma CTX) be used as reference markers.

High calcium intake (1000 to 1500 mg/day), adequate vitamin D (400 to 800 IU/day), sufficient protein intake, and a regular exercise program are helpful in maintaining bone mass and preventing osteoporosis.

In secondary osteoporosis, treatment is directed at the underlying condition. Most therapies for postmenopausal osteoporosis are directed at decreasing osteoclast bone resorption.

Osteomalacia and Rickets

Osteomalacia in adults and **rickets** in children are caused by a mineralization defect that occurs during bone formation, resulting in an increase in osteoid. Osteomalacia or rickets is due to either vitamin D deficiency or phosphate depletion.

Despite supplementation in some countries, of milk, bread, and some cereals with vitamin D, vitamin D insufficiency is common in North America and Europe. Breastfed infants, the elderly, strict vegetarians, and individuals with darker skin pigmentation are at increased risk. Clinical osteomalacia is uncommon. Mild to moderate vitamin D deficiency may be associated with reduced muscle strength, impaired physical performance, and falls—all of which are factors that contribute to osteoporotic fracture. Vitamin D deficiency may develop in cases of malabsorption.

Vitamin D-dependent rickets type I is an inherited defect in 25(OH)D-1α-hydroxylase that causes impaired formation of $1,25(OH)_2D$. The disease is manifested in infancy and can be treated with $1,25(OH)_2D$. Vitamin D-dependent rickets type II is an inherited disorder that is characterized by high plasma concentrations of $1,25(OH)_2D$ caused by resistance to $1,25(OH)_2D$, secondary to defects in the $1,25(OH)_2D$ receptor.

Osteomalacia and rickets may occur as the result of phosphate depletion. In developing countries, dietary calcium deprivation may lead to rickets, without vitamin D or phosphate deficiency.

Clinical manifestations of rickets and osteomalacia are a consequence of the defect in mineralization. In adults, bone pain is the most common symptom, stress fractures and frank skeletal fractures occur.

Vitamin D deficiency is diagnosed by measuring serum 25(OH)D (Table 39.3). Other findings in rickets and osteomalacia include increased ALP due to increased osteoblast activity associated with producing unmineralized osteoid.

Treatment of rickets and osteomalacia is dictated by the cause. Nutritional rickets and osteomalacia are healed by treatment with physiologic doses of vitamin D, whereas higher doses may be required in malabsorption.

Disorders of Bone and Mineral in Chronic Kidney Disease (Renal Osteodystrophy)

CKD is associated with a multitude of disorders of bone and mineral metabolism. Renal bone diseases include both high-turnover bone disease (**osteitis fibrosa** or 2°HPT) and low-turnover bone disease (osteomalacia and **adynamic bone disease [ABD]**). Quantitative histomorphometric analysis of bone biopsies, measurement of bone formation by double tetracycline labeling, and special stains are often necessary for correct diagnosis of patients with osteitis fibrosa, osteomalacia, ABD, and mixed bone disease of renal osteodystrophy.

Osteitis fibrosa (hyperparathyroid bone disease) is the most common high-turnover disease. This disorder is caused by high concentrations of serum PTH, which is a consequence of the hypocalcemia associated with hyperphosphatemia and $1,25(OH)_2D$ deficiency.

Low-turnover bone diseases include osteomalacia and adynamic *(aplastic)* bone disease. Osteomalacia and ABD are distinguished by the extent of unmineralized bone matrix. Osteomalacia in chronic renal failure may reflect vitamin D deficiency caused by decreased renal synthesis of $1,25(OH)_2D$.

Bone pain is the most common complaint of patients with renal osteodystrophy. The weight-bearing bones are the sites of greatest discomfort, with leg, hip, and back pain being common.

The central role of PTH in guiding therapy requires that the PTH assay used can be relied on to measure only the active hormone. This is not true for any of the assays available, and kidney disease leads to accumulation of inactive PTH fragments above concentrations seen in healthy individuals. Treatment guidelines take this into account, but inconsistencies among assays can lead to situations in which opposite therapeutic decisions could be made for a single patient, depending on the assay used.

Biochemical findings in CKD include hyperphosphatemia and hypocalcemia (see Chapter 35). The measured concentration of immunoreactive PTH is increased, and $1,25(OH)_2D$ is decreased. Plasma ALP is increased in patients with hyperparathyroidism or osteomalacia. Because magnesium is cleared by the kidney, modest increases in plasma concentration (2 to 4 mg/dL [0.08 to 0.16 mmol/L]) are common.

Early management of CKD calls for dietary restriction of phosphate and administration of phosphate-binding agents. Calcium supplements are also used. Administration of $1,25(OH)_2D$ or other active forms of vitamin D enhances intestinal calcium absorption and may act directly on the parathyroid glands to reduce PTH secretion. Ultimately dialysis or renal transplantation may be necessary.

Paget Disease of Bone

PDB is a localized disease of bone characterized by osteoclastic bone resorption, followed by replacement of bone in a chaotic fashion. It is most common in Northern Europe, North America, Australia, and New Zealand in persons of Anglo-Saxon descent; prevalence may be up to 5% among people older than 40 years. It is uncommon in Asia, Africa, and Scandinavia. A family history of PDB is reported by 20% to 50% of patients. Environmental factors, especially a possible paramyxovirus infection (measles, canine distemper, respiratory syncytial), may play a role in causation.

PDB may affect one bone (monostotic) or several bones (polyostotic). Signs and symptoms depend on which skeletal site is affected; the skull, femur, pelvis, and vertebrae are common sites. The disease is often diagnosed from radiographs or laboratory tests (total ALP) performed for another reason. Advanced disease can produce deformities such as skull enlargement and anterior bowing of the weight-bearing bones (femur and tibia).

Increases in biochemical markers of bone resorption reflect the osteoclastic nature of the disease.

Therapy is directed at decreasing osteoclastic bone resorption. Patients may require surgery for skeletal deformity that limits mobility, or for arthritic changes, fractures, or nerve compression.

Osteogenesis Imperfecta

Osteogenesis imperfecta (OI), often termed "brittle bone disease," is a heterogeneous group of inherited connective tissue disorders that share similar skeletal abnormalities resulting in low bone mass, minimal trauma fractures, and subsequent bone deformity. In its most severe form it can be lethal in fetal life and in neonates.

Biochemical and molecular studies have identified that milder forms of OI are caused by quantitative defects in type 1 collagen and that more severe types are caused by structural defects in either of the two chains that form the heterotrimer. There is a spectrum of signs and symptoms including frontal bossing of the skull, bluish sclera, yellowish teeth, barrel chest/pectus excavatum, joint laxity, vertebral compression, and growth retardation. Recurrent fractures with minimal trauma may lead to suspicion of child abuse. When a diagnosis is in doubt, biochemical studies of collagen expressed by cultured cells, DNA sequencing of type 1 collagen, and genetic expression plus gene sequencing studies may be of value in confirming the diagnosis.

Involvement of Bone in Malignancies

Bone metastases are the most common skeletal complication of malignancy, occurring in up to 70% of patients with advanced breast or prostate cancer, and in 15% to 30% of patients with carcinoma of lung, colon, stomach, bladder, uterus, rectum, thyroid, or kidney. Metastases can have a markedly osteolytic or osteoblastic character, but often they are mixed. Patients with breast cancer have predominantly osteolytic lesions, but 5% to 20% of metastases have an osteoblastic nature. Metastases of prostate cancer are mainly osteoblastic. In most carcinoma metastases, both bone degradation and formation take place and both biochemical markers of bone formation and resorption are valuable in assessing the presence of bone metastasis.

REVIEW QUESTIONS

1. Which of the following cells is the main producer of RANKL in bone?
   - **a.** Osteoclast
   - **b.** Osteoblast
   - **c.** Osteocyte
   - **d.** Bone lining cell
   - **e.** Chondrocyte
2. Where is the major site of regulation of magnesium excretion in the kidney?
   - **a.** Glomerulus
   - **b.** Proximal convoluted tubule (PCT)
   - **c.** Cortical collecting duct (CCD)
   - **d.** Distal convoluted tubule (DCT)
   - **e.** Thick ascending limb of the Loop of Henle
3. When is measurement of parathyroid hormone-related peptide of greatest clinical value?
   - **a.** In the differential diagnosis of hypocalcemia
   - **b.** In pregnancy
   - **c.** In patients with hypercalcemia thought to be caused by a tumor but with no obvious clear diagnosis of malignancy
   - **d.** In chronic kidney disease
   - **e.** In patients suspected to have primary hyperparathyroidism
4. Biochemical markers of bone metabolism are of clinical value in—
   - **a.** Indicating a response to treatment of metabolic bone disease
   - **b.** Diagnosing osteoporosis
   - **c.** Diagnosing osteomalacia
   - **d.** Indicating a requirement for calcium supplementation
   - **e.** The guidelines for recommending parathyroidectomy
5. Which of the following biochemical tests are of diagnostic value in Paget disease of bone?
   - **a.** Parathyroid hormone
   - **b.** Fibroblast growth factor 23
   - **c.** Alkaline phosphatase
   - **d.** 25 hydroxyvitamin D
   - **e.** Acid phosphatase
6. What can cause a falsely high 25 hydroxyvitamin D measurement in many immunoassays?
   - **a.** Hypercalcemia
   - **b.** Li heparin contamination
   - **c.** Ingestion of magnesium sulfate
   - **d.** High concentration of the 3-epimer of 25OHD in the sample
   - **e.** High concentration of 24,25-dihydroxyvitamin D in the sample

SUGGESTED READINGS

Bilezikian, J. P., Brandi, M. L., Eastell, R., et al. (2014). Guidelines for the management of asymptomatic primary hyperparathyroidism: summary statement from the Fourth International Workshop. *The Journal of Clinical Endocrinology and Metabolism*, *99*, 3561–3569.

Compston, J., Cooper, A., Cooper, C., & National Osteoporosis Guideline Group (NOGG)., et al. (2017). UK clinical guideline for the prevention and treatment of osteoporosis. *Archives of Osteoporosis*, *12*, 43.

Compton, J. T., & Lee, F. Y. (2014). A review of osteocyte function and the emerging importance of sclerostin. *Journal of Bone and Joint Surgery. American Volume*, *96*, 1659–1668.

Couchman, L., Taylor, D. R., Krastins, B., Lopez, M. F., & Moniz, C. F. (2014). LC-MS candidate reference methods for the harmonisation of parathyroid hormone (PTH) measurement: a review of recent developments and future considerations. *Clinical Chemistry and Laboratory Medicine*, *52*, 1251–1263.

Cundy, T. (2017). Paget's disease of bone. *Metabolism* (17). https://doi.org/10.1016/j.metabol.2017.06.010. pii: S0026–0495 30182-8 [Epub ahead of print].

Eastell, R., O'Neill, T. W., Hofbauer, L. C., et al. (2016). Postmenopausal osteoporosis. *Nature Reviews Disease Primers*, *2*, 16069. https://doi.org/10.1038/nrdp.2016.69.

Fink, H. A., Litwack-Harrison, S., Taylor, B. C., & for the Osteoporotic Fractures in Men (MrOS) Study Group., et al. (2016). Clinical utility of routine laboratory testing to identify possible secondary causes in older men with osteoporosis: the Osteoporotic Fractures in Men (MrOS) Study. *Osteoporosis International*, *27*, 331–338.

Florez, H., Peris, P., & Guañabens, N. (2016). Fibrous dysplasia. Clinical review and therapeutic management. *Medicina Clínica*, *147*(12), 547–553.

Fraser, W. D. (2009). Hyperparathyroidism. *Lancet*, *374*(9684), 145–158.

Gonciulea, A. R., & Jan De Beur, S. M. (2017). Fibroblast growth factor 23-mediated bone disease. *Endocrinology and Metabolism Clinics of North America*, *46*(1), 19–39.

KDIGO clinical practice guideline for the diagnosis, evaluation, prevention, and treatment of Chronic Kidney Disease-Mineral and Bone Disorder (CKD-MBD). (2009). Kidney disease: improving global outcomes (KDIGO) CKD-MBD work group. Kidney International. *Supplement*, *113*, S1–S130.

Ketteler, M., Block, G. A., Evenepoel, P., et al. (2017). Executive summary of the 2017 KDIGO Chronic Kidney Disease-Mineral and Bone Disorder (CKD-MBD) Guideline Update: what's changed and why it matters. *Kidney International*, *92*, 26–36.

Marini, J. C., Forlino, A., Bächinger, H. P., et al. (2017). Osteogenesis imperfecta. *Nature Review Disease Primers*, *18*(3), 17052.

Pludowski, P., Holick, M. F., Grant, W. B., et al. (2018). Vitamin D supplementation guidelines. *Journal of Steroid Biochemistry and Molecular Biology*, *175*, 125–135.

Ralston, S. H., & Albagha, O. M. (2014). Genetics of Paget's disease of bone. *Current Osteoporosis Reports*, *12*, 263–271. https://doi.org/10.1007/s11914-014-0219-y.

Reid, I. R. (2017). Vitamin D effect on bone mineral density and fractures. *Endocrinology and Metabolism Clinics of North America*, *46*, 935–945.

Thode, J., Juul-Jorgensen, B., Bhatia, H. M., et al. (1989). Comparison of serum total calcium, albumin-corrected total calcium, and ionized calcium in 1213 patients with suspected calcium disorders. *Scandinavian Journal of Clinical and Laboratory Investigation*, *49*, 217–223.

Whyte, M. P. (2017). Hypophosphatasia: an overview for 2017. *Bone*, *102*, 15–25.

Disorders of the Pituitary Gland

Neil S. Harris, William E. Winter, Ann McCormack, Roger L. Bertholf

OBJECTIVES

1. Describe the anterior pituitary hormones and their release/inhibiting factors.
2. Discuss the levels of control of the anterior pituitary hormones including the long and short feedback loops.
3. Be familiar with the critical role of pulsatility in the release of hypothalamic-releasing hormones.
4. Discuss the factors that contribute to variations in reference intervals for anterior pituitary hormones.
5. Describe the physiologic effects of hormones synthesized by the anterior pituitary gland.
6. Understand the complex interplay between growth hormone and insulin-like growth factor 1 (IGF-1).
7. List the main causes of hyperprolactinemia.
8. Be aware of macroprolactinemia as a common cause of an elevated plasma prolactin.
9. Describe the role of luteinizing hormone (LH) and follicle-stimulating hormone (FSH) in gonadal steroid synthesis and ovulation.
10. State the effects of increased and decreased anterior pituitary hormone secretion and list the laboratory tests used to assess hypersecretion and hypopituitarism.
11. With respect to the posterior pituitary, understand the role of antidiuretic hormone (ADH) and its physiologic control.
12. Discuss the pathophysiology of diabetes insipidus and the syndrome of inappropriate antidiuretic hormone (SIADH).

KEY WORDS AND DEFINITIONS

Acid-labile subunit (ALS) IGF-1 and IGF-2 circulate together with insulin-like growth factor binding protein-3 (IGFBP-3) and the acid-labile subunit (ALS) to form a 150-kDa trimeric protein complex.

Acromegaly A chronic disease of adults caused by hypersecretion of pituitary growth hormone, characterized by enlargement of many parts of the skeleton and soft tissues as well as many organs.

Adrenocorticotropic hormone (ACTH) A 39–amino acid peptide hormone secreted by the anterior pituitary that stimulates the adrenal cortex to secrete cortisol.

Antidiuretic hormone (ADH) A 9-amino acid peptide hormone synthesized in the hypothalamus but stored and released from the posterior pituitary lobe; also known as vasopressin.

Corticotropin-releasing hormone (CRH) A 41–amino acid neuropeptide released by the hypothalamus that stimulates the release of ACTH by the anterior pituitary.

Diabetes insipidus Failure of the kidney tubules to reabsorb sufficient water; caused by either a deficiency of ADH or defective receptor action.

Follicle-stimulating hormone (FSH) A glycoprotein hormone secreted by the anterior pituitary. In women FSH stimulates the growth and maturation of ovarian follicles (eggs), stimulates estrogen secretion, and promotes endometrial changes. In males, FSH stimulates the Sertoli cells; in this way spermatogenesis is supported.

Gigantism Excessive growth caused by increased synthesis of GH before the epiphyses have fused.

Growth hormone (GH) A polypeptide of 191 amino acids produced by the anterior pituitary that affects carbohydrate, lipid, and protein metabolism. GH regulates the secretion of insulin-like growth factor I (IGF-1).

Insulin-like growth factors (IGFs) These are members of the insulin-related peptide family; they include insulin and relaxin. Proinsulin, IGF-1, IGF-2, and relaxin are all composed of two domains joined by a connecting domain. In children, IGF-1 secretion is a major determinant of normal linear growth.

Insulin-like growth factor binding proteins (IGFBPs) IGFBPs have a very high affinity for IGFs and function as inhibitors of IGF action. IGFBPs have higher affinity for IGFs than do the IGF receptors.

Luteinizing hormone (LH) A glycoprotein gonadotropic hormone secreted by the anterior pituitary that acts with FSH to promote ovulation and progesterone production. In males, LH stimulates testosterone secretion.

Oxytocin A nonapeptide hormone synthesized in the hypothalamus and stored in the posterior lobe of the

pituitary. It induces smooth muscle contraction in the uterus and mammary glands.

Polydipsia Chronic excessive intake of water, as in diabetes mellitus or diabetes insipidus.

Polyuria The passage of a large volume of urine in a given period of time, a characteristic of diabetes mellitus and diabetes insipidus.

Pro-opiomelanocortin (POMC) This is the precursor 266–amino acid protein of ACTH as well as the melanocyte-stimulating hormones and beta-endorphin.

Prolactin (PRL) A lactogenic hormone synthesized by the anterior pituitary.

Syndrome of inappropriate antidiuretic hormone (SIADH) A condition where inappropriately excessive ADH secretion produces dilutional hyponatremia with an elevated urine osmolality.

Thyroid-stimulating hormone (TSH) A glycoprotein hormone synthesized by the anterior pituitary that sustains, stimulates, and promotes the growth of the hormonal secretion of the thyroid gland; also called thyrotropin.

Thyrotropin-releasing hormone (TRH) A tripeptide produced in the hypothalamus that stimulates the release of TSH from the anterior pituitary.

Vasopressin See ADH or antidiuretic hormone.

BACKGROUND

The anterior and posterior lobes of the pituitary gland control processes vital for survival of the individual and the species.

CONTENT

This chapter focuses on disorders of the anterior and posterior pituitary that produce deficient or excess hormone activity. The hypophysis is composed of the adenohypophysis (the anterior lobe of the pituitary; ≈75% of the mass of the pituitary) and the neurohypophysis (the posterior lobe of the pituitary, ≈25% of the mass of the pituitary) (Fig. 40.1).

The biology of the adenohypophysis is distinctly different from that of the neurohypophysis; the adenohypophysis is controlled by the hypothalamus via releasing or inhibiting hormones, whereas the cell bodies of the neurohypophysis are anatomically located in hypothalamic nuclei, with **oxytocin** or **antidiuretic hormone (ADH)** reaching the neurohypophysis through nerve axons.

The roles of the various hormones secreted by the pituitary are exceedingly diverse and include regulation of (1) the body's response to stress (**adrenocorticotropic hormone [ACTH]** and **growth hormone [GH]**), (2) the metabolic rate (**thyroid-stimulating hormone [TSH]**), (3) growth (TSH and GH), (4) reproduction (**luteinizing hormone [LH]** and **follicle-stimulating hormone [FSH]**), (5) nourishment for the newborn and infant (prolactin), (6) parturition and milk letdown during breast feeding (oxytocin), and (7) fluid balance and blood pressure regulation in states of stress (ADH and cortisol).

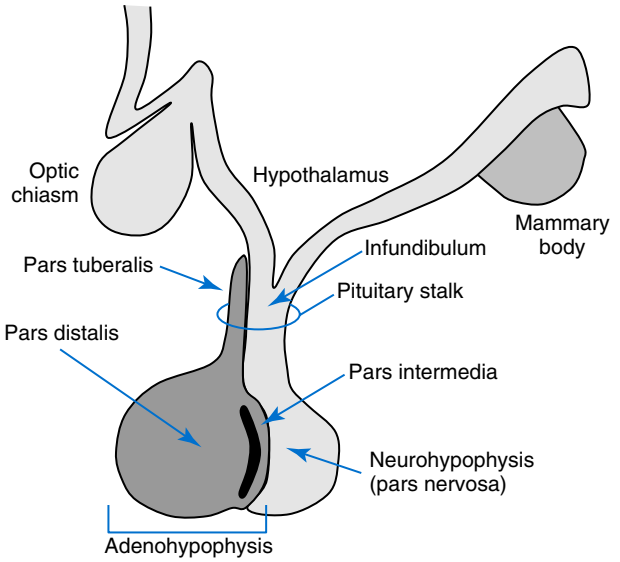


Fig. 40.1 The hypophysis (pituitary gland) is composed of the adenohypophysis (the anterior lobe of the pituitary) and the neurohypophysis (the posterior lobe of the pituitary; pars nervosa). The adenohypophysis has three parts: the pars distalis, where most of the hormone-producing cells are located; the pars tuberalis, which is part of the pituitary stalk; and the pars intermedia.

ANATOMY

The pituitary is located at the base of the brain and is protected by the bony sella turcica. Direct delivery of hypothalamic regulatory hormones to the adenohypophysis occurs through the hypothalamic-pituitary portal system. Anterior and superior to the pituitary is the optic chiasm. These relations are clinically important because pituitary neoplasms can invade or compress these structures.

Regulation of Function of the Adenohypophysis

The synthesis and release of the following anterior pituitary hormones are stimulated by hypothalamic-releasing hormones: ACTH, TSH, GH, LH, and FSH. **Prolactin** is the sole anterior pituitary hormone whose release is predominantly regulated through suppression. Corticotrophs secrete ACTH, thyrotrophs secrete TSH, somatotrophs secrete GH, gonadotrophs secrete both LH and FSH, and lactotrophs secrete prolactin. Except for LH and FSH, each hormone is normally produced by a unique cell type. The molecular composition of the anterior pituitary hormones is summarized in Table 40.1.

Multiple levels of control of the hypothalamic-pituitary-end organ-hormone axis are known (Fig. 40.2). Except for prolactin and LH at the midpoint of the menstrual cycle, negative feedback controls secretion of the adenohypophyseal hormones. The long feedback loop involves suppression

TABLE 40.1 Hypothalamic-Releasing or Hypothalamic-Inhibiting Hormones, Their Target Cells, and the Hormone That Is Regulated

Hypothalamic Hormone/ Abbreviation	Amino Acids	Anterior Pituitary Target Cell	Hormone Regulated	Amino Acids	MW (kDa)
Corticotropin-releasing hormone	41	Corticotroph	ACTH	39	4.5
Thyrotropin-releasing hormone	3	Thyrotroph	TSH[a]	α: 92 β: 118	28
Growth hormone–releasing hormone	44	Somatotroph	GH[c]	191	22
Somatotropin release–inhibiting hormone[b] (SRIH)	14	Somatotroph	GH[c]	191	22
Gonadotropin-releasing hormone	10	Gonadotroph	LH[a]	α: 92 β: 121	32
			FSH[a]	α: 92 β: 111	30
Prolactin release–inhibiting hormone	1	Lactotroph	Prolactin	199	22

[a]All α-glycoprotein chains are identical, including the α-chain of human chorionic gonadotropin.
[b]Also known as somatostatin.
[c]Note: The 20 kDa form of GH is not shown in the table but is secreted together with the 22 kDa form of GH.
ACTH, Adrenocorticotropic hormone; *FSH*, follicle-stimulating hormone; *GH*, growth hormone; *LH*, luteinizing hormone; *MW*, molecular weight; *TSH*, thyroid-stimulating hormone.

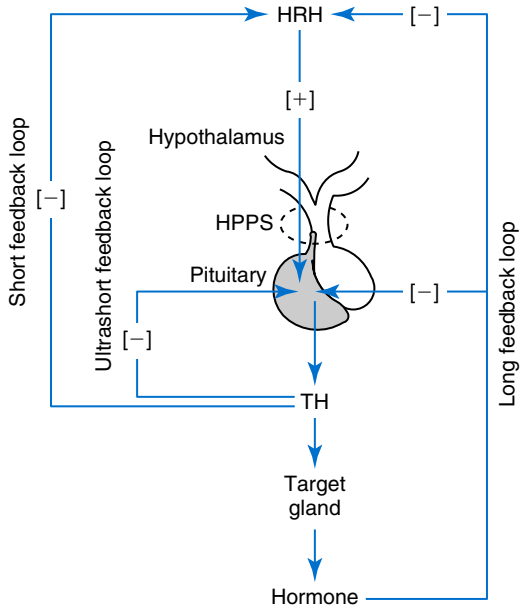


Fig. 40.2 Many anterior pituitary trophic hormones (e.g., adrenocorticotropic hormone, thyroid-stimulating hormone, growth hormone, luteinizing hormone, follicle-stimulating hormone) are regulated by hypothalamic releasing hormones *(HRHs)*. Releasing hormones secreted by the hypothalamus reach the pituitary via the hypothalamic-pituitary portal system *(HPPS)*. Long feedback loops involve negative feedback of the target cell hormone at the pituitary gland and hypothalamus. The short feedback loop involves the anterior pituitary trophic hormone feeding back at the hypothalamus, whereas the ultrashort feedback loop involves the anterior pituitary hormone feeding back at the anterior pituitary. *[+]*, Stimulation; *[−]*, suppression; *TH*, trophic hormone.

of the hypothalamic-releasing hormones and the anterior pituitary trophic hormones by the hormonal product of the target tissue. The major site of negative feedback for cortisol (regulated by ACTH), insulin-like growth factor-I (IGF-1; regulated by GH), and sex steroids and inhibins (regulated by LH and FSH) is the hypothalamus. In contrast, for thyroid hormone (regulated by TSH), the major site of negative feedback is the anterior pituitary. Retrograde flow from the pituitary to the hypothalamus via the portal system permits the existence of short negative feedback loops in which pituitary hormones suppress the secretion of hypothalamic-releasing hormones. Ultrashort feedback loops also exist in which pituitary hormones inhibit their own secretion.

AT A GLANCE

1. The pituitary is a master gland for many vital endocrine functions regulating growth (growth hormone [GH]), thyroid function (thyrotropin), reproduction (luteinizing hormone [LH], follicle-stimulating hormone [FSH], and prolactin), stress responses (ACTH), parturition (oxytocin), and water balance (**antidiuretic hormone [ADH]**).
2. The hypothalamus regulates pituitary hormone release or is the site of synthesis of the posterior pituitary hormones.

POINTS TO REMEMBER

1. Most anterior pituitary hormones are predominantly regulated by stimulatory hypothalamic hormones. The exception is prolactin, which is regulated by suppression.
2. The pulsatility of release of hypothalamic-releasing hormones can regulate the release of anterior pituitary hormones such as luteinizing hormone (LH) and follicle-stimulating hormone (FSH).
3. Reference intervals may depend on the time of the day at which a measurement is obtained, the sex of the subject, his or her age, and, for gonadotropin and sex steroid concentrations in women of reproductive age, the woman's menstrual cycle.

TABLE 40.2 Hypothalamic-Pituitary-End Organ Physiology

Hypothalamic Hormone	Anterior Pituitary Hormone	Target Organ/Tissue	Target Hormone
CRH	ACTH	Adrenal cortex: zona fasciculata and zona reticularis	Cortisol
TRH	TSH	Thyroid follicular cell	Thyroxine (T_4) and 3,5,3′-triiodothyronine (T_3)
GHRH and SRIH	GH	Liver and many tissues of the body	IGF-1, IGFBP-3, and ALS
GnRH	LH, FSH	Gonad	Sex steroids and inhibins
PRIH	Prolactin	Breast	Not applicable

ACTH, Adrenocorticotropic hormone; *ALS*, acid-labile subunit; *CRH*, corticotropin-releasing hormone; *FSH*, follicle-stimulating hormone; *GH*, growth hormone; *GHRH*, growth hormone–releasing hormone; *GnRH*, gonadotropin-releasing hormone; *IGFBP-3*, IGF binding protein-3; *IGF-1*, insulin-like growth factor-1; *LH*, luteinizing hormone; *MW*, molecular weight; *PRIH*, prolactin release–inhibiting hormone; *SRIH*, somatotropin release–inhibiting hormone; *TRH*, thyrotropin-releasing hormone; *TSH*, thyroid-stimulating hormone.

HYPOTHALAMIC REGULATION

Hormones released by the hypothalamus that regulate the anterior pituitary hormones are listed in Table 40.2. The secretion of **corticotropin-releasing hormone (CRH)** is stimulated by systemic physiologic stress. Physiologically, stress, inflammation, and hypoglycemia stimulate ACTH release. The energy state and temperature of the organism influence the secretion of thyrotropin-releasing hormone (TRH). Acute inflammatory cytokines such as interleukin (IL)-1β, IL-6, and tumor necrosis factor (TNF)-α stimulate ACTH release. Stimulators of GH–releasing hormone (GHRH) release include dopamine- and galanin-secreting neurons and brain stem neurons with catecholaminergic inputs. Physiologically, amino acids and hypoglycemia stimulate GH release. In turn, the secretion of IGF-1 in response to GH is influenced by nutrition, sex steroids, thyroid hormone, and the presence or absence of chronic disease. Malnutrition, sex hormone deficiency in adolescents and adults, hypothyroidism, and chronic disease all produce GH resistance to varying extents. The regulation of gonadotropin-releasing hormone (GnRH) is complicated by the fact that GnRH must differentially control the secretion of LH and FSH, which varies greatly during the menstrual cycle. GnRH pulsatility is essential to gonadotroph responsiveness. Tonic release of GnRH will downregulate GnRH receptors on gonadotrophs, which leads to hypogonadism. The rate of pulsatility may influence the relative secretion of LH and FSH.

Anterior Pituitary

In the anterior pituitary, (1) CRH receptors are expressed on corticotrophs, (2) thyrotropin-releasing hormone (TRH) receptors are expressed on thyrotrophs, (3) GHRH receptors are expressed on somatotrophs, (4) GnRH receptors are expressed on gonadotrophs, and (5) prolactin-release inhibiting hormone (PRIH) receptors are expressed on lactotrophs. In pathologically high concentrations, TRH stimulates the release of LH and prolactin. Corticotrophs are stimulated by high concentrations of proinflammatory cytokines, such as IL-1, IL-6, and TNF-α.

The hormonal products of each anterior pituitary target cell (if applicable) are listed in Table 40.1 together with a summary of each system. Deficiency of an individual pituitary hormone is typically called hypopituitarism, whereas deficiency of all anterior pituitary hormones is termed panhypopituitarism.

Growth Hormone and Insulin-Like Growth Factors

Linear growth is the consequence of (1) genetic potential, (2) nutrition, (3) the presence or absence of disease, and (4) hormonal effects. Many hormones influence growth, but the most important are GH, thyroid hormone, and sex steroids. Excess glucocorticoids can impair growth in children. GH deficiency produces morbidity in adults, thus GH is essential for health throughout life.

Growth Hormone

Two circulating forms of GH are present: a 22-kDa form that represents 85% to 90% of circulating GH and a 20-kDa GH that results from alternative splicing of the GH mRNA transcript. Circulating GH also exists as aggregates and oligomers. "Big GH" is a dimer of GH monomers, and "big, big GH" is GH associated with its binding protein (GHBP). GHBP is the external domain of the GH receptor (GHR), which binds GH with high affinity and is produced by cleavage of the GHR (Fig. 40.3).

GH has both indirect and direct activity. The indirect activity of GH is mediated by IGF-1. To initiate its direct and indirect activity, GH binds to receptors (GHR) that appear to be expressed by all tissues.

Insulin-Like Growth Factors

Proinsulin, **insulin-like growth factor** (IGF)-1, IGF-2, and relaxin are all composed of two domains (A and B) joined by a connecting domain. Less than 1% of the total IGF-1 is free (the biologically active form).

Most of the circulating IGF-1 is produced by hepatocytes, but it is also produced locally throughout the body and thus acts as a paracrine and an autocrine hormone.

Similar to IGF-1, IGF-2 does not normally produce hypoglycemia. However, tumors that secrete a larger than normal form of IGF-2 and can cause hypoglycemia have been described (see Chapter 22).

Receptors for Insulin-Like Growth Factors

Two types of receptors for IGFs have been identified: type I and type II. Structurally, the type I IGF receptor is similar to

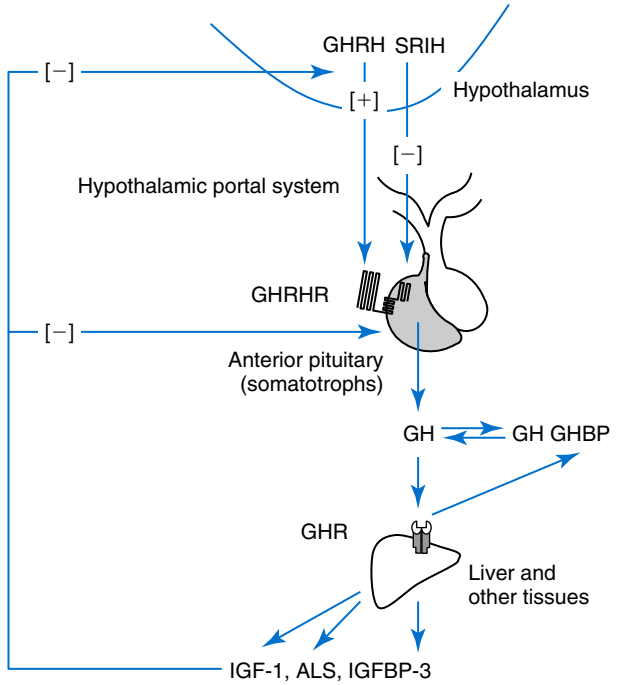


Fig. 40.3 The hypothalamus secretes growth hormone–releasing hormone *(GHRH)* and somatotropin release–inhibiting hormone (*SRIH*; somatostatin), which regulate growth hormone *(GH)* release. The receptor for GHRH *(GHRHR)* is illustrated because mutations in this receptor can cause some forms of inherited GH deficiency. GH circulates unbound and bound to its binding protein *(GHBP)*. GHBP is the extracellular domain of the GHR, which is cleaved from the GHR and circulates in the plasma. GH releases IGF-1, IGF binding protein-3 *(IGFBP-3)*, and the **acid-labile subunit (ALS)**. IGF-1 negatively feeds back at the anterior pituitary somatotrophs and at the hypothalamus. Because transection of the pituitary stalk leads to GH deficiency, the predominant hypothalamic control of GH is stimulatory via GHRH.

the insulin receptor. Because this receptor does not exclusively bind IGF-1, the terminology "IGF-1 receptor" is not recommended.

The type I IGF receptor binds IGF-1 with higher affinity than IGF-2, and affinity for insulin is lower than the affinity of the receptor for IGF-1 or IGF-2. The affinities of the insulin receptor are the opposite: insulin ≫IGF-2 >IGF-1. The type II IGF receptor is structurally dissimilar from the type I IGF receptor and the insulin receptor. The type II IGF receptor is monomeric and similar to the epidermal growth factor (EGF) receptor.

Regulation of Growth Hormone Secretion

GH ultimately stimulates the release of IGF-1, which negatively feeds back to regulate GH release via two hypothalamic hormones: somatotropin release–inhibiting hormone (SRIH) (or somatostatin) and GHRH (Fig. 40.3). Somatotrophs of the anterior pituitary gland have receptors for both hormones. Somatostatin inhibits GH release, whereas GHRH promotes the release of GH. In addition to its hypothalamic negative feedback effects, IGF-1 directly suppresses pituitary release of GH.

Physiologically, GH secretion is episodic and pulsatile. Consequently random measurements cannot exclude GH deficiency or confirm GH excess; between pulses, GH concentrations can be quite low and do not distinguish GH insufficiency from normal GH production. During daytime hours, the plasma concentration of GH in healthy adults remains stable and relatively low (<2 ng/mL; <2 μg/L), with several secretory spikes occurring approximately 3 hours after meals (particularly meals high in protein and arginine) and after exercise. In contrast, during the evening hours, adults and children show a marked rise in GH concentrations approximately 90 minutes after the onset of sleep, reaching a peak during the period of deepest sleep. GH is also increased by stress and hypoglycemia. Normal GH secretion requires thyroxine and age-appropriate concentrations of testosterone or estrogen.

GH is suppressed by increased blood glucose. One of the tests for GH excess measures GH after an oral glucose load (e.g., 75 g in adults and 1.75 g/kg in children); the normal response is a GH of less than 1 to 2 ng/mL (1–2 μg/L), depending on the lower limit of detection of the GH immunoassay (Box 40.1). GH also declines with increases in free fatty acid concentrations and aging. In the presence of abnormally high concentrations of glucocorticoids, GH secretion is suppressed.

Physiologic Actions

GH *directly* raises blood glucose by stimulating gluconeogenesis and reducing insulin sensitivity. Also, GH causes adipose tissue lipolysis, and the resulting free fatty acids provide an alternative energy source. Therefore, when glucose and free fatty acid concentrations are raised with stress—in

BOX 40.1 Protocol for Glucose Suppression of Growth Hormone Test

Rationale

Normal subjects show suppression of serum growth hormone (GH) concentrations after oral administration of glucose. Subjects with acromegaly fail to exhibit appropriate GH suppression.

Procedure

The test should be performed after an overnight fast with the patient maintained at bed rest. After a baseline blood specimen is collected for GH and glucose measurement, a solution of 75 g of glucose is given orally (in children, 1.75 g/kg to a maximum dose of 75 g). Glucose and serum GH are measured again on specimens collected 30, 60, 90, and 120 min later.

Interpretation

Serum GH concentrations in normal individuals fall to less than 1 to 2 ng/mL (1 to 2 μg/L) (i.e., below the lower limit of detection of the GH assay used). Subjects with acromegaly fail to show this suppression and sometimes show a paradoxical increase in GH concentration. Patients with liver disease, uremia, or heroin addiction may have false-positive results with this test (failure to suppress serum GH concentrations after oral glucose load).

partnership with epinephrine, glucagon, and cortisol—fuels for the fight-or-flight response are provided. GH directly stimulates the production of IGF-1, **insulin-like growth factor binding protein** (IGFBP)-3, and the ALS.

Indirect effects of GH are mediated through IGF-1 production, which (together with GH) is necessary for linear growth in childhood and promotes growth in soft tissue, cartilage, and bone. IGF-1 is mitogenic and antiapoptotic. GH (through IGF-1) and insulin induce growth in a similar manner because both have protein anabolic effects and stimulate the transport of amino acids into peripheral cells. Their respective effects on glucose homeostasis, however, oppose each other.

IGF-1 concentrations vary widely with age and gender. IGF-1 rises during childhood and puberty and then shows a gradual decline, reaching a steady state in the third decade of life.

GH is not the only determinant of IGF-1 concentration in the circulation. Therefore a decreased IGF-1 concentration is not necessarily synonymous with GH deficiency.

Clinical Significance

Clinically important states of GH excess or deficiency are uncommon and are often difficult to diagnose. GH measurements are best determined as part of dynamic testing: physiologic or pharmacologic provocative stimuli are used to diagnose GH deficiency, whereas GH suppression (or lack thereof) following glucose administration is useful in evaluating GH excess.

A single measurement of IGF-1, which has a much longer half-life than GH, reflects GH-IGF-1 production more accurately irrespective of the time of the day or meals.

Growth Hormone Excess

Growth hormone excess results in **acromegaly** in adults. Even less common is **gigantism**, which results from GH excess in childhood. The clinical features of acromegaly involve overgrowth of the skeleton and soft tissue, producing (1) acral (extremity) enlargement, (2) organomegaly (enlarged heart and/or liver), (3) facial coarsening, (4) intestinal polyposis, (5) premature cardiovascular disease, (6) increased sweating, (7) skin tags, (8) joint disease, (9) myopathy with weakness, (10) insulin resistance, and often (11) diabetes mellitus. Premature cardiovascular disease is the most common cause of death from acromegaly. Gigantism is characterized by extreme tall stature in addition to the clinical features of acromegaly, as pathologic GH excess occurs before epiphyseal fusion is complete.

Most cases of acromegaly (≈95%) result from anterior pituitary GH-secreting tumors (somatotropinomas). Somatotropinomas are usually large macroadenomas (>10 mm in diameter) by the time they come to clinical attention. Approximately 5% of GH-secreting tumors are familial and caused by disorders such as multiple endocrine neoplasia type 1 syndrome, familial acromegaly, Carney syndrome, McCune-Albright syndrome, and familial isolated pituitary adenoma. Approximately 5% of acromegaly cases result from GHRH-secreting hypothalamic tumors, and less than 1% result from extrapituitary somatotropinomas.

In severe or advanced cases of GH excess, the diagnosis may be established on the basis of physical appearance alone. In less severe or early cases, the physical changes may be subtle and gradual, so a high degree of clinical suspicion is needed. The reversibility of tissue changes depends largely on the duration of the disease. In addition to soft tissue changes, acromegaly may cause severe disability or death from cardiac, pulmonary, or neurologic sequelae. The most important requirement for the diagnosis of acromegaly is the demonstration of inappropriate and excessive GH secretion.

Essentially all patients with acromegaly have an abnormal GH response to oral glucose load (see Box 40.1). Patients with acromegaly typically show no change in their basal GH level or demonstrate a paradoxical increase in GH; in contrast, normal individuals show suppression of GH after a 75-g oral glucose load. Serum IGF-1 is also elevated in acromegaly.

Growth Hormone Deficiency and Growth Retardation

GH deficiency is not a common cause of short stature with a low growth velocity. Approximately 50% of children evaluated for growth retardation have no specific organic cause; approximately 15% have an endocrine disorder, of which approximately half have GH deficiency. Children with significantly reduced height and low growth velocities with no clear explanation should be screened for GH deficiency after other endocrine disorders have been excluded.

GH deficiency in children is characterized by (1) short stature, (2) low growth velocity, (3) immature facial appearance, (4) retarded bone age on radiologic examination, and (5) increased adiposity. In cases of congenital GH deficiency, size at birth is usually normal because in utero IGF-1 does not appear to be under GH control. GH replacement therapy forms an important part of the clinical care of GH-deficient children.

Adults with GH deficiency experience (1) reduced muscle mass, (2) increased central adiposity, (3) osteoporosis with decreased bone density, (4) an increase in fractures, (5) decreased quality of life, (6) dyslipidemia, and (7) an increased risk for cardiovascular disease. Whether GH therapy is required in GH-deficient adults is somewhat controversial.

GH insufficiency can be a consequence of (1) hypothalamic disease, (2) disruption of the portal system between hypothalamic nuclei and the anterior pituitary, (3) GHRHR loss-of-function mutations, or (4) somatotroph disease. GH deficiency can occur in isolation or together with other pituitary deficiencies. Patients with isolated GH deficiency should be followed clinically for the development of other pituitary hormone deficiencies because multiple pituitary hormone deficiencies can evolve over time. In most affected children, the cause of GH deficiency is unknown (idiopathic GH deficiency). Approximately one in four children with proven GH deficiency has an organic origin for his or her GH deficiency; half of these children will be diagnosed with a central nervous system (CNS) tumor. Biochemical stimulation testing is necessary to establish the diagnosis of GH deficiency, GH resistance, or multiple pituitary hormone deficiencies.

Investigation of growth hormone deficiency

Children. A staged approach for the evaluation of GH deficiency in children is advised. Initial screening can involve one of the following tests: (1) measurement of IGF-1 (with or without IGFBP-3) or (2) GH measurement after exercise or pharmacologic stimulation. If a GH screening test is abnormal (e.g., the GH and/or IGF-1 concentration is reduced), definitive testing should be pursued. All forms of GH testing should be performed after the subject has fasted overnight.

Exercise physiologically enhances GH release. A baseline GH measurement is not required. The child exercises vigorously for approximately 20 minutes and GH is then measured. GH is also measured 20 minutes later to capture delayed GH release. Screening for GH deficiency can also be performed by measuring GH 60 to 90 minutes following clonidine, glucagon, or L-dopa administration. (Note: L-dopa is not currently available in the United States.)

If IGF-1 is within its reference interval for age and sex in children, GH deficiency is excluded. If IGF-1 is low, definitive GH testing is required. Because IGF-1 concentrations can be depressed in states other than GH deficiency, a low IGF-1 concentration does not confirm GH deficiency. IGFBP-3 is less dependent on good nutrition to achieve normal concentrations, so it may be a better marker of GH deficiency than IGF-1.

GH responses to insulin-induced hypoglycemia (insulin tolerance test [ITT]) and GH responses to centrally acting pharmacologic or biologic agents are considered definitive tests. The stimuli are usually administered sequentially or, less commonly, on different days. The classic diagnosis of pediatric GH deficiency requires that GH responses to two different stimuli be deficient. Stimulated GH concentrations less than 7 to 10 ng/mL (7 to 10 μg/L) define GH deficiency in children. The cutoff defining GH deficiency is method-dependent and subject to interpretation by clinicians.

Of children with appropriate stature for age, approximately 80% will have a normal GH response to one stimulus and 95% will have normal GH responses to one or both of the two stimuli.

Adults. A history of childhood GH deficiency, CNS disease, trauma, or irradiation is an indication to test adults for GH deficiency, because not all adults with childhood GH deficiency remain GH-deficient as adults. In adults, a single deficient GH response to a stimulus is diagnostic of GH deficiency if the deficiency is congenital or genetic or if multiple pituitary hormone deficiencies are due to organic disease. In adults, GH deficiency is present when stimulated GH is generally less than 3 to 5 ng/mL (3 to 5 μg/L).

IGF-1 measurements in adults are often not diagnostically helpful. For reasons that are unclear, IGF-1 concentrations can be normal in GH-deficient adults. If the IGF-1 concentration is low and suspicion for GH deficiency is high, some experts would diagnose GH deficiency in the absence of GH testing in adults.

Growth Hormone Resistance

In children with short stature and low growth velocity, if (1) IGF-1 is below the reference interval for the child's bone age and gender, (2) the random GH concentration is normal or elevated, and (3) non–GH-dependent causes of IGF-1 deficiency have been excluded, GH resistance should be considered. GH resistance is rare.

GH resistance can be congenital, resulting from loss-of-function GHR mutations or GHR signaling defects (STAT5b mutations) or from defects in the production of IGF-1 itself. Most cases of GHR deficiency display low or absent concentrations of GHBP. In acquired GH resistance, which is far more common, the IGF-1 is low (despite sufficient GH secretion) because of malnutrition, malabsorption, chronic disease, hypothyroidism, or sex hormone deficiency.

POINTS TO REMEMBER

1. Secretion of growth hormone (GH) by the pituitary gland is episodic and pulsatile, and transient elevations have been observed in normal healthy subjects.
2. A single basal or random concentration of GH provides limited diagnostic information.
3. The diagnosis of GH deficiency usually requires repeated stimulation tests, although some endocrinologists will diagnose GH deficiency based solely on the measurement of IGF-1.
4. GH excess is diagnosed by the failure of GH suppression after an oral glucose load and by elevated IGF-1 levels.

Prolactin

Prolactin is secreted by lactotrophs. Prolactin stimulates and sustains lactation in postpartum mammals after the mammary glands have been prepared by other hormones, including estrogens, progesterone, GH, corticosteroids, and insulin.

Biochemistry

The hypothalamic PRIH is dopamine. Occasionally and independent of the presence or absence of disease, IgG autoantibodies can bind to prolactin, forming macroprolactin, which is biologically inactive. Macroprolactin elevates the total prolactin concentration as the result of lower clearance. Failure to recognize macroprolactin can lead to the inappropriate diagnosis of hyperprolactinemia.

Physiology

Prolactin is necessary for lactation after delivery of the newborn. Prolactin is stimulated by breast-feeding, chest wall disease, and stress. Prolactin secretion by lactotrophs is controlled predominantly through suppression by PRIH.

It is unclear if prolactin has a physiologic function in men. Postpartum, a positive feedback loop occurs between suckling and milk production. Transmitted via nerve fibers from the nipple to the CNS, suckling reduces PRIH, which increases prolactin release. The positive feedback loop of suckling, prolactin secretion, and milk production is a "stimulus secretion" reflex. However, with continued breast-feeding, prolactin concentrations decline.

Hyperprolactinemia

Hyperprolactinemia is the most common hypothalamic-pituitary disorder encountered in clinical endocrinology. Prolactin may be elevated in women who have only subtle alterations in fertility, such as (1) anovulation with or without menstrual irregularity, (2) amenorrhea and galactorrhea, or (3) galactorrhea alone. In men, prolactin excess usually manifests as a result of low serum testosterone, with reduced libido and central weight gain. Hyperprolactinemia can also cause galactorrhea in men.

An irregular menstrual period may reveal a microadenoma (≤10 mm in diameter) in women. Elevated prolactin concentrations are observed in as many as 30% of women with polycystic ovarian syndrome and patients with clinically silent pituitary adenomas. If a borderline elevation of prolactin is found, it is advisable to repeat the measurement on at least two other occasions, taking care to obtain a morning specimen under conditions of minimal excitement or stress to the patient (i.e., no trauma and no breast stimulation). Ideally, the patient should not be on any medication that could stimulate prolactin release (see later in this section). The differential diagnosis of hyperprolactinemia is extensive (Table 40.3).

The higher the prolactin concentration, the greater the likelihood that hyperprolactinemia is the result of a prolactinoma. Idiopathic hyperprolactinemia is a diagnosis of exclusion. Early diagnosis of a prolactinoma is critical because therapy with oral dopamine agonists, such as bromocriptine, can reduce tumor size and control tumor progression. Surgical excision of a prolactinoma is usually considered if there is tumor growth or dopamine agonist therapy fails (the "dopamine-resistant prolactinoma"), if there is failure to quickly reverse associated visual loss, or if the patient is intolerant to dopamine agonists.

A diagnostic challenge in the investigation of hyperprolactinemia is the presence of a pituitary incidentaloma (a radiologic image consistent with a mass that is not necessarily pathologic) in a patient with elevated prolactin that could potentially lead to inappropriate medical or surgical treatment. Unless a pituitary tumor can be demonstrated by magnetic resonance imaging (MRI), prolactin-secreting microadenoma (≤10 mm in diameter) is diagnosed by exclusion. Because half of all prolactin-secreting microadenomas are too small to be detected by imaging methods, differentiating between a small pituitary tumor, prolactin-cell hyperplasia, and idiopathic hyperprolactinemia may not be possible.

Medications that stimulate prolactin release (through PRIH suppression) are the most common cause of hyperprolactinemia in otherwise healthy individuals. When a significant elevation of prolactin is confirmed, a careful history should rule out the possibility that medications are the cause. A pregnancy test should be performed in women of reproductive age because pregnancy is a cause of hyperprolactinemia.

TABLE 40.3 Differential Diagnosis of Hyperprolactinemia

PRIH (dopamine) deficiency	Hypothalamic disease
	Interruption in the hypothalamic-pituitary portal system
Drugs	Dopamine antagonists
	Cholinergic antagonists
	Serotonergic antagonists
	Antipsychotic medication
Hormones	Estrogen, pregnancy
Neurogenic	Nursing (nipple stimulation)
	Chest wall disease
	Spinal cord injury
Other diseases	Hypothyroidism (pathologically elevated TRH can release prolactin)
	Chronic renal disease
	Cirrhosis

PRIH, Prolactin release–inhibiting hormone; *TRH,* thyrotropin-releasing hormone.

Prolactin Deficiency

Prolactin is of great clinical importance in the postpartum period because it is required for lactation. However, other than the necessity of breast-feeding, prolactin deficiency in humans may not have adverse consequences.

POINTS TO REMEMBER

1. Compared with the measurement of other anterior pituitary hormones, prolactin can be measured in the absence of stimulation or suppression.
2. There are many causes of hyperprolactinemia that should be considered in the evaluation of hyperprolactinemia.
3. The first line of treatment for prolactinomas is medical therapy through the administration of drugs with dopamine-like action.

Adrenocorticotropic Hormone and Related Peptides

ACTH is secreted by the adenohypophysis as a derivative of pro-opiomelanocortin (**POMC**). ACTH acts on the adrenal cortex, stimulating its growth and the secretion of corticosteroids (specifically cortisol). ACTH production is increased during physiologic or psychologic stress.

Many variables affect the secretion of ACTH, which is both pulsatile and circadian in nature. Cortisol is the major negative feedback hormone for the inhibition of CRH and ACTH secretion. Regulation of ACTH—and a discussion of adrenal disorders including disorders of ACTH secretion—is found in Chapter 41.

Gonadotropins (Follicle-Stimulating Hormone, Luteinizing Hormone)

LH and FSH are synthesized by gonadotrophs. Their actions have already been described. LH and FSH control the functional activity of the gonads. In males and females, gonadotropin secretion is regulated via GnRH. For more details on FSH and LH action and regulation, refer to Chapter 43.

POINTS TO REMEMBER

1. Luteinizing hormone (LH) and follicle-stimulating hormone (FSH) are secreted from pituitary gonadotrophs in response to pulsatile gonadotropin-releasing hormone (GnRH).
2. FSH stimulates growth of ovarian follicles and the secretion of estrogen (estradiol) in females and spermatogenesis in males.
3. LH in females assists in the formation of the corpus luteum and together with FSH promotes progesterone secretion; in males, LH stimulates Leydig cells to produce testosterone.

Thyroid-Stimulating Hormone

TSH, which is synthesized in thyrotrophs, promotes the growth of thyroid follicular cells and sustains and stimulates the hormonal secretion of thyroid gland hormones 3,5,3′,5′ tetraiodothyronine (thyroxine; T4) and 3,5,3′-triiodothyronine (T3).

TSH binds to TSH receptors (TSHRs) located on the surfaces of thyroid follicular cells. TSH (1) stimulates growth and vascularity of the thyroid gland, (2) stimulates the growth of thyroid follicular cells, (3) promotes thyroid hormone synthesis by increasing the uptake of iodine (via the sodium-iodide transporter), (4) promotes the organification (reduction) of iodine, (5) promotes the coupling of tyrosines, and (6) promotes the proteolytic release of stored thyroid hormone from thyroglobulin. TSH release is stimulated by TRH and suppressed by thyroid hormone (principally circulating free T4). TRH is a modified tripeptide produced by the hypothalamus. Details concerning the regulation and clinical significance of these hormones are discussed in detail in Chapter 42.

Antidiuretic Hormone

Disorders of ADH involve excess hormone (**syndrome of inappropriate ADH [SIADH]**) or deficient ADH action (**diabetes insipidus** [DI]). DI can result from ADH deficiency, ADH resistance, or renal tubular disease; the latter two conditions are termed nephrogenic DI. Disorders of oxytocin secretion have not been described; therefore oxytocin is not discussed further in this chapter.

Biochemistry

ADH is a nonapeptide consisting of a cyclic hexapeptide and a three-amino-acid side chain. The ADH receptor in the renal collecting ducts is termed the arginine **vasopressin** receptor 2 (V2 receptor). The action of ADH is to stimulate the movement of aquaporin-2 from the cytoplasm to the apical plasma membrane (Fig. 40.4). In this way ADH allows the

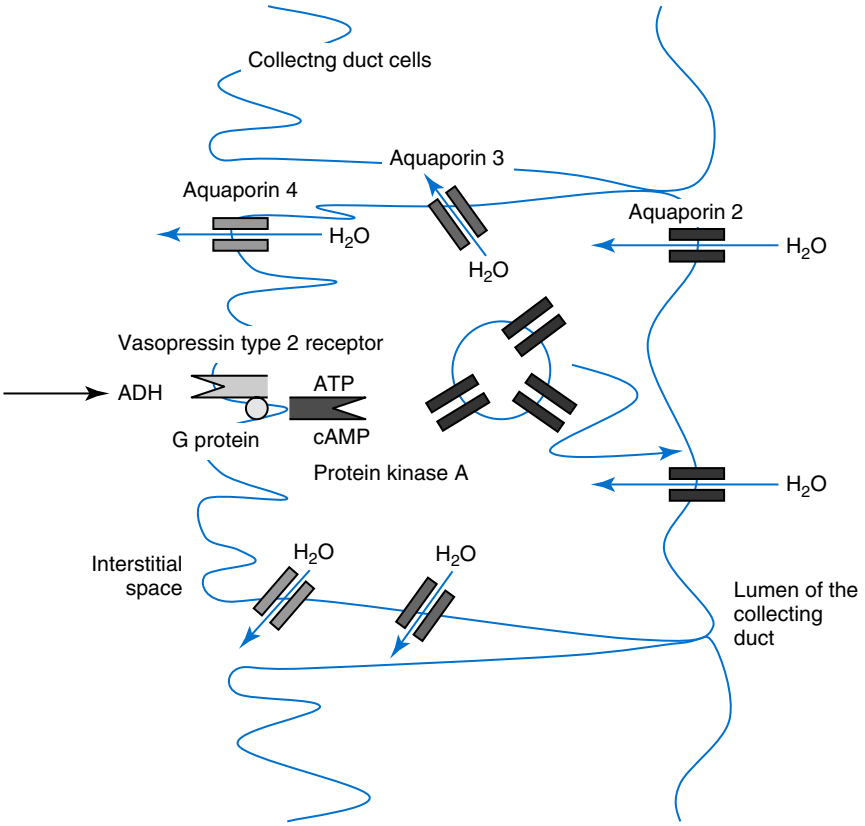


Fig. 40.4 Vasopressin type 2 receptors on collecting duct cells bind antidiuretic hormone *(ADH)*. Through a G-protein system, adenosine triphosphate *(ATP)* is converted to cyclic adenosine 3′,5′-monophosphate *(cAMP)* via adenylate cyclase with protein kinase A activation. This leads to translocation of aquaporin-2 water channels from an intracellular pool to the apical plasma membrane, allowing free water uptake by cells of the collecting duct. Via the basolateral plasma membrane aquaporin-3 and aquaporin-4 water channels, free water then leaves these cells.

reabsorption of water from the collecting duct. As electrolyte is not also reabsorbed, the action of ADH can be described as increasing "free water" reabsorption.

Regulation of Antidiuretic Hormone Secretion

ADH secretion is controlled predominantly by plasma osmolality, which is sensed by osmoreceptors in the hypothalamus. Increased osmolality results in ADH release. Plasma osmolality above 280 mOsm/kg (mmol/kg) is thought to be the osmotic threshold for triggering ADH release.

In addition to the osmoreceptor mechanism of vasopressin release, physiologic regulation of ADH secretion involves a pressure–volume mechanism that is distinct from the osmotic sensor. High-pressure arterial baroreceptors of the aortic arch and carotid sinus and low-pressure volume receptors in the pulmonary venous system and atria also regulate ADH release. Therefore ADH is secreted in response to decreased circulating blood volume or decreased blood pressure.

The thirst center is regulated by many of the same factors that determine ADH release. This center has a higher set point than the osmoreceptors and responds to osmolalities above 290 mOsm/kg. Responses involving ADH, thirst, and renal reabsorption of sodium and water are coordinated in a complex scheme that maintains plasma osmolality in healthy individuals within a narrow interval (~285 to ~295 mOsm/kg [mmol/kg]).

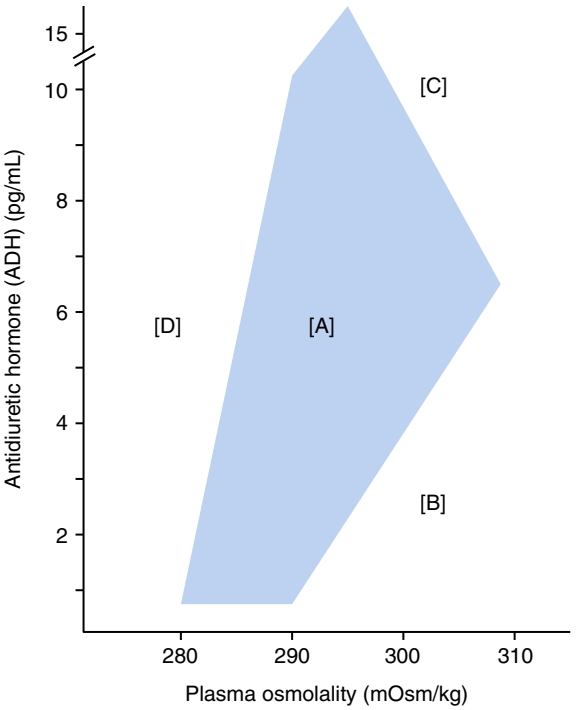


Fig. 40.5 Antidiuretic hormone (ADH) concentration versus plasma osmolality is plotted. The shaded area *[A]* depicts the normal physiologic relationship between ADH and plasma osmolality. Area *B* illustrates cases of diabetes insipidus with ADH deficiency. Area *C* encompasses cases of ADH resistance. Area *D* includes cases of SIADH (syndrome of inappropriate ADH secretion).

Physiologic Activity

The actions of ADH are to conserve free water (via V2 receptors) and stimulate vasoconstriction (via V1a receptors). These effects combine to maintain proper osmolality of the extracellular space (the major action of ADH) and blood pressure through maintenance of circulating blood volume and prevention of dehydration and excessive loss of water.

ADH increases the permeability of renal collecting ducts to water, thereby increasing water reabsorption and concentrating the urine to a higher specific gravity.

Clinical Significance

Disorders of ADH activity have been divided into hypofunction (DI) and hyperfunction (SIADH) (Fig. 40.5).

Polyuric states and diabetes insipidus. Exclusive of diabetes mellitus, polyuric states are divided into two main categories: (1) deficient ADH action, producing DI (sometimes called "water diabetes"), and (2) excessive oral water intake (psychogenic **polydipsia**). Osmotic diuresis (e.g., secondary to hyperglycemia) may also produce **polyuria** and polydipsia.

In DI, polyuria results from excessive loss of water into the urine. Under normal circumstances, urine output is predominantly dependent on fluid intake. When urine output is greater than 2.5 L/d, investigation is usually indicated. In the absence of ADH, urine output may approach 1 L/h. Increased osmolality normally stimulates thirst. Therefore, if a patient with DI has an intact thirst mechanism and access to water, excessive urinary loss of water should be matched by excessive intake of fluids. The major laboratory finding in DI is urine of low osmolality and serum sodium within the upper half of the reference interval, with a corresponding high-normal serum osmolality. If water is not available or the individual with DI is physically impaired or lacks a normal thirst mechanism, serum osmolality rises, serum sodium rises (producing hypernatremia), urine osmolality remains low, and polyuria continues. With dehydration, weight loss is acute (because of fluid loss) and blood pressure falls, inducing tachycardia.

Central DI can result from any destructive hypothalamic lesion or infundibular lesion (Table 40.4). DI resulting from such lesions can equally be termed hypothalamic, neurogenic, central, or cranial DI. When the thirst center is also abnormal, severe dehydration can occur.

In 30% of patients, central DI occurs without apparent cause. The remaining cases are associated with (1) neoplastic disease, (2) neurologic surgery, (3) head trauma, (4) ischemic or hypoxic disorders, (5) granulomatous disease, (6) infection, or (7) autoimmune disorders. DI can result from acquired or genetic tubular diseases that affect the responsiveness of renal tubules to ADH (nephrogenic DI). In the broadest sense, any form of tubular injury with impaired water reabsorption, including (1) polycystic kidney disease, (2) medullary cystic kidney, (3) chronic pyelonephritis, (4) acute tubular necrosis, (5) obstructive uropathy, (6) sickle cell nephropathy, and (7) renal amyloidosis can cause nephrogenic DI. A large number

TABLE 40.4 Causes of Central Diabetes Insipidus

Congenital	Midline malformations: septo-optic dysplasia, holoprosencephaly, single central incisor, cleft lip and/or palate
	Malformation of the pituitary (ectopia or hypogenesis)
	Diabetes insipidus–diabetes mellitus–optic atrophy syndrome (Wolfram syndrome)
	Familial diabetes insipidus (autosomal dominant and recessive forms)
Acquired	Tumors (craniopharyngioma, germinoma, pinealoma, optic glioma, pituitary adenoma, metastatic tumor, leukemia)
	Trauma (e.g., stalk section)
	Infarction (e.g., septic shock, Sheehan syndrome, hypoxic injury)
	Infiltrative disease (sarcoidosis, hypophysitis, histiocytosis)
	Cysts and aneurysms
	Drugs (opiates, alcohol, phenytoin, alpha-adrenergic agents, etc.)
	Infection
	Increased metabolism of ADH (vasopressinase in pregnancy)

ADH, Antidiuretic hormone.

of drugs can cause nephrogenic DI, including (1) lithium, (2) various antimicrobials, and (3) several antineoplastic drugs. In psychogenic polydipsia (primary polydipsia), excessive water intake eventually begins to impair the concentrating ability of the kidney.

Investigation of polyuria. Assuming that diabetes mellitus is excluded, the examination of polyuric states is undertaken with measurements of plasma and urine osmolality and, if the findings are equivocal, plasma ADH concentrations. A screening test for ADH sufficiency is the measurement of the urine specific gravity in the first morning-voided urine sample where the specific gravity should be 1.010 or greater. Failure to demonstrate an appropriate urine specific gravity and response in a subject with a history of polyuria and polydipsia should trigger assessment of a basic metabolic panel, urinalysis, serum osmolality, and urine osmolality.

Urine osmolality less than 300 mOsm/kg (mmol/kg) combined with serum osmolality more than 300 mOsm/kg (mmol/kg) (or with hypernatremia) is diagnostic for DI. If urine osmolality is above 600 mOsm/kg (mmol/kg), and serum osmolality is below 270 mOsm/kg (mmol/kg), DI is unlikely.

If the diagnosis of DI is unclear, a water deprivation test should be performed. Water deprivation testing should not be carried out if the subject is dehydrated at baseline or has renal insufficiency. The overnight water deprivation test is usually conducted in a hospital setting because of the immediate concerns of profound hypotension and possible mortality. The water deprivation test is usually begun on the morning after an overnight fast unless the history describes large volumes of water ingested and urine produced, in which case the test should begin after breakfast. The subject remains fasting throughout the entire test. A heparin lock is inserted intravenously, so that serial blood samples are obtained easily. Baseline laboratory tests include (1) sodium, (2) potassium, (3) chloride, (4) serum carbon dioxide (CO_2), (5) urea (blood urea nitrogen [BUN]), (6) creatinine, (7) glucose, (8) calcium, and (9) serum osmolality. Potassium and calcium are measured to exclude hypokalemia and hypercalcemia as causes of nephrogenic DI (see later).

Measurements of serum or plasma sodium, chloride, CO_2, and urine pH provide an assessment of the patient's renal tubular acid-base function. Renal function and hydration status are evaluated with creatinine and urea (BUN) measurements. A urine specimen is obtained for measurement of urine sodium, urine osmolality, and urine specific gravity. Body weight and vital signs at baseline are recorded. Thereafter, each hour, serum and urine tests are repeated with measurement of hourly urine output, body weight, and urine volume. If ADH measurements are requested, they can be performed at the beginning, middle, and completion of the test; however, ADH measurements are not required for making the diagnosis of DI. The test is continued for 8 to 10 hours unless the diagnosis of DI is confirmed before the full time period has elapsed.

During the water deprivation test, if urine osmolality exceeds 600 mOsm/kg (mmol/kg) on two samples 1 hour apart or a single urine sample exceeds 1000 mOsm/kg (mmol/kg), DI is ruled out and the test can be concluded. If urine osmolality is less than 600 mOsm/kg (mmol/kg), and serum osmolality is more than 300 mOsm/kg (mmol/kg), DI is diagnosed and the test can be concluded. If serum osmolality does not exceed 300 mOsm/kg (mmol/kg), the test should be continued. If mental status changes or hypotension occurs or weight loss exceeds 3%, the test should be terminated.

If DI is diagnosed during a water deprivation test, aqueous vasopressin is injected subcutaneously (1 U/m^2 or, according to some sources, 5 U total), or des-amino-D-arginine vasopressin (desmopressin) is given intravenously or intramuscularly (2 μg). A decline in urine volume with doubling of urine osmolality over the next 1 to 2 hours identifies the DI as central in origin with the patient being ADH deficient. Failure to respond to exogenous ADH defines nephrogenic DI. Most patients with psychogenic polydipsia have normal urine osmolality after water deprivation, but some fail to produce concentrated urine unless the water deprivation is prolonged.

Syndrome of inappropriate antidiuretic hormone secretion. SIADH is a common cause of hyponatremia in hospitalized patients. The autonomous sustained production of ADH in the absence of recognized and appropriate stimuli (such as hyperosmolality) is termed SIADH. In this syndrome, plasma ADH concentrations are "inappropriately" elevated relative to decreased plasma osmolality and to normal or increased plasma volume.

TABLE 40.5 Causes of the Syndrome of Inappropriate Antidiuretic Hormone

CNS disease	Brain tumor Infection (e.g., meningitis, encephalitis, abscess) Prolonged seizure Psychiatric disease Stress (e.g., prolonged nausea)
Non-CNS tumor (e.g., leukemia)	
Pulmonary disease	Hypoxia (e.g., neonatal) Infection (e.g., pneumonia, emphysema)
Nonpulmonary infection (e.g., AIDS)	
Drugs	Drugs with CNS effects (anticonvulsants, antiparkinson drugs, antipsychotics, antipyretics, antidepressants) Angiotensin-converting enzyme inhibitors Antineoplastic drugs First-generation sulfonylureas

AIDS, Acquired immunodeficiency syndrome; *CNS*, central nervous system.

SIADH may be the result of one of several factors (Table 40.5), including (1) production of ADH by a malignancy (such as small cell carcinoma of the lung), (2) the presence of acute or chronic disease of the CNS, (3) pulmonary disorders, or (4) as a side effect of certain drug therapies.

In SIADH, primary excess of ADH coupled with unrestricted fluid intake promotes increased reabsorption of water by the kidney. The consequences are decreased urine volume and increased urine osmolality. The increase in intravascular volume causes hemodilution accompanied by dilutional hyponatremia and low plasma osmolality. It is important to recognize that although osmoregulation is tightly controlled, volume regulation always takes precedence over osmoregulation. Via suppression of the renin-angiotensin-aldosterone axis, volume expansion also decreases renal sodium reabsorption; thus the urine sodium concentration is typically 20 to 40 mEq/L (mmol/L) despite dilutional hyponatremia in the blood. This hypervolemia-induced urinary sodium excretion coupled with water loss explains why patients with SIADH are euvolemic and not edematous.

Clinical manifestations of hyponatremia are nonspecific and include nausea, weakness, and apathy in mild cases and CNS changes such as lethargy, coma, and seizures in more severe cases. No signs or symptoms are specific for SIADH. History, physical examination, and routine laboratory test results often suggest that hyponatremia is dilutional (decreased urea, hemoglobin, or albumin) or depletional (increased urea, hemoglobin, or albumin). Measurements of sodium and osmolality in blood and urine combined with clinical assessment of volume status usually permit the appropriate differential diagnosis of hyponatremic conditions.

TABLE 40.6 Disorders of Overproduction and Underproduction of Pituitary Hormones

Pituitary Hormone	Consequences of Hormone Excess	Consequences of Hormone Deficiency
ACTH	Cushing disease	Cortisol deficiency
TSH	Central hyperthyroidism	Central hypothyroidism
GH	Children: gigantism Adults: acromegaly	Children: short stature Adults: adult GH deficiency
LH, FSH	Alpha-chain overproduction	Hypogonadism
Prolactin	Galactorrhea, hypogonadism	Inadequate lactation or lactation failure in mothers after delivery
ADH	SIADH	DI

ACTH, Adrenocorticotropic hormone; *ADH*, antidiuretic hormone; *DI*, diabetes insipidus; *FSH*, follicle-stimulating hormone; *GH*, growth hormone; *LH*, luteinizing hormone; *SIADH*, syndrome of inappropriate ADH; *TSH*, thyroid-stimulating hormone.

Diagnostic Studies

Serum osmolality can be estimated from serum sodium, glucose, and urea (BUN) but is more reliably measured directly. Various formulas are available for estimating serum osmolality (see Chapter 24 for detail).

POINTS TO REMEMBER

1. Antidiuretic hormone (ADH) and oxytocin-synthesizing cells are located in the hypothalamus.
2. ADH deficiency or resistance causes diabetes insipidus (DI) with the excess loss of free water.
3. ADH excess causes the syndrome of inappropriate ADH (SIADH), which results in the excess retention of free water.

SUMMARY OF PITUITARY-RELATED DISORDERS

Disorders that result from the over- or underproduction of pituitary hormones are tremendously diverse (Table 40.6). Pituitary adenomas may be secretory or nonsecretory. Corticotropinomas secrete ACTH, somatotropinomas secrete GH, and prolactinomas secrete prolactin. Gonadotropinomas usually do not secrete intact LH and/or FSH but may secrete free α subunits. Pituitary adenomas, whether functional or not, can lead to deficiencies of other pituitary hormones through the compression and destruction of the adjacent pituitary cells. These topics are reviewed in depth in many chapters of this textbook.

REVIEW QUESTIONS

1. Which one of the following anterior pituitary hormones is regulated through suppression via the secretion of a hypothalamic hormone?
   a. ACTH
   b. Prolactin
   c. Growth hormone
   d. LH, FSH
   e. IGF-1
2. Which one of the following hormones exerts a predominant negative feedback via the anterior pituitary? The other hormones predominantly feed back at the level of the hypothalamus.
   a. IGF-1
   b. Sex steroids
   c. Inhibins
   d. Cortisol
   e. Thyroid hormone
3. Which one of the following statements is correct?
   a. IGF-1 and IGF-2 circulate bound to albumin.
   b. IGF-1 and IGFBP-3 levels are independent of one another.
   c. IGF-1 and IGF-2 circulate together bound to IGFBP-3 and the acid-labile subunit (ALS).
   d. The acid-stable subunit binds to IGF-1.
   e. IGF-1 and insulin form a complex.
4. Which of the following statements is correct regarding growth hormone (GH)?
   a. Physiologically, GH secretion is episodic and pulsatile.
   b. GH concentrations reach a very low levels during the period of deepest sleep.
   c. GH actions are mediated through cortisol.
   d. GH secretion is stable throughout the day.
   e. A single measurement of GH is sufficient for diagnostic purposes.
5. Which one of the following statements is true?
   a. GH lowers blood glucose and can cause hypoglycemia.
   b. GH raises blood glucose by stimulating gluconeogenesis and reducing insulin sensitivity.
   c. GH is suppressed by hypoglycemia.
   d. GH inhibits lipolysis.
   e. GH is the only determinant of IGF-1 concentrations.
6. What is the most common case of acromegaly?
   a. IGF-1–secreting tumors
   b. Macro GH
   c. Anterior pituitary GH-secreting tumors (somatotropinomas)
   d. Multiple endocrine neoplasia type 1 syndrome
   e. Gonadotropin-secreting tumors
7. Which one of the following statements is true regarding macroprolactin?
   a. This is a complex between prolactin and an immunoglobulin.
   b. This is a very high-molecular-weight form of prolactin due to alternative splicing.
   c. Macroprolactin lowers the plasma prolactin concentration.
   d. Macroprolactin heralds a dangerous pituitary tumor.
   e. Macroprolactin cannot be detected in the laboratory.
8. Which one of the following statements is true?
   a. In men, FSH stimulates testosterone production by Leydig cells. LH directs Sertoli cells to nourish developing sperm during spermatogenesis.
   b. LH and FSH are secreted by separate and distinct pituitary cells.
   c. LH and FSH secretion is stimulated by steady high concentrations of GnRH.
   d. LH and FSH are single-chain polypeptides.
   e. LH and FSH are secreted episodically.
9. Which one of the following statements is true?
   a. Cortisol is directly regulated by CRH.
   b. The major site of cortisol feedback is the anterior pituitary gland.
   c. ACTH directly regulates CRH.
   d. Increased ACTH increases cortisol secretion.
   e. ACTH is a glycoprotein hormone made up of alpha and beta subunits.
10. Which of the following suggests diabetes insipidus (DI)?
   a. Urine osmolality >600 mOsm/kg (mmol/kg)
   b. Serum osmolality <270 mOsm/kg (mmol/kg)
   c. Urine osmolality <300 mOsm/kg (mmol/kg) plus serum osmolality >300 mOsm/kg
   d. Edema and hypertension
   e. Very high urine sodium

SUGGESTED READINGS

Agarwal, M., Das, A., & Singh, A. S. (2010). High-dose hook effect in prolactin macroadenomas: a diagnostic concern. *Journal of Human Reproductive Sciences*, *3*, 160–161.

Alter, D., & Winter, W. E. (2016). Chapter 37. Disorders of the Adrenal Cortex and Medulla. In W. A. Clarke (Ed.), *Contemporary practice in clinical chemistry* (pp. 535–556). Washington, DC: AACC Press.

Amar, A. P., & Weiss, M. H. (2003). Pituitary anatomy and physiology. *Neurosurgery Clinical of North America*, *14*, 11–23, v.

Ascoli, P., & Cavagnini, F. (2006). Hypopituitarism. *Pituitary*, *9*, 335–342.

Ball, S. G. (2007). Vasopressin and disorders of water balance: the physiology and pathophysiology of vasopressin. *Annals of Clinical Biochemistry*, *44*, 417–431.

Bernard, V., Young, J., Chanson, P., et al. (2015). New insights in prolactin: pathological implications. *Nature Reviews Endocrinology*, *11*, 265–275.

Guillemin, R. (2005). Hypothalamic hormones a.k.a. hypothalamic releasing factors. *Journal of Endocrinology*, *184*, 11–28.

Ho, K. K. (2007). Consensus guidelines for the diagnosis and treatment of adults with GH deficiency II: a statement of the GH Research Society in association with the European Society for Pediatric Endocrinology, Lawson Wilkins Society, European Society of Endocrinology, Japan Endocrine Society, and Endocrine Society of Australia. *European Journal of Endocrinology*, *157*, 695–700.

Mullis, P. E. (2007). Genetics of growth hormone deficiency. *Endocrinology Metabolism Clinical of North America*, *36*, 17–36.

LeRoith, D. (2008). Clinical relevance of systemic and local IGF-I: lessons from animal models. *Pediatric Endocrinology Reviews*, *5*(Suppl. 2), 739–743.

Rogol, A. D., & Hayden, G. F. (2014). Etiologies and early diagnosis of short stature and growth failure in children and adolescents. *Journal of Pediatrics*, *164*, S1–S14.

Rosenbloom, A. L., & Connor, E. L. (2007). Hypopituitarism and other disorders of the growth hormone-insulin-like growth factor-I axis. In F. Lifshitz (Ed.), *Pediatric endocrinology* (pp. 65–99). New York: Informa Healthcare.

Richmond, E. J., & Rogol, A. D. (2008). Growth hormone deficiency in children. *Pituitary*, *11*, 115–120.

Ciccone, N. A., & Kaiser, U. ,B. (2009). The biology of gonadotroph regulation. *Current Opinion Endocrinology Diabetes and Obesity*, *16*, 321–327.

Toogood, A. A., & Stewart, P. M. (2008). Hypopituitarism: clinical features, diagnosis, and management. *Endocrinology Metabolism Clinical of North America*, *37*, 235–261.

Tomycz, N. D., & Horowitz, M. B. (2009). Inferior petrosal sinus sampling in the diagnosis of sellar neuropathology. *Neurosurgery Clinical of North America*, *20*, 361–367.

Vance, M. L. (2008). Pituitary adenoma: a clinician's perspective. *Endocrine Practice*, *14*, 757–763.

Winter, W. E., & Rosenbloom, A. L. (2010). Biochemistry of Growth. In D. Dietzen, M. J. Bennett, & E. C. C. Wong (Eds.), *Biochemical basis of pediatric disease* (4th ed.) (pp. 171–194). Washington, DC: AACC Press.

Winter, W. E., Schatz, D., & Bertholf, R. L. (2015). Thyroid Disorders. In Carl A. Burtis, & David E. Bruns (Eds.), *Tietz fundamentals of clinical chemistry and molecular diagnostics* (7th ed.) (pp. 806–823). Elsevier.

Winter, W. E., Hardt, N. S., & Harris, N. S. (2016). Disorders of the Anterior and Posterior Pituitary. In W. A. Clarke (Ed.), *Contemporary practice in clinical chemistry* (pp. 493–512). Washington, DC: AACC Press.

Winter, W. E., McCormack, A., & Bertholf, R. L. (2017). Pituitary Function and Pathophysiology. In N. Rafai (Ed.), *Tietz textbook of clinical chemistry and molecular diagnostics* (6th ed.) (pp. 1492–1529). St. Louis: Elsevier.

41

Adrenal Cortex

Roger L. Bertholf, Mark Cooper, William E. Winter

OBJECTIVES

1. Define the following:
   - Androgen
   - Angiotensin
   - Glucocorticoid
   - Incidentaloma
   - Mineralocorticoid
   - Renin
   - Steroid hormone
2. State how adrenocortical steroid hormones are synthesized from cholesterol.
3. Describe the anatomy and physiology of the adrenal cortex.
4. For each of the following hormones, state the function, regulation (including anterior pituitary or hypothalamic hormones involved, if any), site of synthesis, circulatory transport, metabolism, and urinary metabolite:
   - Aldosterone
   - Androstenedione
   - Cortisol
   - Dehydroepiandrosterone (DHEA)
5. Describe the following dynamic test of adrenal function, including the reasons for performing it, protocol followed, and interpretation of results:
   - Adrenocorticotropic hormone (ACTH) (cosyntropin) stimulation
   - Corticotropin-releasing hormone (CRH) stimulation
   - Dexamethasone suppression
   - Insulin-induced hypoglycemia stimulation
   - Metyrapone stimulation
   - Mineralocorticoid stimulation/suppression
6. Discuss the following adrenocortical disorders including the hormone(s) involved, symptoms, primary/secondary/tertiary causes, and screening and suppression/stimulation tests used for assessment:
   - Addison disease
   - Conn syndrome
   - Cushing syndrome/Cushing disease
   - Congenital adrenal hyperplasia (CAH)
7. Summarize the circadian variation of cortisol and aldosterone; list the preanalytical variables that must be considered when assessing aldosterone.
8. Diagram the renin-angiotensin-aldosterone axis.
9. List and describe the laboratory analyses used to assess adrenocortical function, including methodology, specimen type, collection, and storage requirements for the following analytes:
   - Aldosterone
   - Angiotensin
   - Cortisol (serum and urine)
   - 17-Hydroxyprogesterone
   - Renin
   - Urinary free cortisol (UFC)
10. Assess and solve case studies related to disorders of the adrenal cortex and their laboratory assessment.

KEY WORDS AND DEFINITIONS

Addison disease Deficiency of adrenocortical hormones secondary to disease of the adrenal glands, characterized by hypotension and a bronze-like hyperpigmentation of the skin. Also called *primary hypoadrenalism* to distinguish it from secondary hypoadrenalism (deficiency of pituitary adrenocorticotropic hormone).

Adrenal androgens A class of sex hormones that produce masculinization.

Adrenocorticotropic hormone (ACTH) A 39-amino-acid anterior pituitary hormone that acts primarily on the adrenal cortex, stimulating its growth and the secretion of corticosteroids. Also called *corticotropin*.

Aldosterone The major mineralocorticoid steroid hormone secreted by the adrenal cortex.

Androstenedione An androgenic steroid produced by the testis, adrenal cortex, and ovary.

Angiotensin II A small (eight-amino-acid) polypeptide hormone; among its functions, it stimulates release of aldosterone and other hormones, constricts blood vessels, and controls arterial pressure.

Angiotensin-converting enzyme (ACE) Enzyme that catalyzes the removal of two amino acids from angiotensin I, thus converting it to the active hormone angiotensin II.

Angiotensin-converting enzyme inhibitors Pharmaceuticals that are competitive inhibitors of the angiotensin-converting enzyme. They are used in the treatment of hypertension.

Angiotensinogen A serum globulin formed by the liver that is cleaved by renin to produce angiotensin I.

Conn syndrome A condition of primary aldosteronism arising from oversecretion of aldosterone by an adrenocortical adenoma.

Congenital adrenal hyperplasia (CAH) A group of inherited disorders in which deficiencies of enzymes that catalyze the biosynthesis of cortisol result in compensatory hypersecretion of corticotropin and subsequent adrenal hyperplasia as well as excessive androgen production.

Cortisol The major adrenal glucocorticoid synthesized in the zone fasciculata (and, to a lesser extent, the zona reticularis) of the adrenal cortex.

Cushing syndrome A condition characterized by an increased concentration of adrenal glucocorticoid hormone in the bloodstream and its effects on the body.

Dehydroepiandrosterone (DHEA) A weak androgenic steroid secreted by the adrenal cortex. It is the major androgen precursor in females.

Glucocorticoids Any of the group of C21 steroids produced by the adrenal cortex that regulate carbohydrate, fat, and protein metabolism.

Hyperaldosteronism A condition in which the adrenal gland secretes and releases increased quantities of aldosterone.

Hypovolemia A condition characterized by an abnormal decrease in the volume of circulating blood in the body.

Mineralocorticoids Any of the group of 21-carbon corticosteroids (principally aldosterone) that contribute to the regulation of (1) water, (2) acid-base, and (3) electrolyte balance in the body.

Renin A hydrolase enzyme that catalyzes cleavage of the leucine-leucine bond in angiotensinogen to generate angiotensin I.

Steroidogenic acute regulatory protein (StAR) A transport protein that functions to regulate cholesterol transfer within the mitochondria.

Waterhouse-Friderichsen syndrome Adrenal gland failure caused by bleeding into the adrenal gland; it is a fulminating complication of bacterial infections, notably meningococcemia, characterized by a sudden onset and short course, cyanosis with petechial hemorrhages of the skin and mucous membranes, fever, and hypotension, which can lead to shock and coma.

Zona fasciculata The thick middle layer of the adrenal cortex that contains large lipid-laden cells. It is the major source of glucocorticoids and, to a lesser extent, adrenal androgens.

Zona glomerulosa The thin outer layer of the adrenal cortex. It is the source of mineralocorticoids.

Zona reticularis The inner layer of the adrenal cortex. Its cells resemble those of the zona fasciculata except that they contain less lipid. The zona reticularis is the major source of adrenal androgens and produces glucocorticoids to a lesser degree.

The adrenal cortex produces steroid hormones that regulate salt and water balance (mineralocorticoids); metabolism of fat, proteins, and carbohydrates (glucocorticoids); and sexual development (androgens). Secretion of adrenocortical hormones is influenced by multiple physiologic signals originating in the kidney, liver, lungs, pituitary, hypothalamus, and elsewhere. Disease can result from overproduction as well as deficiency of adrenocortical hormones. Moreover, excess or deficiency of these hormones can be primary (i.e., resulting from adrenal gland dysfunction) or secondary (i.e., resulting from dysregulation of stimulating hormones). Mineralocorticoid excess causes hypertension and hypokalemia (called **Conn syndrome**), whereas deficiency causes salt wasting, sometimes in the form of **Addison disease** (primary adrenal insufficiency). Glucocorticoid excess causes **Cushing syndrome**, characterized by hypertension, abdominal obesity, weakness, and abnormal fat distribution. Glucocorticoid deficiency, as in Addison disease, causes weakness, hypotension, and hyperpigmentation of the skin among other symptoms. Because regulation of adrenocortical hormones is multifactorial, laboratory investigation of adrenal disorders often requires measuring the response to various stimulating factors, both chemical and physiologic. Analytic methods are available to measure most adrenocortical hormones.

ANATOMY

The adrenal (or suprarenal) glands are pyramid-shaped, 2 to 3 cm wide, 4 to 6 cm long, and about 1 cm thick. The adrenal cortex and the gonads share several metabolic pathways for synthesis of steroid hormones because both are derived from mesodermal anlagen. The postnatal adrenal cortex is composed of three layers: the glomerulosa (10% to 15% of the cortex), the fasciculata (up to 75% of the cortex), and the reticularis (5% to 10% of the cortex). At birth the adrenal gland is nearly equal in weight to that of an adult (8 to 12 g). The large size of the fetal adrenal gland (≈250 mg/100 g body weight) may explain its propensity to occasionally be traumatized during delivery. Between 3 and 18 months of age, the adrenal glands involute to approximately half their size at birth. Later in life, the adrenal glands are less susceptible to trauma and represent less than 50 mg/100 g of body weight.

TABLE 41.1 Anatomy and Products of the Adrenal Gland

Adrenal Layer	Major Product(s)	Action
Cortex		
Zona glomerulosa	Aldosterone	Mineralocorticoid
Zona fasciculata	Cortisol	Glucocorticoid
Zona reticularis	Dehydroepiandrosterone Androstenedione	Adrenal androgen
Medulla	Epinephrine	Catecholamine

Adult adrenal glands weigh 4 to 6 g each. Table 41.1 lists the location and action of hormones produced by the adrenal gland. Hormones produced by the adrenal cortex include mineralocorticoids, glucocorticoids, and adrenal androgens.

Mineralocorticoids (Aldosterone)

Mineralocorticoids bind to the mineralocorticoid receptor (MR) in the distal convoluted tubule (DCT) and collecting duct of the nephron, the colon, and the salivary glands to promote sodium reabsorption and potassium and hydrogen ion excretion. The principal mineralocorticoid is **aldosterone**, but other compounds with mineralocorticoid activity include 11-deoxycorticosterone (DOC), 18-hydroxycorticosterone, corticosterone, and **cortisol**. Although aldosterone binds to the MR in the cytoplasm, with subsequent transit of the complex to the nucleus, some free MR is present in the nucleus. The MR binds cortisol and DOC with affinity equal to that of aldosterone. However, the MR is protected from cortisol and DOC by 11 beta-hydroxysteroid dehydrogenase-2 (HSD11B2), which converts cortisol to cortisone. Cortisone does not bind to the MR. The actions of mineralocorticoids include sodium retention (causing water to be retained), and potassium and hydrogen ion excretion.

Glucocorticoids (Cortisol)

Glucocorticoids bind to the glucocorticoid receptor (GR), which is expressed in a variety of tissues including lymphocytes, hepatocytes, and bone. Because of the wide distribution of the GR, glucocorticoid effects are diverse, including changes in intermediary metabolism and immunoregulation. Glucocorticoids raise blood glucose concentrations by enhancing the synthesis of gluconeogenic enzymes, increase liver glycogen content by activation of glycogen synthase, and inhibit glycogen phosphorylase, producing insulin resistance in both muscle and adipose tissue, further raising blood glucose concentrations. Protein catabolism causes thinning of the skin and loss of strength in connective tissues. Excess glucocorticoids may result in bone loss from collagen catabolism and loss of osteoid, leading to fractures and compressed vertebrae.

Glucocorticoids promote the redistribution of adipose tissue centrally to the trunk, neck, and face. Insulin resistance raises very low-density lipoprotein (VLDL) and triglyceride concentrations and lowers high-density lipoprotein concentrations. Insulin resistance decreases the activity of hormone-sensitive lipase in adipose tissue, allowing triglyceride hydrolysis to free fatty acids. In the liver, those free fatty acids provide substrate for the synthesis of hepatic triglycerides and subsequent VLDL production and export. Glucocorticoids increase appetite, thereby increasing caloric intake and weight gain.

Glucocorticoids are powerful anti-inflammatory hormones. They suppress immune hypersensitivity by inhibiting the production of histamine by basophils and mast cells. Modest doses of glucocorticoids can improve mood, but at therapeutic doses they can cause psychosis.

Adrenal Androgens (Dehydroepiandrosterone and Androstenedione)

The **adrenal androgens dehydroepiandrosterone (DHEA)**, dehydroepiandrosterone sulfate (DHEA-S), and **androstenedione** are peripherally converted to testosterone, which binds to the androgen receptor. Between ages 7 and 8 years, urinary excretion of 17-ketosteroids (breakdown products of adrenal androgens) increases, signaling that puberty will begin within 3 to 5 years.

In males, adrenal androgens such as DHEA and androstenedione are mostly unimportant because testosterone, produced in the testes, is a much more potent androgen. In pubertal and adult women, however, the adrenal androgens produce axillary and pubic hair. DHEA is highly sulfated; circulating concentrations of DHEA-S exceed those of DHEA by at least 100-fold. DHEA-S decreases slightly in women following menopause.

PHYSIOLOGY AND REGULATION OF ADRENOCORTICAL HORMONES

Steroid hormones are not stored and therefore must be produced as needed. Steroids are lipophilic and pass through cell membranes to exit the hormone-producing cells, enter the circulation to be distributed throughout the body, and enter target cells by passing through the membrane into the cytoplasm, where they bind to receptors. Translocation of the cytosolic hormone-receptor complex into the nucleus results in hormonal activity. Evidence suggests that steroid hormones may affect more than DNA transcription alone. In the circulation, steroids exist as free and protein-bound species. (For a more detailed discussion of free vs. bound hormone, refer to the section titled Adrenocortical Hormones in the Circulation.)

Aldosterone

The production and secretion of aldosterone are regulated by the renin-angiotensin system (Fig. 41.1). The rate-limiting component in this system is **renin**, the release of which is regulated by the juxtaglomerular ("next to the glomerulus") apparatus. The juxtaglomerular apparatus is composed of the juxtaglomerular cells of the afferent arteriole, which immediately leads to the glomerulus, lacis cells, and macula densa. Juxtaglomerular cells are modified smooth muscle cells that synthesize and secrete renin in response to decreased renal

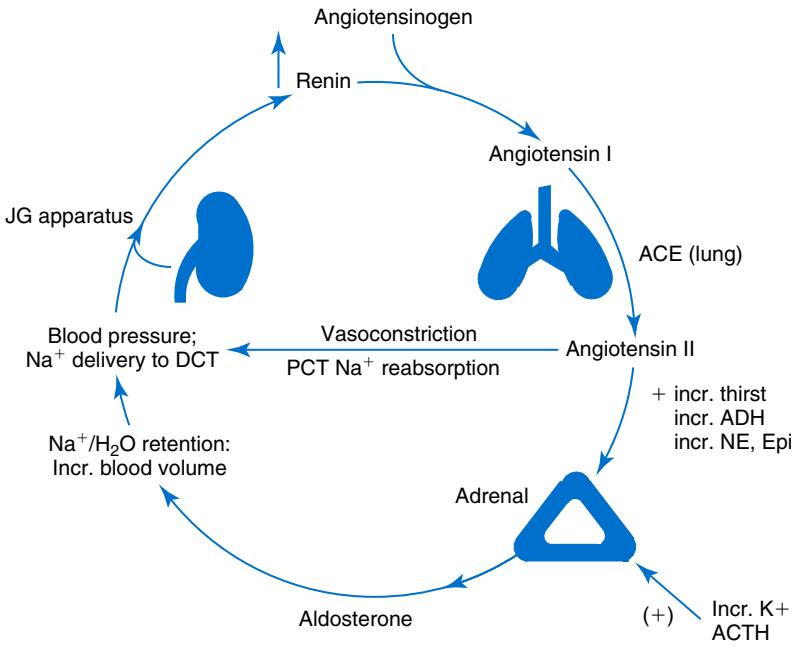


Fig. 41.1 The juxtaglomerular *(JG)* apparatus monitors hydrostatic pressure in the glomerulus and the sodium concentration in the distal convoluted tubule *(DCT)*. Renin is released in cases of decreased renal perfusion or decreased sodium concentration in the DCT. Renin cleaves angiotensinogen to angiotensin I. Angiotensin-converting enzyme *(ACE)*, predominantly in the lung, converts angiotensin I to angiotensin II. Angiotensin II has powerful vasoconstrictive effects, promotes sodium reabsorption by the proximal convoluted tubule *(PCT)*, and stimulates thirst, antidiuretic hormone *(ADH)* release, catecholamine release, and aldosterone synthesis and secretion. Aldosterone stimulates sodium and water reabsorption, increasing blood volume and blood pressure. Overall these effects enhance renal perfusion. Adrenocorticotropic hormone *(ACTH)* and increased potassium have minor effects in stimulating aldosterone synthesis and secretion. *Epi*, Epinephrine; *NE*, norepinephrine.

perfusion. The macula densa monitors the sodium concentration in the DCT. If a decline in sodium concentration is detected, the macula densa signals the juxtaglomerular cells via the hormone prostacyclin to release renin.

Sodium delivery to the DCT is decreased in hyponatremia and when renal perfusion declines, both of which stimulate renin release. The catecholamines norepinephrine and dopamine also stimulate renin release. Upright posture and catecholamine release thereby stimulate renin release. Potassium also directly stimulates renin release. Overall, renin release is physiologically stimulated by **hypovolemia**, reduced cardiac output, systemic vasodilation, selectively reduced renal perfusion, hyponatremia, and stress (mediated by catecholamines).

Angiotensinogen is an approximately 60-kDa alpha$_2$-globulin synthesized in the liver. Renin cleaves 10 N-terminal residues from angiotensinogen to form the decapeptide angiotensin I, a prohormone that has no innate activity. **Angiotensin-converting enzyme (ACE)**, expressed mostly in the lung, removes the two C-terminal residues from angiotensin I to generate the octapeptide **angiotensin II**. Further degradation of angiotensin II by aminopeptidase A yields the heptapeptide angiotensin III. The ratio of angiotensin II to angiotensin III is usually 4 to 1. An arginyl aminopeptidase generates angiotensin IV from angiotensin III. Angiotensin III and IV are active but overall have less potency than angiotensin II.

Angiotensin II preserves circulating blood volume and maintains blood pressure through several mechanisms: (1) stimulation of aldosterone synthase to produce aldosterone; (2) direct vasoconstriction; (3) increased release of epinephrine and norepinephrine, which also act as vasoconstrictors, from the adrenal medulla; (4) stimulation of sodium reabsorption in the proximal convoluted tubule (PCT); (5) stimulation of thirst; and (6) stimulated release of antidiuretic hormone (ADH). Aldosterone secretion occurs at approximately one-tenth the rate of cortisol secretion on a weight basis. The half-life of circulating aldosterone is less than 15 minutes.

Cortisol

Cortisol is regulated through a traditional hypothalamic-pituitary-end organ negative feedback system (Fig. 41.2). Corticotropin-releasing hormone (CRH) is released from the hypothalamus by stress, exercise, and hypoglycemia. Examples of physiological stress include pain, trauma, surgery, and hemorrhage. Examples of psychological stress include severe anxiety and major depression. Secreted CRH reaches the anterior pituitary gland via the hypothalamic pituitary portal system and stimulates the release of **adrenocorticotropic hormone** (**ACTH**, also called corticotropin). ACTH is also released in response to ADH, but to a lesser degree than CRH. The proinflammatory cytokines interleukin (IL)-1, IL-6, and tumor necrosis factor-alpha also stimulate ACTH release.

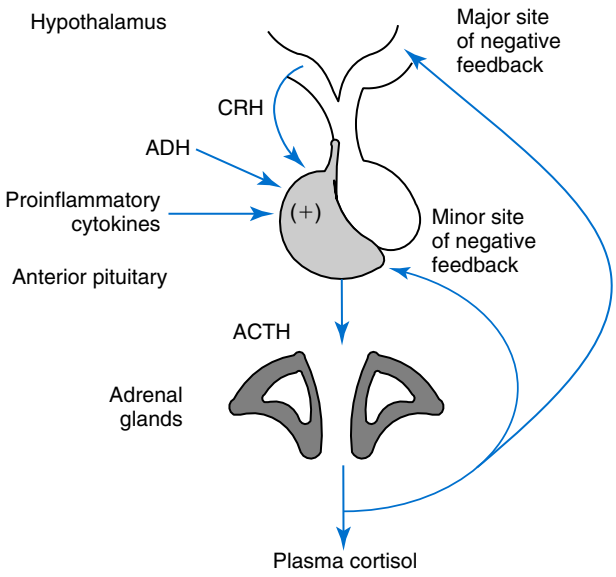


Fig. 41.2 Cortisol synthesis and release are controlled through a traditional hypothalamic–pituitary–end organ negative feedback loop. Corticotropin-releasing hormone *(CRH)* released from the hypothalamus is delivered to the anterior pituitary via the hypothalamic-pituitary portal system. CRH releases adrenocorticotropic hormone *(ACTH)* from anterior pituitary corticotrophs. Increased concentrations of antidiuretic hormone *(ADH)* and proinflammatory cytokines (e.g., interleukin [IL]-1, IL-6, and tumor necrosis factor-alpha) also stimulate ACTH release. ACTH stimulates the synthesis and release of cortisol from the adrenal cortex. Cortisol feeds back negatively at the pituitary, but the major site of negative feedback is the hypothalamus.

Hyperpigmentation of the skin occurs with ACTH excess, and this appears to be a direct consequence of the melanocyte stimulating hormone (MSH)-like activity of ACTH. Certain amino acid sequences in MSH are contained within ACTH. Pro-opiomelanocortin (POMC) is the precursor polypeptide of ACTH, the MSHs, and beta-endorphin.

Corticotropin circulates systemically and binds to ACTH receptors on cells in the adrenal cortex, resulting in cortisol synthesis and release. Cortisol inhibits hypothalamic and pituitary release of CRH and ACTH, respectively. Other negative feedback loops include ACTH suppression of hypothalamic CRH and an ultrashort feedback loop whereby ACTH suppresses its own release. There is diurnal variation in the secretion of cortisol; highest concentrations occur approximately 2 hours before awakening and lowest concentrations shortly after falling asleep.

Adrenal Androgens

Regulation of adrenal androgen synthesis and secretion is not well understood. The existence of a pituitary adrenal androgen-stimulating hormone or an adrenocortical androgen-stimulating hormone remains doubtful despite many years of research seeking its existence. The best-characterized regulator of androstenedione and DHEA secretion is ACTH. Diurnal rhythm in adrenal androgen concentrations parallels cortisol variations. Nevertheless, ACTH regulation of adrenal androgens does not explain the increase in adrenal androgen synthesis that occurs in both boys and girls during puberty; ACTH does not increase prior to puberty. Evidence indicates that changes in the expression and activity of factors that regulate enzymes involved in androgen synthesis are responsible for prepubertal and pubertal increases in adrenal androgens.

The synthetic pathways for aldosterone, cortisol, and adrenal androgens are illustrated in Fig. 41.3.

ADRENOCORTICAL HORMONES IN THE CIRCULATION

Steroid hormones are 90% to 98% bound to specific carrier proteins or albumin. Steroids that are sulfated or glucuronidated circulate unbound in the plasma. Aldosterone is carried primarily by albumin because cortisol, corticosterone, and 17-hydroxyprogesterone occupy most of the binding sites on corticosteroid-binding globulin (CBG; transcortin). In the blood, 80% to 90% of cortisol is bound to CBG, 7% is loosely bound to albumin, and 2% to 3% is unbound (free). Total cortisol concentration normally exceeds the aldosterone concentration by about 1000-fold, which explains why little aldosterone is carried on CBG. Because more cortisol than aldosterone is bound to CBG, the half-life of cortisol is longer (60 to 80 minutes) than the half-life of aldosterone (20 to 30 minutes). When total cortisol rises in Cushing syndrome, the increased fraction of free cortisol is excreted in the urine, increasing the urinary free cortisol (UFC) concentration. At normal concentrations in the blood, only 0.25% to 0.5% of total cortisol is excreted in the urine.

Transcortin (CBG) concentrations rise in response to increasing estrogen. In pregnancy, CBG concentration may rise two- to threefold. CBG concentrations are also increased in some patients with chronic active hepatitis. CBG concentrations are reduced in nephrosis (as a result of CBG loss in the urine), cirrhosis (because of decreased production), and hyperthyroidism (due to increased metabolism) as well as with glucocorticoid treatment (probably as a result of catabolism). Even with changes in total cortisol due to variations in CBG, the concentration of free (active) cortisol remains stable if the hypothalamic-pituitary-adrenal axis is functioning normally.

DHEA and its sulfated form, DHEA-S, and estradiol are mostly bound to albumin. In contrast, testosterone and dihydrotestosterone (DHT) are carried by sex hormone–binding globulin (SHBG). Estrogens and thyroid hormone increase SHBG production, whereas insulin, growth hormone, glucocorticoids, androgens, and progestins inhibit SHBG synthesis. SHBG concentrations are higher in children than in adults.

DISORDERS OF THE ADRENAL CORTEX

Thomas Addison first reported hypofunction of the adrenal cortex in 1855. During the past 65 years, many diseases associated with abnormal adrenal function have been described.

Diseases of the adrenal cortex result from either hypofunction or hyperfunction.

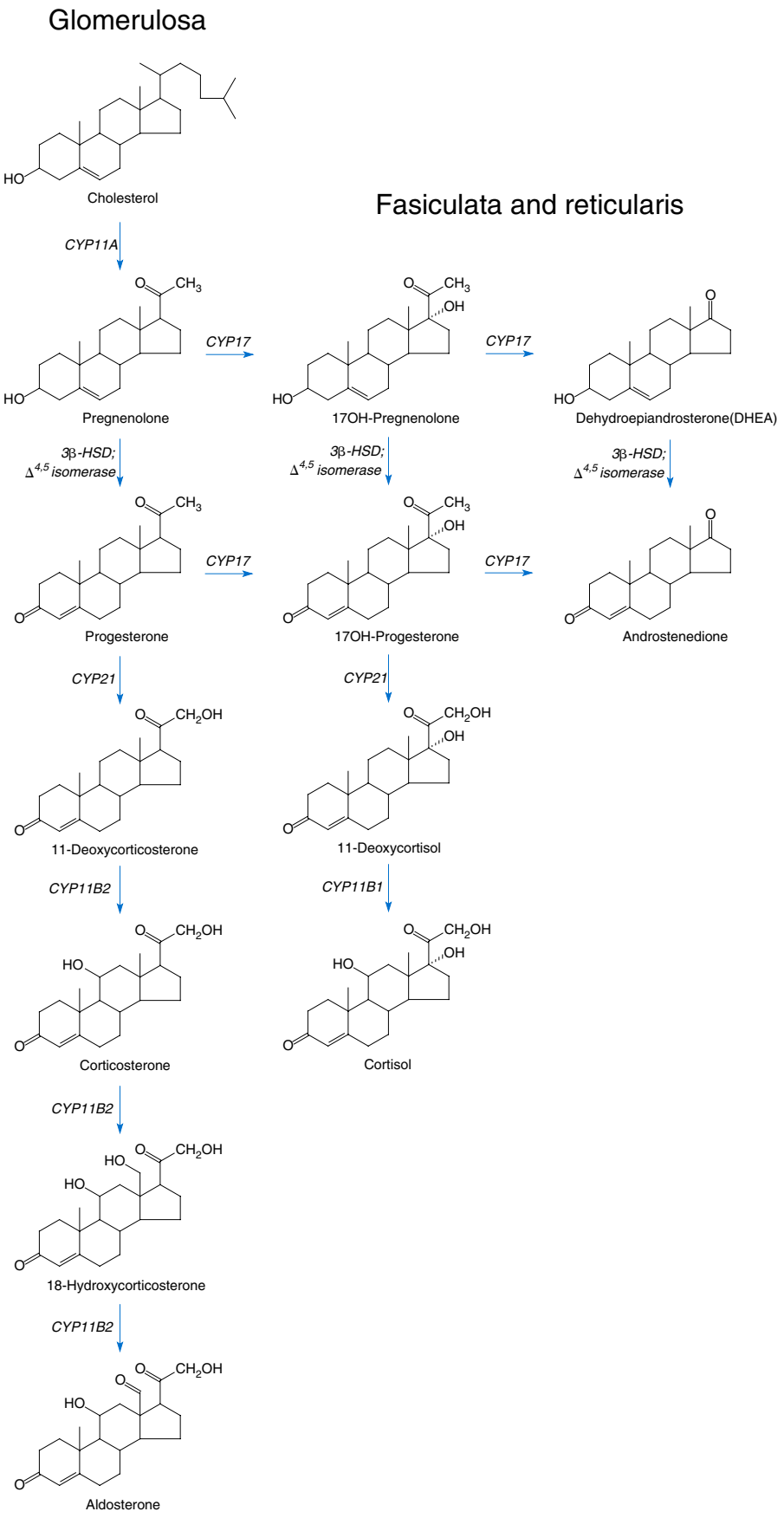


Fig. 41.3 The zona glomerulosa is the site of aldosterone synthesis. CYP11B2 is predominantly controlled by angiotensin II, regulating aldosterone synthesis and secretion. In the fasciculata and reticularis layers, cortisol and the adrenal androgens DHEA and androstenedione are produced. CYP11A, 3 beta-HSD, CYP17, CYP21, and CYP11B1 activities are controlled by ACTH.

Adrenal Insufficiency (Addison Disease)

Adrenal insufficiency causing combined mineralocorticoid and glucocorticoid deficiency is a rare disorder with a prevalence of only 4 to 11 cases per 100,000. Untreated adrenal insufficiency can be fatal. Cortisol deficiency is classified as primary, secondary, or tertiary. Primary adrenal insufficiency, also known as Addison disease, results from progressive destruction or dysfunction of the adrenal glands by a systemic disorder, an inborn error of metabolism (endogenous causes), or an exogenous cause such as infection. Worldwide, infectious disease is the most common cause of primary adrenal insufficiency. Human immunodeficiency virus (HIV) infection increases the risk of adrenalitis. Autoimmune adrenalitis accounts for more than 70% of cases reported in the Western world, with adrenal autoantibodies measurable in more than 75% of cases.

The most common inborn steroidogenic defects involve synthesis of cortisol with or without concurrent aldosterone deficiency, producing **congenital adrenal hyperplasia (CAH)**. Because such disorders commonly lead to adrenal androgen overproduction, they are discussed in a later section concerning hyperfunction of the adrenal cortex.

Because the entire adrenal cortex is affected in primary adrenal insufficiency, all classes of adrenal steroids are deficient. The onset of clinical manifestations is usually gradual, and the degree and severity of symptoms depend on the extent of adrenal failure. Complete glucocorticoid deficiency can manifest in a variety of ways, including fatigue, weakness, weight loss, gastrointestinal disturbance, and fasting hypoglycemia. Mineralocorticoid deficiency leads to dehydration, with hypotension, acidosis, hyponatremia, and hyperkalemia. Excessive pituitary release of ACTH unchecked by the negative feedback system may cause hyperpigmentation of the skin and mucous membranes. Basal plasma ACTH concentrations greater than 150 pg/mL (33 pmol/L), along with serum cortisol concentrations less than 5 μg/dL (138 nmol/L), are diagnostic of adrenal insufficiency.

The **ACTH stimulation test,** sometimes referred to as the *cosyntropin test,* assesses the functional capacity of the adrenal glands to synthesize cortisol. Cosyntropin (other names are Cortrosyn and Synacthen) is the N-terminal 24–amino acid sequence of ACTH, which includes the biologically active domain. Cosyntropin is a potent stimulator of cortisol secretion and has a very brief half-life and minimal antigenicity. The protocols for the 1-hour (also known as the "short Synacthen test") and multiple-day ACTH stimulation test (or "long or prolonged Synacthen test") are shown in Boxes 41.1 and 41.2, respectively.

A subnormal cortisol response in the cosyntropin stimulation test supports the diagnosis of primary adrenal insufficiency. A normal cortisol response to cosyntropin stimulation establishes that the adrenal cortex is capable of releasing cortisol and excludes primary adrenal failure.

Mineralocorticoid and glucocorticoid deficiency may occur in cases of primary adrenal insufficiency. In contrast, when ACTH deficiency (secondary or tertiary) is the cause of adrenal insufficiency, mineralocorticoid secretion is typically normal. In secondary and tertiary adrenal insufficiency, inadequate cortisol production may be due to destructive processes in the hypothalamus and/or pituitary that result in a decreased ability to secrete CRH (tertiary) or ACTH (secondary). However, the most common cause of tertiary insufficiency is long-term pharmacologic administration of glucocorticoids that suppress CRH, leading to a decrease in both ACTH release and cortisol secretion. Mineralocorticoid deficiency, hyperkalemia, and ACTH excess are not seen in secondary or tertiary adrenal insufficiency.

ACTH and cortisol concentrations are often not useful for establishing the diagnosis of secondary adrenal insufficiency. Episodic secretion and circadian variation result in ACTH and cortisol concentrations that overlap between normal subjects and individuals with secondary or tertiary adrenal insufficiency. Elevated basal plasma ACTH in a patient with an abnormal response to acute ACTH stimulation suggests primary adrenal failure, whereas low ACTH and normal or subnormal response to the short Synacthen test suggest a hypothalamic or pituitary disorder. Apart from the prolonged ACTH stimulation test (see Box 41.2), several dynamic tests can be used to distinguish primary from secondary and/or tertiary (central) causes of adrenal insufficiency.

In the **CRH test,** administration of ovine CRH stimulates ACTH secretion in normal subjects within 60 to 180 minutes. Individuals with tertiary disease produce elevations in ACTH following CRH administration, whereas individuals with secondary disease demonstrate only minimal changes in ACTH concentration.

BOX 41.1 Protocol for the 1-Hour (Rapid) Cosyntropin Stimulation Test

Rationale

Administration of adrenocorticotropic hormone (ACTH) to normal subjects causes a rapid rise in the serum cortisol concentration. Patients with adrenal destruction (e.g., Addison disease) show no change or an inadequate change in serum cortisol concentration after ACTH stimulation. Patients with atrophy of the adrenal cortex caused by exogenous glucocorticoid treatment or dysfunction of the pituitary gland or hypothalamus may show a slight rise in serum cortisol concentration but not one of normal magnitude.

Procedure

A morning (baseline) blood specimen is collected for determination of serum cortisol concentration; then 250 μg of cosyntropin is administered intramuscularly or intravenously. Blood specimens for serum cortisol determination are collected 30 and 60 min after injection.

Interpretation

A peak serum cortisol concentration of 18 to 20 μg/dL (500–550 nmol/L) or greater is a normal response. The expected change (delta) in cortisol concentration is 7 to 10 μg/dL (193–276 nmol/L). The peak cortisol value is more important than the incremental change. The incremental change may not be seen in patients who are tested at times of stress, when their adrenal output of cortisol is already maximally stimulated by endogenous ACTH. Also, it should be noted that clinical thresholds depend on the specific quantitative method used and may vary based on the reference material with which the analytical method is standardized.

BOX 41.2 Protocol for the Multiple-Day Adrenocorticotropic Hormone Stimulation Test

Rationale

Multiple-day adrenocorticotropic hormone (ACTH) stimulation testing for assessment of adrenal cortex function is required occasionally to evaluate the responsiveness of adrenal cortisol. A common situation is adrenal insufficiency treated with glucocorticosteroids before the cause has been established. Prolonged ACTH stimulation is used to distinguish primary from secondary and/or tertiary (central) causes of adrenal insufficiency.

Procedure

A total of 250 μg of cosyntropin is injected daily for 3 days. This is followed by an 8-h infusion of 250 μg of cosyntropin. Urinary free cortisol and serum cortisol are measured daily.

Interpretation

Serum cortisol values of 18 to 20 μg/dL (500–550 nmol/L) or greater exclude primary adrenal insufficiency. Glucocorticoid withdrawal would be required before secondary or tertiary adrenal insufficiency could be assessed. Little or no increase in cortisol secretion, even over successive days, is seen in primary adrenal failure. A progressive (staircase) rise in cortisol is seen over 2 to 3 days in adrenal insufficiency caused by pituitary or hypothalamic disease or corticosteroid suppression. Little or no response is seen in congenital adrenal hyperplasia caused by 21- and 17-hydroxylase deficiencies. Two- and 5-day variations of this test have been used in clinical practice. Also, it should be noted that clinical thresholds depend on the specific quantitative method used and may vary based on the reference material with which the analytic method is standardized.

BOX 41.3 Protocol for the Overnight Single-Dose Metyrapone Stimulation Test

Rationale

Metyrapone inhibits 11-beta-hydroxylase (CYP11B1), the enzyme that catalyzes the step immediately preceding cortisol synthesis. As the blood concentration of cortisol falls, the negative feedback effect is diminished, causing the release of adrenocorticotropic hormone (ACTH) from the pituitary gland. The stimulatory effect of ACTH on the adrenal cortex leads to a rise in 11-deoxycortisol, the compound immediately preceding cortisol in the biosynthetic pathway.

Procedure

Metyrapone (30 mg/kg body weight) is given orally at midnight with milk or a snack (to delay absorption). At 08:00 h the following morning, blood is collected for measurement of 11-deoxycortisol, cortisol, and ACTH concentrations.

Interpretation

In normal subjects, 11-deoxycortisol increases from less than 1 μg/dL (29 nmol/L) to greater than 7 μg/dL (200 nmol/L) after metyrapone stimulation, and ACTH values exceed 150 pg/mL (33 pmol/L). No response or an impaired response may be observed in pituitary or hypothalamic disease combined with inadequate enzyme blockade (plasma cortisol >3 μg/dL [>83 nmol/L]) or with Cushing syndrome caused by adrenal tumors or nonendocrine ACTH-secreting tumors. Exaggerated responses may be seen in pituitary Cushing syndrome.

Note: This test is not recommended in cases of suspected Addison disease for fear that suppression of cortisol production might precipitate an addisonian crisis. However, patients with Addison disease have an inadequate rise in 11-deoxycortisol in response to metyrapone.

In the **insulin-induced hypoglycemia stimulation test,** insulin is administered to stimulate CRH release by hypoglycemia; then plasma ACTH or cortisol concentrations are measured. This test involves a risk of hypoglycemia (obtundation, seizure, coma, and death) and should be performed only in highly monitored settings. The test should not be performed in patients with a seizure disorder or coronary artery disease. Venous access must be maintained for immediate administration of intravenous glucose if hypoglycemic complications occur.

The **metyrapone stimulation test** is an alternative although less commonly used indirect test of the function of the hypothalamic-pituitary-adrenal axis. It involves the administration of metyrapone, an inhibitor of the 11-beta-hydroxylase enzyme that converts 11-deoxycortisol to cortisol. Several protocols have been designed for metyrapone stimulation testing: a simple and relatively safe protocol for outpatient testing is described in Box 41.3.

The evolution of Addison disease passes through sequential stages: (1) normal to low aldosterone and elevated renin concentrations, (2) a deficient cortisol response to ACTH injection, (3) increased basal ACTH concentrations, and (4) deficient basal cortisol and aldosterone secretion. Autoantibodies against the adrenocortical cytoplasmic antigens or steroidogenic enzymes precede the clinical appearance of Addison disease; however, their predictive power is more evident in children than in adults. High-titer autoantibodies are also more predictive of progression to clinical disease. Approaches to the diagnosis of Addison disease are provided in the boxes titled "At a Glance: Investigation of Clinically Suspected Addison Disease" and "At a Glance: Investigation of Suspected ACTH Deficiency."

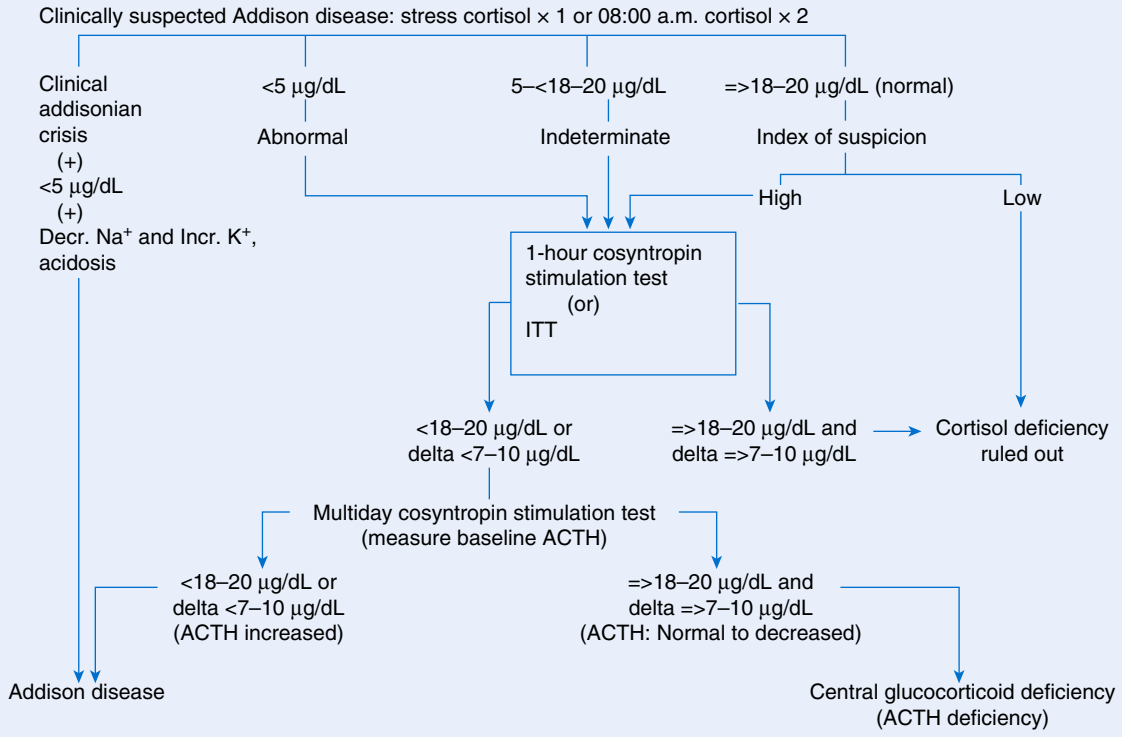
Hypoaldosteronism

Deficient aldosterone production occurs in conditions other than Addison disease (Table 41.2). Isolated aldosterone deficiency accompanied by normal cortisol production is seen in patients with (1) inadequate production of renin by the kidney, which leads to secondary aldosterone deficiency (hyporeninemic hypoaldosteronism); (2) inherited enzyme defects in aldosterone biosynthesis; and (3) acquired forms of primary aldosterone deficiency (heparin therapy and after surgery). The resulting metabolic changes are hyperkalemia and hyponatremia, often with mild hyperchloremic metabolic acidosis. Mild or moderate volume depletion, often with postural or unprovoked hypotension, may also occur. Hyporeninemic hypoaldosteronism can be established by demonstrating failure of both plasma renin and aldosterone to increase in response to furosemide stimulation or upright posture.

AT A GLANCE

Investigation of Clinically Suspected Addison Disease

In cases of clinically suspected Addison disease—at the time of an addisonian crisis with compatible hyponatremia, hyperkalemia, and acidosis—cortisol concentrations less than 5 μg/dL (<138 nmol/L) are diagnostic. Otherwise further testing guided by the cortisol concentration is required (either the "stress" cortisol concentration or the average of two cortisol levels at 08:00 h). Further testing can include the 1-hour cosyntropin stimulation test or an insulin tolerance test (ITT). If the stress or average morning cortisol is 18 to 20 μg/dL (500–550 nmol/L) or greater, and the index of suspicion for Addison disease is low, no further testing is required. However, if the index of suspicion for Addison disease is high despite a stress or average 08:00-hour cortisol of 18 to 20 μg/dL (500–550 nmol/L) or greater, further testing should be considered. As depicted, results of the 1-hour cosyntropin stimulation test or the ITT determine whether cortisol deficiency is excluded or, alternatively, a multiple-day cosyntropin stimulation test is indicated. Failure to demonstrate increased cortisol production in response to multiple-day cosyntropin stimulation supports the diagnosis of Addison disease (primary adrenal insufficiency). Not shown but still important: an elevated adrenocorticotropic hormone (ACTH) in the setting of proven cortisol deficiency also supports the diagnosis of Addison disease. If cortisol rises in response to prolonged cosyntropin stimulation, ACTH deficiency is evident (e.g., central ACTH deficiency). Note that cortisol assay results will exhibit bias based on calibration materials and other assay characteristics, so decision thresholds may vary and should be established by the laboratory.



Patients with primary adrenal insufficiency almost always suffer from aldosterone deficiency in addition to cortisol deficiency. Tests to confirm aldosterone deficiency in such patients are not usually necessary because hyperkalemia and hypotension support the diagnosis of primary adrenal insufficiency in patients with cortisol deficiency, elevated ACTH, and hyperpigmentation.

Glucocorticoid Excess (Cushing Syndrome)

Endogenous Cushing syndrome is the result of autonomous excessive production of cortisol, leading to the classic symptoms that are characteristic of this disorder. Exogenous Cushing syndrome is caused by excessive glucocorticoid therapy. The common clinical picture includes truncal obesity, moon face, a "buffalo" hump on the upper back below the neck, supraclavicular fat pads, purple striae, myopathy, hypertension, hirsutism, hypokalemic alkalosis, carbohydrate intolerance, disturbed reproductive function, and neuropsychiatric symptoms.

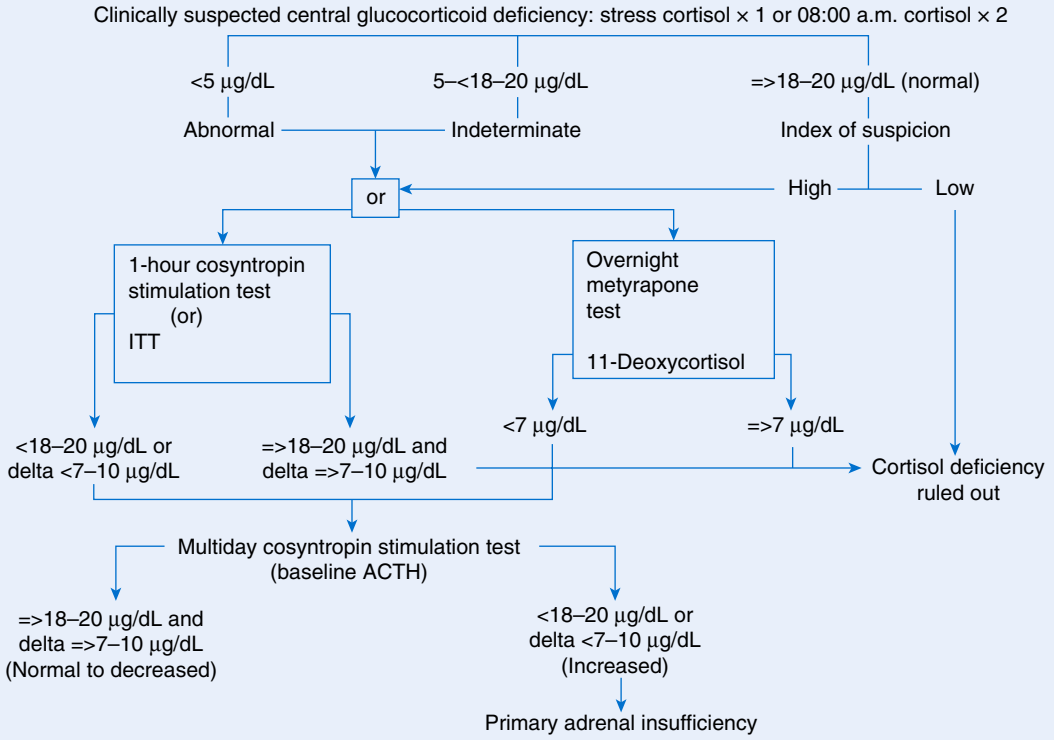
Endogenous disorders that cause hypersecretion of cortisol and Cushing syndrome may be classified as ACTH-dependent or ACTH-independent (Fig. 41.4).

ACTH-independent Cushing syndrome is caused by an adrenal adenoma or, less commonly by adrenal hyperplasia, leading to cortisol excess and suppressed pituitary secretion of ACTH. In bilateral micronodular hyperplasia (also called *primary nodular hyperplasia* or *adrenocortical dysplasia*), nodules 0.1 to 0.3 cm in diameter are observed that may be pigmented. The clinical outcome is similar to that of adrenal adenoma. Affected patients can display a spotty skin pigmentation pattern. In bilateral macronodular hyperplasia (also

AT A GLANCE

Investigation of Suspected Adrenocorticotropic Hormone Deficiency

When adrenocorticotropic hormone (ACTH) deficiency is suspected (e.g., central glucocorticoid deficiency), unless the 08:00-hour cortisol is 18 to 20 μg/dL (500–550 nmol/L) or greater and the index of suspicion for ACTH deficiency is low, the 1-hour cosyntropin stimulation test, an insulin tolerance test, or an overnight metyrapone test is performed. (Note: Metyrapone is not currently available in the United States.) Normal responses to such tests rule out cortisol deficiency. An abnormal response to such testing triggers testing via the multiple-day cosyntropin stimulation test. If cortisol rises in response to prolonged cosyntropin stimulation, ACTH deficiency is evident. Failure to respond to prolonged cosyntropin stimulation indicates primary adrenal failure, and concurrent mineralocorticoid deficiency should be investigated. Note that cortisol assay results will exhibit bias based on calibration materials and other assay characteristics, so decision thresholds may vary and should be established by the laboratory.



called *massive macronodular hyperplasia* or *macronodular adrenal dysplasia*), large nodules ranging from 0.2 to more than 4 cm in diameter are evident.

Cushing disease is the pituitary-dependent form of Cushing syndrome. In Cushing disease, hypersecretion of ACTH by a pituitary microadenoma is the primary defect that leads to bilateral adrenal hyperplasia and cortisol overproduction. In ectopic ACTH syndrome, nonendocrine tumors develop the ability to secrete ACTH or modified ACTH, resulting in bilateral adrenal hyperplasia, unregulated cortisol secretion, and suppression of pituitary ACTH release. Rarely, ectopic secretion of CRH occurs. In ectopic ACTH syndrome, the clinical presentation is usually dominated by the presence of cancer, and the patient may not display classic clinical findings of Cushing syndrome, such as centripetal obesity.

Screening Tests for Cushing Syndrome

Cushing syndrome is an uncommon disorder, but many of the usual signs and symptoms of this syndrome are seen in patients with normal adrenal function. The initial diagnosis of Cushing syndrome, particularly in mild or early disease, rests on laboratory evidence of excessive and autonomous cortisol production. Three simple screening tests are available for detecting Cushing syndrome—namely, 24-hour UFC, the 1-mg overnight dexamethasone suppression test, and salivary cortisol (see the following section titled Analytical Methods). The 2008 Endocrine Society guidelines recommend that two of the aforementioned three screening tests should be abnormal before proceeding to more definitive evaluation.

Under normal circumstances, less than 2% of secreted cortisol appears in urine as free cortisol. A **24-hour UFC** excretion of less than 50 μg/d (1380 nmol/d) excludes the diagnosis of Cushing syndrome. Diagnostic accuracy is greater than 90%.

The **overnight low-dose dexamethasone suppression test** is a reliable and convenient screening test for Cushing syndrome. Elevated cortisol normally inhibits ACTH

TABLE 41.2 Causes of Primary Adrenal Insufficiency or Failure of Aldosterone Production

Endogenous Causes	
Autoimmune disease	Sporadic
	Autoimmune polyglandular syndrome type 1 (Addison disease, candidiasis, hypoparathyroidism, and primary gonadal failure)
	Autoimmune polyglandular syndrome type 2 (Addison disease, primary hypothyroidism, primary hypogonadism, diabetes, and pernicious anemia)
Inborn errors	Congenital adrenal hyperplasia
	Congenital adrenal hypoplasia (*DAX-1* mutation)
	Demyelinating disorders: adrenoleukodystrophies
	Childhood X-linked recessive adrenoleukodystrophy (ALD; Brown-Schilder disease)
	Neonatal autosomal recessive ALD
	X-linked recessive adrenomyeloneuropathy (sudanophilic leukodystrophy)
	Familial (isolated) glucocorticoid deficiency (degeneration of fasciculata-reticularis layers)
	Wolman disease (lysosomal acid lipase deficiency)
	Steroidogenic factor-1 mutations
	Mitochondrial forms of Addison disease (Kearns-Sayre syndrome)
	Smith-Lemli-Opitz syndrome (sterol delta-7-reductase mutations)
	Adrenocorticotropic hormone resistance syndromes
Vascular disorders	Intra-adrenal hemorrhage (Waterhouse-Friderichsen syndrome): infection (caused by *Neisseria meningitidis, Pseudomonas aeruginosa, Haemophilus influenzae, Streptococcus pyogenes, Streptococcus pneumoniae*) or anticoagulants
Glandular infiltration	Neoplastic
	Leukemia, lymphoma, carcinoma of the lung, carcinoma of the breast
	Nonneoplastic
	Amyloid
	Hemochromatosis
Exogenous Causes	
Infection	Granulomatous disease
	Tuberculosis, sarcoidosis, histoplasmosis, cryptococcosis, blastomycosis (North and South American), sporotrichosis, and coccidiomycosis
	Other infections
	Cytomegalovirus, human immunodeficiency virus
Drugs	Blockers of steroid synthesis: mitotane, aminoglutethimide, trilostane, ketoconazole, metyrapone
	Glucocorticoid receptor blockers: RU-486
Abdominal Irradiation	
Bilateral adrenalectomy	Intra-adrenal thrombosis: renal vein thrombosis in neonates, heparin-induced thrombocytopenia, the antiphospholipid syndrome
Causes of Deficient Mineralocorticoid Activity	
Failure of Aldosterone Production (Decreased Aldosterone)	
Hyperreninemia hypoaldosteronism (synthetic deficiency)	Primary adrenal insufficiency
	Selective aldosterone deficiency
	Inborn errors (e.g., *CYP11B2* mutations)
	Drug-induced aldosterone suppression
	Heparin (direct inhibition of aldosterone secretion)
	Angiotensin-converting enzyme inhibitors
	Angiotensin receptor blockers
	Hypoaldosteronism in critical illness/hypotension (e.g., selective zona glomerulosa injury)
Hyporeninemia hypoaldosteronism (deficient aldosterone stimulation)	Renin deficiency (e.g., diabetes, renal failure)
Deficient Aldosterone Action (Resistance to Aldosterone; Elevated Aldosterone)	
Pseudohypoaldosteronism type 1	Renal
	Multiple-target-organ defects
	Early childhood hyperkalemia

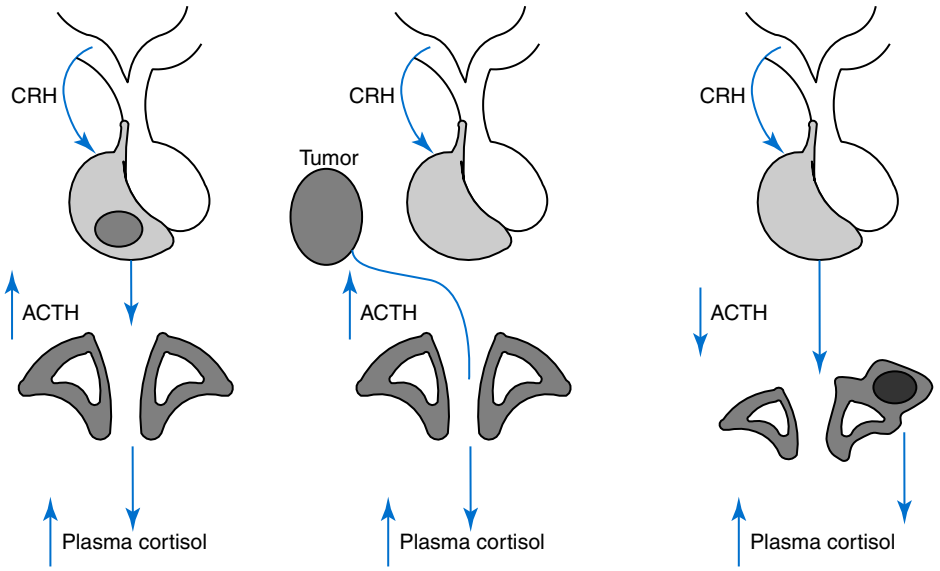


Fig. 41.4 Endogenous Cushing syndrome can be classified as Cushing disease *(left),* ectopic adrenocorticotropic hormone *(ACTH)* syndrome *(center),* or adrenal Cushing syndrome *(right).* ACTH is elevated in cases of Cushing disease and ectopic ACTH syndrome. ACTH is suppressed in cases of adrenal Cushing syndrome. *CRH,* Corticotropin-releasing hormone.

release from the pituitary gland, resulting in decreased production of cortisol and other ACTH-dependent steroids by the adrenal cortex. The integrity of this feedback mechanism is tested by administering a potent glucocorticoid such as dexamethasone and assessing ACTH secretion by measuring serum or urine cortisol concentrations. Dexamethasone does not significantly cross-react with cortisol immunoassays, so the secreted endogenous glucocorticoid, cortisol, can be distinguished from the exogenous glucocorticoid, dexamethasone. Several dexamethasone suppression tests are available for clinical use. Their protocols and interpretation are explained in Boxes 41.4 through 41.6.

Differential Diagnosis of Cushing Syndrome

Hypercortisolism supports the diagnosis of endogenous Cushing syndrome. **Plasma ACTH** concentrations are low in patients with an adrenal tumor (<10 pg/mL; <2.2 pmol/L) and are normal or moderately elevated in patients with Cushing disease (generally >50 pg/mL; >11 pmol/L). Plasma ACTH is often markedly elevated in patients with nonendocrine ACTH-secreting tumors (ectopic ACTH syndrome) because the tumor source of the ACTH does not respond to negative feedback from cortisol. On the other hand, the ACTH-secreting pituitary adenoma of Cushing disease has an elevated set point for negative feedback; if the cortisol (or other glucocorticoid) concentration is elevated high enough, some measure of negative feedback will occur (e.g., 60% of patients with Cushing disease achieve ACTH and UFC suppression when treated with high-dose dexamethasone). Plasma ACTH concentrations greater than 300 pg/mL (66 pmol/L) usually suggest a nonendocrine ACTH-secreting tumor (ectopic ACTH syndrome).

BOX 41.4 Protocol for the Overnight Low-Dose Dexamethasone Suppression Test

Rationale

In low doses, dexamethasone, a synthetic glucocorticoid and cortisol analog, suppresses adrenocorticotropic hormone (ACTH) and cortisol production in normal subjects but not in patients with Cushing syndrome.

Procedure

One milligram of dexamethasone is given orally between 22:00 and 24:00 h. Blood is collected for measurement of serum cortisol at 08:00 h the next morning.

Interpretation

In normal subjects, the serum cortisol concentration is suppressed to 2 μg/dL (55 nmol/L) or less after administration of 1 mg of dexamethasone. A postdexamethasone cortisol cutoff below 5 μg/dL (138 nmol/L) at 08:00 h is more sensitive for the detection of Cushing syndrome but less specific. Most patients with Cushing syndrome do not show adequate suppression, and cortisol concentrations at 08:00 h are usually 10 μg/dL (276 nmol/L) or greater. Serum cortisol above 2 μg/dL (55 nmol/L) may be seen in cases of stress, obesity, infection, acute or chronic illness, alcohol abuse, severe depression, oral contraceptive use, pregnancy, estrogen therapy, failure to take dexamethasone, or treatment with phenytoin or phenobarbital (which can enhance dexamethasone metabolism).

In a **CRH stimulation test,** plasma cortisol usually increases by 20% and plasma ACTH increases by 50% following CRH (1.0 μg/kg) administration. In cases of Cushing syndrome or ectopic ACTH secretion, glucocorticoid excess inhibits ACTH release in response to

BOX 41.5 Protocol for the Classic Low-Dose/High-Dose Dexamethasone Suppression Test

Rationale

Normal subjects show a decline in urinary free cortisol (UFC) concentrations in response to low-dose dexamethasone. However, cortisol in patients with excess cortisol production, regardless of the cause, is not suppressed by low-dose dexamethasone. Patients with Cushing syndrome caused by an adrenocorticotropic hormone (ACTH)–producing pituitary adenoma (e.g., Cushing disease from a corticotropinoma) may show suppression of cortisol with high-dose dexamethasone. Patients with Cushing syndrome from other causes—adrenocortical adenoma, carcinoma, or ectopic production of ACTH or corticotropin-releasing hormone (CRH)—usually do not demonstrate cortisol suppression with high-dose dexamethasone treatment.

Procedure

Twenty-four-hour urine samples are collected daily for 6 consecutive days. Free cortisol and creatinine are measured in each 24-h urine sample. UFC is regarded as a more specific measure of cortisol secretion than 17-hydroxycorticosteroids. The first 2 days provide baseline measurements of the excretion of cortisol. Dexamethasone 0.5 mg ("low-dose dexamethasone") is given orally every 6 h starting at 08:00 h on day 3 (for a total of eight doses: days 3 and 4). Dexamethasone 2.0 mg ("high-dose dexamethasone") is then given orally every 6 h starting at 08:00 h on day 5 (for a total of eight doses: days 5 and 6).

Interpretation

On the first 2 days, urinary excretion of cortisol (i.e., the baseline measurements) should be elevated in cases of Cushing syndrome. In cases of true endogenous Cushing syndrome, baseline UFC is usually at least two to three times higher than the upper limit of the reference interval. In normal individuals, baseline UFC excretion should be within the reference interval. Should a normal individual be stressed psychologically or physiologically and the baseline UFC is elevated, UFC on days 3 and 4 should be suppressed to less than 50% of the baseline level—that is, UFC suppression was achieved. All subjects with true endogenous hypercortisolism should not suppress on day 3 or 4. If suppression of UFC excretion occurs on days 5 and 6 (or on day 6), Cushing disease is diagnosed. However, not all cases of Cushing disease suppress on high-dose dexamethasone. Therefore failure to suppress on high-dose dexamethasone leaves three diagnostic possibilities: Cushing disease, adrenal adenoma (or, less commonly, adrenal carcinoma), and ectopic ACTH or CRH syndrome. Patients taking phenytoin or phenobarbital (or both) may metabolize dexamethasone more rapidly than normal subjects and may not show appropriate suppression.

BOX 41.6 Protocol for the High-Dose Overnight Dexamethasone Suppression Test

Rationale

The rationale for this test is similar to the rationale for the multiple high-dose dexamethasone suppression test.

Procedure

Four milligrams of dexamethasone are given orally at 23:00 and 24:00 h. Blood is collected for measurement of plasma cortisol at 07:00 or 08:00 h the next morning, or 8 to 9 h after dexamethasone is given.

Interpretation

Most patients with Cushing disease show at least 50% suppression of the baseline cortisol concentration. If suppression is not greater than 50%, the test should be repeated with 8 to 24 mg of dexamethasone.

exogenous CRH. In about 90% of patients with Cushing disease, however, the CRH stimulation test produces a normal or exaggerated ACTH or cortisol response or both.

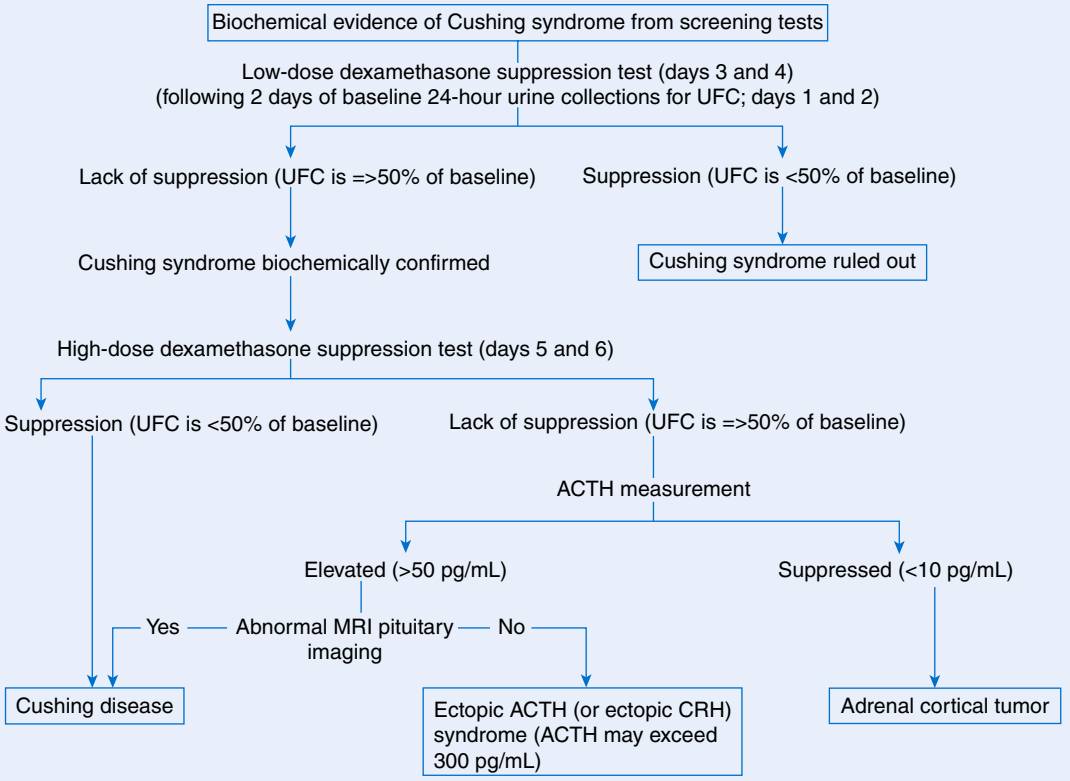
To differentiate pituitary-dependent Cushing disease from ectopic ACTH syndrome, a variation of the CRH stimulation test can be performed that measures ACTH in blood collected from the two inferior petrosal sinuses (IPSs). Blood specimens are collected from both right and left IPSs and from a peripheral vein (e.g., the inferior vena cava) at −30, 0, +2, +5, +10, and +30 minutes after intravenous administration of ovine CRH (1 μg/kg body weight) over 20 to 60 seconds. The ratio of the IPS concentration to the peripheral venous concentration of plasma ACTH is used to predict the location of excess ACTH secretion. An IPS-to–peripheral vein ratio greater than 2.0 is consistent with a pituitary lesion (Cushing disease), whereas a ratio less than 1.4 to 1.7 supports the diagnosis of an ectopic ACTH syndrome.

In patients with Cushing syndrome, plasma DHEA-S is usually normal or moderately elevated (plasma DHEA-S ~500 μg/dL; 13.5 μmol/L). Patients with an adrenal adenoma usually have low (age-adjusted) plasma DHEA-S levels. Plasma DHEA-S in patients with nonendocrine ACTH-secreting tumors is normal to elevated. In patients with CAH, adrenal androgens are usually suppressed after administration of 0.75 mg/d of dexamethasone for 2 to 3 weeks, but suppression does not occur in patients with adrenal tumors and nonendocrine ACTH-secreting tumors.

AT A GLANCE

Investigation of the Causes of Hypercortisolism

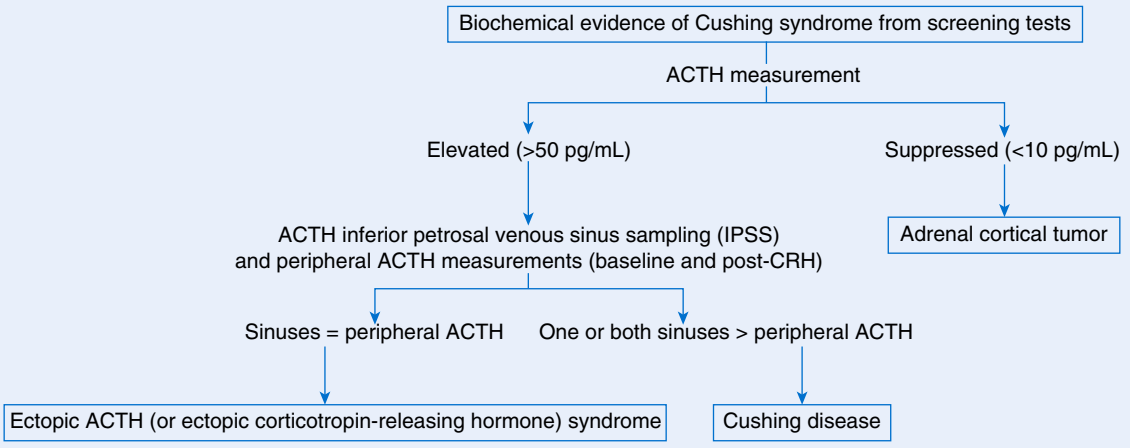
When hypercortisolism is established, low-dose/high-dose dexamethasone suppression testing (see Box 41.5 for protocol) can clarify the cause. Suppression of urinary free cortisol (UFC) on low-dose dexamethasone excludes cortisol excess. Suppression on high-dose dexamethasone establishes the diagnosis of Cushing disease, but not all patients with Cushing disease are suppressed on high-dose dexamethasone. Lack of cortisol suppression on high-dose dexamethasone triggers the measurement of ACTH. Elevated ACTH (>50 pg/mL or >10 pmol/L) favors a central or ectopic source, whereas a suppressed ACTH (<10 pg/mL or <2.2 pmol/L) favors an adrenal lesion.



AT A GLANCE

Differentiation of Endogenous Cushing Syndromes

A suppressed ACTH (<10 pg/mL or <2.2 pmol/L) establishes the diagnosis of adrenal Cushing syndrome (i.e., a cortisol-secreting adrenal cortical tumor). If ACTH is elevated (>50 pg/mL or >10 pmol/L), the pituitary is examined via inferior petrosal sinus sampling (and magnetic resonance imaging) to resolve Cushing disease versus ectopic ACTH syndrome.



Conditions That Mimic Cushing Syndrome

Alcohol abuse can induce a "pseudo–Cushing syndrome" that mimics the clinical and biochemical features of the disease. The abnormalities are reversible when alcohol consumption stops. HIV infection, anorexia nervosa, and depression are associated with elevated serum cortisol concentrations, and patients with these disorders may have abnormal low-dose, overnight dexamethasone suppression tests (morning cortisol ≥5 μg/dL; ≥138 nmol/L). Cortisol measurements at 24:00 hours may help to exclude Cushing syndrome in these circumstances. However, the clinical features of Cushing syndrome patients with HIV and anorexia nervosa are not typical. Measurement of UFC and plasma cortisol with the dexamethasone suppression test improves the predictive value in the diagnosis of both Cushing syndrome and depression. In patients with depression, it is best to repeat testing after the depression has been treated.

Obese patients may also display clinical features that mimic Cushing syndrome. Truncal obesity, striae, and enhanced excretion of 17-hydroxycorticosteroids (17-OHCS) are features of Cushing syndrome that occur in normal obese subjects. However, UFC is normal in obese individuals and can effectively differentiate normal subjects from those with true Cushing syndrome. Likewise, 48 hours of low-dose dexamethasone (0.5 mg orally every 6 hours) should suppress UFC excretion by more than 50% in obese patients.

Congenital Adrenal Hyperplasia

CAH is the most common cause of adrenocortical insufficiency in newborns. However, CAH is discussed under adrenal hormone overproduction syndromes because CAH commonly leads to overproduction of adrenal androgens. CAH presents a mixed picture of cortisol deficiency (hypofunction) and adrenal androgen overproduction (hyperfunction). Rarely, adrenal hemorrhage from trauma or sepsis causes adrenal failure in newborns, but CAH is the most common cause of adrenal failure in this age group.

CAH results from loss-of-function mutations in adrenocortical enzymes responsible for cortisol synthesis. With insufficient cortisol production, ACTH concentrations increase and stimulate fetal adrenal hyperplasia in utero. About 95% of CAH cases result from 21-alpha-hydroxylase deficiency, and 11-beta-hydroxylase deficiency accounts for most of the remaining cases. Other causes of CAH are very rare. The incidence of 21-alpha-hydroxylase CAH in western societies varies from 1 in 5000 to 15,000 live births. The incidence of 11-beta hydroxylase deficiency is approximately 1 in 100,000. Diagnosis of both conditions depends on finding elevated concentrations of 17-alpha-hydroxyprogesterone (17-OHP) in plasma or serum. Newborn screening for 21-alpha-hydroxylase and 11-beta-hydroxylase deficiency is standard practice in the United States and many other countries.

The metabolic block in cortisol biosynthesis leads to an accumulation of precursors, which are shunted into production of adrenal androgens. Measurements of the precursor steroids in blood help identify the specific enzyme defect and to monitor the response to cortisol replacement therapy. A partial block may cause marked or subtle clinical manifestations, whereas a complete enzyme block can be fatal without treatment. The closer the enzyme block to the final cortisol product, the less life-threatening the symptoms.

Because 21-alpha-hydroxylase and 11-beta-hydroxylase CAH produce elevations of adrenal androgens (DHEA and androstenedione) in utero, virilization of the external genitalia of a female fetus occurs, producing sexual ambiguity (ambiguous genitalia). Only rarely does CAH lead to inadequate male sexual development.

Cortisol deficiency in affected individuals leads to malaise, failure to thrive, hypoglycemia, and vascular instability. In approximately half of infants afflicted with 21-alpha-hydroxylase deficiency, insufficient aldosterone production can cause hyponatremia, hyperkalemia, acidosis, dehydration, and hypotension. In its most severe form, the "salt losing" 21-alpha-hydroxylase deficiency presents clinically at 10 to 14 days of age with addisonian crisis and a high mortality if left untreated. The non–salt-losing form of this disease is customarily referred to as *simple virilizing* 21-alpha-hydroxylase deficiency.

In males, adrenal androgen overproduction in 21-alpha-hydroxylase deficiency produces precocious pseudopuberty and some degree of hyperpigmentation of the male external genitalia during infancy or childhood; in the neonatal period they present with salt-losing crisis. Untreated, postnatal males will experience early-onset puberty with advanced secondary sex characteristics, and despite an early acceleration in their growth rate and tall stature, they may ultimately suffer from short stature because of premature closure of their epiphyses. The more profound the defect in cortisol biosynthesis, the greater the degree of precocious pseudopuberty in males.

Because the conversion of DOC to corticosterone in the **zona glomerulosa** is predominantly catalyzed by CYP11B2, mineralocorticoid levels are not affected in 11-beta-hydroxylase (CYP11B1) deficiency. Also, elevated ACTH may have some effect in stimulating DOC production, and DOC may be elevated in 11-beta-hydroxylase deficiency. Whereas females with 11-beta-hydroxylase deficiency have ambiguous genitalia and males undergo precocious pseudopuberty, salt loss does not occur. Thus electrolyte disturbances and acidosis do not occur, nor does addisonian crisis in the newborn period or later.

Once a child is diagnosed with CAH, fetal DNA testing in subsequent pregnancies allows prenatal diagnosis of CAH, which can prompt maternal treatment with high-dose dexamethasone to suppress excess fetal adrenal androgen production. This treatment prevents ambiguous genitalia in females and the androgenic effects of excess fetal adrenal androgens on the fetal brain.

A late-onset or attenuated (nonclassical) form of 21-alpha-hydroxylase deficiency can occur in female adolescents, presenting as hirsutism or virilization associated with menstrual irregularities, oligomenorrhea, amenorrhea, and a polycystic ovary–like syndrome. Late-onset 21-alpha-hydroxylase deficiency CAH is much milder than the classical prenatal virilizing form of 21-alpha-hydroxylase deficiency.

Early-morning 17-OHP or random androstenedione measurements are used to assess the response to glucocorticoid replacement in CAH patients. With adequate glucocorticoid replacement, 17-OHP or androstenedione should not exceed the upper limit of the reference interval for age and sex.

17-Alpha-hydroxylase deficiency is a rare cause of CAH. *CYP17* mutations influence sex steroid production of both androgens and estrogens in males in utero and in females at the time of puberty. Absence of 17-alpha-hydroxylase activity impedes cortisol biosynthesis and, as a result, ACTH concentrations increase. Elevated ACTH stimulates mineralocorticoid overproduction in the zona glomerulosa. Salt and water retention suppress aldosterone release. Salt-wasting is not present in 17-alpha-hydroxylase deficiency; in fact, excessive salt and water retention from excessive mineralocorticoid activity frequently produces hypertension in childhood. When sex steroids are not produced, genetic males will appear as sexually ambiguous or phenotypic females with delayed puberty. The uterus, fallopian tubes, and upper third of the vagina are absent because of anti-Müllerian hormone secretion by the Sertoli cells of the testes. Virilization of the external genitalia is reduced or absent in males in utero because of the block in testosterone biosynthesis. In genetic females, genitalia are not ambiguous; however, puberty and menses will not occur spontaneously because estrogens will not be produced.

The least common but most severe form of CAH is not an enzyme deficiency but a disorder involving **steroidogenic acute regulatory protein (StAR)**. StAR mediates the transport of cholesterol across the mitochondrial membrane. A deficiency of StAR is potentially lethal and causes congenital lipoid adrenal hyperplasia when the adrenal glands are saturated with cholesterol and cholesterol esters that are not further metabolized. In this disorder, deficiencies of all steroid hormones are present. Failure of androgen production in males results in ambiguous genitalia.

Functioning Adrenocortical Tumors

Plasma DHEA-S, DHEA, androstenedione, and testosterone concentrations are elevated in patients with virilizing adrenal adenomas and Cushing syndrome. The plasma concentrations of DHEA may also be elevated in women with virilizing ovarian tumors. Aldosterone-producing adenomas, referred to as *Conn syndrome,* are typically small microadenomas found in the zona glomerulosa that hypersecrete aldosterone, producing a syndrome characterized by low renin hypertension.

Adrenal carcinomas are rare, with an incidence of only 1 per million; they are usually larger than 4 cm and may cause only virilization, without the typical features of Cushing syndrome. Adrenal carcinomas are more common in women than in men by a 2.5:1 ratio. Most adrenal carcinomas are functional, producing glucocorticoids alone or glucocorticoids and androgens. Plasma DHEA-S, DHEA, and androstenedione concentrations are markedly elevated in patients with functional adrenal carcinomas. Peripheral conversion of adrenal androgens to testosterone results in hirsutism and virilization. DHEA-S concentrations can exceed 1000 μg/dL (27 μmol/L) in patients with adrenal carcinoma and usually are diagnostic for this malignancy. High-dose glucocorticoids do not suppress elevated androgen production.

Feminizing adrenocortical carcinomas are rare. These tumors result in elevated plasma DHEA-S, DHEA, androstenedione, estrone, and estradiol concentrations. Gynecomastia and sexual dysfunction occur in males; precocious pseudopuberty occurs in females. Dexamethasone does not suppress cortisol production in these cases.

Nonfunctioning Adrenocortical Tumors

Approximately 2% of the population have an adrenal tumor; most of these tumors are nonfunctioning and benign and are called *incidentalomas* (or *incidentomas*). The tumors are usually found when computed tomography (CT) or magnetic resonance imaging (MRI) scans of the abdomen are performed; these easily detect tumors 1 cm in diameter or 5 g in weight. No virilizing tumor smaller than 1 cm in diameter has been reported. Carcinomas usually weigh more than 30 g.

Mineralocorticoid Excess (Hyperaldosteronism)

Primary **hyperaldosteronism**, also known as *Conn syndrome,* is a syndrome caused by aldosterone excess. In primary hyperaldosteronism, excessive aldosterone production originates within the adrenal gland, whereas in secondary hyperaldosteronism, a stimulus outside the adrenal gland activates the renin-angiotensin system. (Note: The term *aldosteronism* is often used synonymously with *hyperaldosteronism.*) The interaction of renin, angiotensin, and aldosterone is important in the regulation of extracellular fluid volume and blood pressure, the balance of sodium and potassium ions, and acid-base balance.

Primary Hyperaldosteronism

Primary hyperaldosteronism is characterized by elevated plasma aldosterone and suppressed renin release, along with hypertension and hypokalemic alkalosis. Historically it was believed that no more than 1% of hypertensive patients had primary hyperaldosteronism. However, results from a 2007 study identified primary hyperaldosteronism in 10% or more of cases otherwise defined as "essential" hypertension.

Overproduction of aldosterone may be due to autonomous and inappropriate secretion of aldosterone by (1) an adenoma of one adrenal gland (aldosterone-producing adrenal adenoma [APA], or Conn syndrome), (2) hyperplasia of aldosterone-producing cells in both glands (bilateral idiopathic adrenal hyperplasia [IAH]), (3) an aldosterone-producing adrenal carcinoma, or (4) a rare familial condition known as *glucocorticoid-suppressible hyperaldosteronism* (familial hyperaldosteronism type 1). There are other rarer familial causes of hyperaldosteronism.

Clinical features of primary hyperaldosteronism are generally related to the consequences of aldosterone overproduction. Increased retention of sodium, expansion of extracellular fluid volume, and increased tubular secretion of potassium and hydrogen ions are the cardinal manifestations.

Hypokalemic alkalosis results from excessive renal excretion of potassium and hydrogen ions. Sodium retention causes a modest expansion of extracellular fluid volume along with an increase in arterial blood pressure.

Elevated aldosterone and low renin activity are key diagnostic features in primary hyperaldosteronism. Other factors influence the secretion of renin and aldosterone, however, and these factors must be considered before further laboratory evaluation. Drugs such as ACE inhibitors, beta blockers, and spironolactone alter renin release, and patients should be withdrawn from these medications for several weeks before the plasma aldosterone-to-renin ratio is determined. When patients with suspected primary hyperaldosteronism are evaluated, it is helpful to compare ambulatory plasma renin activity (PRA) to the patient's urinary sodium excretion or stimulate renin production with a potent diuretic such as furosemide.

Unusual conditions that suggest aldosterone excess but do not involve disorders of renin, angiotensin II, or aldosterone include apparent mineralocorticoid excess (AME), type 1 pseudohyperaldosteronism (Liddle syndrome), type 2 pseudohyperaldosteronism, and cortisol resistance.

Secondary events that disrupt normal renin release are much more common than primary abnormalities of renin secretion (Table 41.3). For example, a decrease in effective plasma volume or mean arterial pressure from sudden blood loss leads to release of renin from the juxtaglomerular cells of the kidneys; more angiotensin I and angiotensin II are formed, and production of aldosterone by the adrenal glands is increased. Angiotensin II promotes sodium reabsorption in the PCT, increases thirst, and increases release of ADH and catecholamines. Overall, angiotensin II promotes retention of sodium and water, an increase in extracellular volume, hydrogen ion wasting, and a decrease in the serum potassium due to enhanced renal potassium excretion. Secondary hyperaldosteronism is common in patients with congestive heart failure, nephrotic syndrome, cirrhosis of the liver, other hypoproteinemic states, or conditions that chronically deplete plasma volume.

Non–Mineralocorticoid-Deficient Causes of Urinary Sodium Loss

Bartter syndrome is a consequence of sodium loss from dysfunction of the ascending, thick loop of Henle or the DCT. Sodium loss evokes a compensatory hyperaldosteronism with increased renin and aldosterone, producing hypokalemia and alkalosis. Bartter syndrome in adults may be asymptomatic, but infants it can manifest as failure to thrive, salt craving, polyuria, and polydipsia. A variety of genetic mutations cause urinary sodium loss leading to secondary hyperaldosteronism with hypokalemia and alkalosis.

In pseudohypoaldosteronism (PHA; aldosterone resistance), clinical symptoms of hypoaldosteronism are concurrent with elevated concentrations of aldosterone. PHA may be classified as renal type 1, multiple target organ defects type 1, early childhood hyperkalemia type 1, Spitzer-Weinstein syndrome (adolescent hyperkalemic syndrome), subtype 3 RTA (renal tubular acidosis, IV type 2 PHA), and Gordon syndrome (mineralocorticoid-resistant hyperkalemia, chloride shunt syndrome type 2 PHA).

TABLE 41.3 Causes of Functional Hypermineralocorticoidism

Hyperaldosteronism	Primary (hyporeninemic) • Aldosterone-producing adenoma • Aldosterone-producing carcinoma • Idiopathic adrenal hyperplasia • Unilateral adrenocortical hyperplasia • Glucocorticoid remediable hypertension Secondary (hyperreninemic) • Renin-secreting tumors • Renovascular hypertension • Compensatory (nephrosis, cirrhosis, CHF) • Bartter syndrome • Gitelman syndrome
Deoxycorticosterone (DOC) excess	DOC-secreting adrenal adenoma 11 Beta-hydroxylase (CYP11B1) deficiency 17-Hydroxylase (CYP17) deficiency
Glucocorticoid excess	Cushing syndrome Cortisol resistance
Increased mineralocorticoid action without circulating mineralocorticoid excess	Apparent mineralocorticoid excess (HSD11B2 deficiency) Acquired HSD11B2 inhibition (glycyrrhizic acid induced) • Liddle syndrome (type 1 pseudohyperaldosteronism) • MR gain-of-function mutation (type 2 pseudohyperaldosteronism)

CHF, Congestive heart failure; *MR,* mineralocorticoid receptor.

LABORATORY EVALUATION OF ADRENOCORTICAL FUNCTION

Laboratory evaluation of adrenal cortical function centers on the measurement of cortisol, the primary corticosteroid; aldosterone, the primary mineralocorticoid; and the sex hormones, of which the adrenal cortex is a secondary source. Underproduction or overproduction of cortisol and aldosterone produces characteristic clinical syndromes, and genetic defects in steroid hormone biosynthesis produce abnormalities in sex hormone concentrations associated with clinical symptoms as well. Therefore measurement of steroid hormones produced in the adrenal cortex is essential to the clinical diagnosis of a variety of endocrine disorders.

Steroid hormones such as those produced in the adrenal cortex share considerable structural similarity based on their terpenoid sterane core, which is modified by functional groups at various locations on the tetracyclic backbone. In general, steroids are lipophilic compounds, but they require additional modification—esterification to fatty acids—to facilitate their transport across cellular membranes. Transport in the bloodstream typically involves carrier

proteins (e.g., transcortin, SHBG). These properties influence the choice of analytical methods for measuring steroid hormone concentrations; most procedures begin with enzymatic or chemical hydrolysis of steroid esters and displacement of the steroid from carrier proteins through addition of a protein-binding agent such as 8-anilino-1-naphthalene-sulfonic acid (ANS) or salicylate. Photometric and fluorometric measurements of steroids lack specificity unless extraction and purification procedures are used. Immunoassays are practical for the measurement of steroid hormones and are widely available on a variety of automated chemistry platforms, but cross-reactivity with structurally similar hormones often results in bias between various immunochemical methods. Endogenous interferences also occur—for example, near the time of birth, when placental steroid concentrations are very high. Steroids can be extracted into organic solvents and analyzed by chromatographic methods coupled with mass spectrometric detection, which provides sensitive and specific quantitative results for these compounds. In contrast to peptide hormones, most of which exhibit considerable heterogeneity as the result of posttranslational modifications, steroid hormones constitute a relatively homogenous group of compounds that are calibrated against gravimetrically validated synthetic standards.

Choice of Specimen

Steroid hormones have been measured in urine, blood, saliva, and hair. The choice of specimen depends on the application. For clinical assessment of adrenal function, blood (plasma) measurements are convenient, but conditions that cause the binding protein concentration to increase can produce clinically irrelevant elevations in the total cortisol concentration. Urinary concentrations are helpful in determining time-integrated hormone production and provide useful estimates of free hormone concentrations in blood if an appropriately timed urine specimen collection is available. Measurement of steroid hormones in saliva has been proposed as an alternative to urinary hormone assays to estimate free hormone concentrations in blood, the principal advantage being ease of collection. Hair analysis, while not widely used, may be useful for assessing adrenal hormone production over an extended time.

Urine

Urinary concentrations of a hormone or its metabolites provide an approximation of the amount secreted over the time the urine is collected, typically 24 hours. Thus urinary hormone or metabolite measurements are useful when the hormone has a very short biological half-life and/or its secretion is pulsatile. Timed urine collections suffer variability associated with factors such as completeness of collection and impaired renal function. Urinary metabolites may reflect only a fraction of the active steroid hormones; these metabolites, in turn, may be affected by metabolic disease as well as diet and medications. Despite these limitations, UFC and measurement of urinary free estradiol, estrone, and testosterone are most useful for estimating the rates at which these steroids are produced and are not affected by variations in the concentrations of proteins that bind hormones, such as CBG and albumin.

Blood

Circulating concentrations of steroids in the blood provide the most direct measure of endocrine activity mediated by these hormones. Immunoassays and chromatographic methods have been developed for measuring most of the clinically relevant steroids in blood; mass spectrometric methods for steroid profiling have recently been reviewed. Provocative testing (stimulation and suppression tests, discussed earlier) may require blood samples because of the rapid changes that occur in hormone activity after stimulus or suppression. Many steroid hormones are released in a pulsatile or rhythmic fashion, so their concentrations in isolated blood specimens should take these variables into account.

Saliva

Most steroids of clinical interest have been measured in saliva, and for some steroids—such as cortisol, estriol, testosterone, and progesterone—salivary concentrations appear to correlate with the free hormone concentration in plasma. For other steroids (e.g., 17α-hydroxyprogesterone, estradiol, aldosterone), the clinical usefulness of salivary measurements has not been established. Measurement of steroids in saliva is attractive because specimens are simple to collect; this consideration is particularly important in pediatric patients.

Hair

Hair grows at an average rate of 1 cm per month; therefore it can reveal historical blood concentrations of substances that are deposited in the forming hair matrix. Several reports describe methods for measuring steroid hormones in hair. These methods have shown potential as diagnostic tools but will require additional validation studies before they can be applied generally.

Basal Peptide and Steroid Hormone Concentrations

Concentrations of adrenal mineralocorticoids and elements of the renin-angiotensin system are routinely measured in body fluids by various immunoassay and mass spectrometric methods. In healthy subjects, consuming a low-sodium diet, maintaining an upright posture, and using diuretics all increase plasma aldosterone concentrations, whereas consuming a high-sodium diet and lying in the supine position decrease aldosterone secretion. Standardized procedures for obtaining blood and urine specimens are required for proper interpretation of test results. It is useful to interpret aldosterone concentrations together with urinary sodium excretion because salt intake can profoundly influence the aldosterone concentration. Normally the aldosterone concentration is inversely proportional to the urinary sodium concentration.

Because it has enzymatic activity, renin can be measured either by its activity or its mass, and both methods are available; a MALDI-TOF method for measuring renin was recently

described. PRA is measured by its stimulation of angiotensin I release, which is measured using immunoassay techniques.

Renin release is controlled by many physiologic factors. Consuming a low-sodium diet, maintaining an upright posture, and using diuretic medications all serve to increase renin release (causing a rise in aldosterone) and should be controlled or eliminated before testing of the renin-angiotensin-aldosterone axis is begun. Because plasma renin varies with sodium balance, it is helpful to interpret ambulatory renin measurements together with urinary sodium excretion. An inverse relationship exists (similar to aldosterone), allowing the identification of low-, normal-, and high-plasma renin groups from a nomogram.

Even when time of sampling, posture, drug intake, and diet are controlled, it is often difficult to identify disorders of mineralocorticoid secretion on the basis of basal hormone concentrations alone. A variety of dynamic tests have been developed to assess hypersecretion or hyposecretion of aldosterone and renin.

The diagnosis of adrenal insufficiency in critically ill patients is controversial. On one hand, if the diagnosis of cortisol and/or aldosterone deficiency is missed and replacement therapy is not initiated, the patient can die of an addisonian crisis, as might occur in **Waterhouse-Friderichsen syndrome** (bilateral adrenal hemorrhage). However, inappropriate administration of cortisol in a septic patient can produce immunosuppression that may be fatal. Some investigators believe that free cortisol measurements are diagnostically helpful because CBG concentrations frequently decline from liver disease or acute malnutrition, which could lead to the misdiagnosis of cortisol deficiency. In equivocal cases of adrenal insufficiency, some authors caution that a single low cortisol concentration should not be used to make the diagnosis of cortisol deficiency. Just as free testosterone can be estimated from total testosterone and SHBG concentrations, a free cortisol index has been described as the simple ratio of total cortisol to CBG.

Free Versus Bound Steroids

Many hormones bound to carrier proteins circulate in the blood. Hormone-binding proteins serve two functions: they solubilize lipophilic compounds such as steroid hormones to facilitate their transport in the primarily aqueous circulatory environment and buffer the concentration of free active hormone in tissue. For hormones that are highly protein-bound, the concentration of the binding protein is the principal determinant of the total concentration of hormone. Therefore total hormone concentrations vary significantly, whereas the free, active hormone concentration remains within physiologically appropriate limits. Measurement of free (unbound) hormone concentrations provides the best indicator of hormone activity, but analytic methods that isolate free hormone concentrations without disrupting the free/bound equilibrium are difficult to design. Antibodies that recognize steroid hormones may have binding affinities that compete with binding proteins and detect significant portions of the bound fraction. Preanalytical methods to separate bound and free fractions often disrupt the physiologic equilibrium between bound and free hormone. Urinary and salivary steroid hormone measurements have been widely used to estimate free hormone concentrations.

Analytical Methods

Sensitive immunoassays for measuring steroid hormones have been available for decades and include most, if not all, of the common labels (radioisotope, enzyme, fluorescent, chemiluminescent) and immunoassay designs (heterogeneous, homogeneous, competitive, noncompetitive). However, analytical methods involving liquid chromatography coupled with tandem mass spectrometry are becoming more common.

Measurement of Cortisol

Approximately 90% of circulating cortisol is bound to plasma proteins, primarily to CBG. Cortisol also binds weakly to albumin. The concentration of CBG in the plasma rises in hyperestrogenic states, including pregnancy and oral contraceptive use; therefore total cortisol in serum is increased in these conditions, although free cortisol remains normal. Analytic methods involving extraction or protein precipitation estimate the free and protein-bound cortisol in the circulation, but immunoassays for direct measurement of cortisol have mostly replaced extraction/chromatographic methods for routine cortisol determinations.

Chromatographic methods. Gas chromatography (GC), thin-layer chromatography (TLC), and high-performance liquid chromatography (HPLC) have been used to measure cortisol. GC and HPLC coupled to mass spectrophotometric detectors (GC/MS, LC/MS, and LC/MS-MS) provide the most specific methods for measuring corticosteroids. Cortisol methods based on capillary electrophoresis are also available. All of these methods demonstrate high specificity for cortisol but have relatively low throughput and require sample preparation steps before analysis. For example, most HPLC methods require preanalytical extraction of cortisol with solid-phase extraction columns or liquid-liquid extraction prior to separation and detection by reversed- or normal-phase HPLC and a fluorescence or ultraviolet absorption detector. A reference method for serum cortisol using isotope dilution GC/MS has been proposed.

Liquid chromatography–tandem mass spectrometry (LC/MS-MS) methods are becoming increasingly common. Liquid chromatography has the advantage of not requiring volatile derivatives, so specimen preparation can be vastly simplified. LC/MS-MS methods have been developed to measure cortisol, cortisone, and other corticosteroids. LC/MS-MS appears poised to become the method of choice for corticosteroid profiling owing to its greater analytical specificity.

Immunoassay. In routine laboratory practice, automated immunoassays are the most common methods used for measuring cortisol in serum and urine, and they are widely available on various automated immunoassay platforms. Most cortisol immunoassays in routine use are heterogeneous competitive-binding assays that do not require extraction of

steroids prior to analysis. Cortisol is displaced from CBG and other endogenous binding proteins by protein-binding agents such as ANS or salicylate, low pH, or heat treatment. The efficiency of cortisol displacement from proteins by agents such as ANS may be influenced by the concentration of binding proteins in the specimen. For example, the amount of ANS that is adequate for plasma obtained from healthy men and nonpregnant women may be insufficient to displace cortisol from CBG in pregnancy when CBG concentrations are elevated.

Most contemporary automated cortisol immunoassays use nonisotopic labels. Enzyme labels used in these immunoassays include horseradish peroxidase, alkaline phosphatase, or β-galactosidase, and enzyme activity has been measured using photometric, fluorescent, or chemiluminescent substrates. Nonenzymatic approaches typically involve fluorescent and chemiluminescent labels, and these are the most common methods currently in use. Both heterogeneous and homogeneous designs have been developed. Most current heterogeneous immunoassays for cortisol involve chemiluminescent labels and magnetic separation. Homogeneous immunoassays for cortisol include the enzyme-multiplied immunoassay technique and the cloned enzyme donor immunoassay. These assays do not require separation of bound and free fractions. Most immunoassay methods for cortisol are available on fully automated immunochemistry platforms.

Specimen collection and storage. Cortisol can be measured in serum, heparinized plasma, or ethylenediaminetetraacetic acid (EDTA) plasma, although some methods recommend against using EDTA plasma because of assay interference. In serum or plasma specimens, cortisol is stable for 7 days at room temperature or refrigerated and is stable for 3 months frozen at −20°C.

Comments. Blood cortisol concentrations parallel ACTH concentrations, with episodic and diurnal minima and maxima observed throughout the day. Serum cortisol concentrations are highest in the early morning hours and vary between 7 and 25 μg/dL (193 and 690 nmol/L) between 04:00 and 12:00 hours. Late afternoon concentrations are about half the morning concentrations and are frequently less than 5 μg/dL (138 nmol/L) between 22:00 and 02:00 hours. A midnight cortisol less than 5 μg/dL (138 nmol/L) is considered normal. Serum cortisol combined with plasma ACTH improves the diagnostic accuracy of basal values, although ACTH concentrations display even greater variation throughout the day than cortisol concentrations.

Increased cortisol is associated with stress, glucocorticoid therapy, pregnancy, depression, hypoglycemia, and hyperthyroidism. No significant difference in cortisol concentrations has been observed between men and women, and reference intervals for cortisol are not age-dependent. The half-life of cortisol in the circulation is approximately 60 minutes, so concentrations of this hormone in blood can change rapidly. In newborns, a transient rise in cortisol occurs immediately after delivery, but after 12 to 48 hours, cortisol declines to concentrations below umbilical cord blood concentrations; it then increases to a stable reference interval by about 1 week of age. Renal failure does not directly affect serum cortisol, but metabolites that are not cleared in the urine have the potential to cross-react with immunoassays, causing overestimation of blood cortisol concentrations. Extraction of cortisol into an organic solvent may eliminate interference from hydrophilic metabolites.

Measurement of Free Cortisol

Free or unbound cortisol is the biologically active form of the circulating hormone; its concentration is buffered by the large reservoir bound to its transport proteins: CBG and albumin. Various methods have been developed to measure the free fraction of cortisol in serum, including ultrafiltration, equilibrium dialysis, and gel filtration, but these assays are technically demanding and expensive, and they are not in general use. Algorithms have been suggested for estimating the free cortisol concentration on the basis of CBG and albumin concentrations. Measurement of urinary cortisol provides an estimate of the free hormone concentration and has been considered the gold standard screening test for Cushing disease, although salivary cortisol has been recommended as the first-line screening test. Approximately 2% of total cortisol is excreted into the urine, and urinary cortisol may be used as a screening test for cortisol hypersecretion.

Measurement of salivary cortisol (or oral fluid) is a practical and convenient way to assess the free hormone concentration. Salivary cortisol reflects the concentration of free cortisol in blood because CBG and albumin, the primary cortisol-binding proteins, do not appear in saliva. Most immunoassay kits for total serum cortisol can be used for the measurement of cortisol in saliva; extraction is not required because saliva contains virtually no cortisol-binding proteins or other cortisol metabolites. The glycoproteins in saliva can be precipitated by freezing and thawing, followed by centrifugation, producing a clear fluid that is free of protein interference. At this point, the challenge in measuring salivary cortisol involves achieving an adequate lower limit of detection because the concentration of cortisol in saliva is less than in serum or plasma. Typical salivary cortisol concentrations in healthy individuals are 0.14 to 1.0 μg/dL (4 to 28 nmol/L) at 07:00 hours and 0.07 to 0.22 μg/L (2 to 6 nmol/L) at 22:00 hours.

Measurement of Aldosterone

Measurement of aldosterone is technically challenging because the concentration of this hormone in blood is very low, nearly one-thousandth that of the cortisol concentration. Immunoassays for measuring aldosterone in blood and urine are available.

The cross-reactivity of aldosterone antibodies with other adrenal steroids (e.g., DOC, corticosterone) is relatively low (<0.01%). Nevertheless, the concentration of potentially cross-reacting steroids sometimes is very high, requiring some purification of aldosterone before measurement. Unconjugated plasma aldosterone will extract into an organic solvent and can be isolated by chromatography. The addition

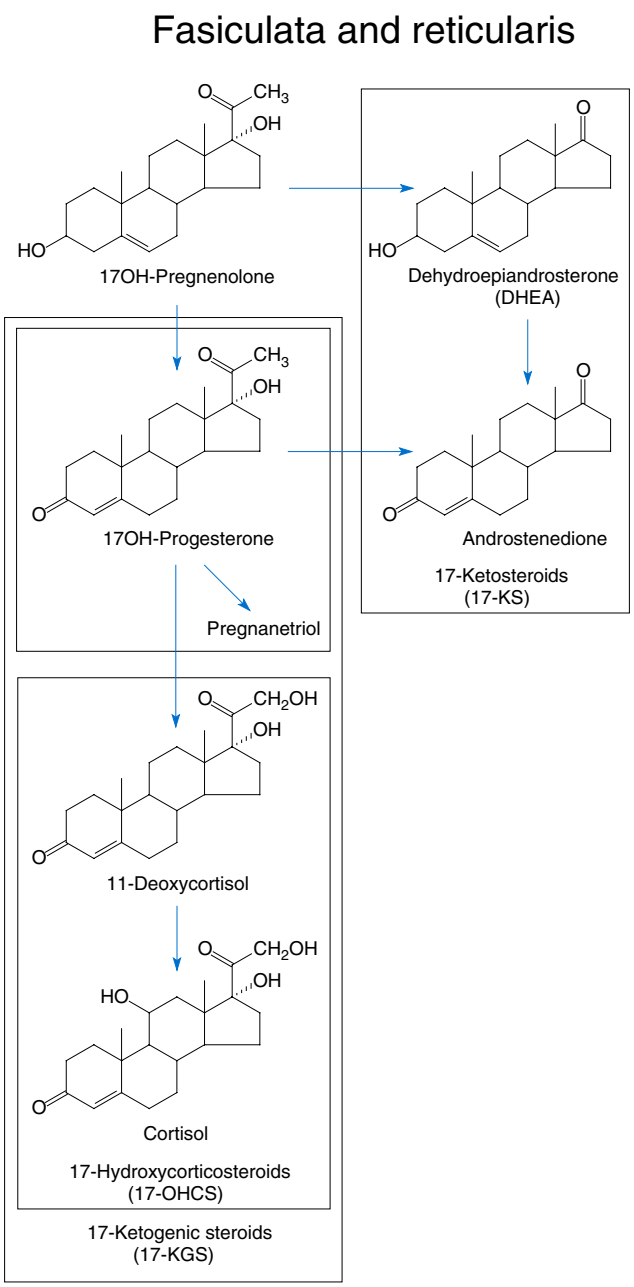


Fig. 41.5 Urinary metabolites generated from steroids involved in cortisol and adrenal androgen synthesis.

of tritiated aldosterone as an internal standard corrects for incomplete extraction in radioimmunoassay methods for aldosterone. In urine, unconjugated steroids are first extracted into an organic solvent such as ethyl acetate or methylene chloride after acid hydrolysis of the conjugates. The solvent is evaporated, and the dried extract is reconstituted in a buffer before analysis. Whether the specimen requires purification depends on the diagnostic kit being used and the type of patient being evaluated. For example, specificity is not a significant concern in hypertensive adults without adrenal disease, whereas greater specificity may be needed for newborns and young infants, patients with adrenal disease, and pregnant women, in whom high concentrations of potentially interfering steroids are likely.

Measurement of 17-Hydroxyprogesterone

Measurement of serum or plasma 17-OHP can be used to diagnose almost all cases of CYP21 deficiency, the most common cause of CAH. In this disorder, 17-OHP concentrations may reach several hundred times the upper limit of normal.

Radioimmunoassays for 17-OHP that use antibodies against 17-hydroxyprogesterone-3-carboxymethyloxime-BSA are available; these methods can be used with serum, plasma, saliva, and amniotic fluid. Monoclonal antibody–based methods have also been described. In addition to radioiodinated tracers, nonisotopic labels used in 17-OHP immunoassays include enzymes with photometric, fluorescent, or chemiluminescent substrates, and fluorescence-based immunoassays using fluorescein or streptavidin-europium labels, and chemiluminescent labels. Despite the use of highly specific antisera, most immunoassays are susceptible to interference by other corticosteroids that may be present in neonatal and infant plasma specimens. Methods for measuring 17-OHP in serum by GC/MS and LC/MS-MS are also available.

Measurement of 11-Deoxycortisol

Serum or plasma 11-deoxycortisol (compound S) measurements are used to detect 11β-hydroxylase (or C-11 hydroxylase) deficiency or as part of the metyrapone stimulation test. Metyrapone inhibits the 11β-hydroxylase enzyme, and a 40- to 80-fold increase in plasma 11-deoxycortisol is observed after metyrapone stimulation in patients with normal pituitary-adrenal reserve. As a consequence, analytical methods for 11-deoxycortisol in metyrapone stimulation tests do not require particularly high sensitivity.

Nonisotopic methods for measuring 11-deoxycortisol in serum have been described, including enzyme immunoassays, fluorometric methods, and fluorescence polarization.

LC/MS-MS methods for measuring 11-deoxycortisol have been described, mostly as part of corticosteroid profiles. These methods typically involve liquid-liquid extraction into an organic solvent, although solid-phase extraction has been used as well.

Measurement of Urinary Metabolites of Adrenocortical Steroids

Measurement of the urinary excretion of adrenocortical steroids and their metabolites is common in the laboratory assessment of adrenal disease. Immunoassays for the major circulating steroid hormones are widely available.

Measurement of 17-OHCS has been used to differentiate CYP21 deficiency from CYP11B1 deficiency. In CYP21 deficiency, 17-OHCS would not be elevated because both 11-deoxycortisol and cortisol are low. However, because CYP11B1 deficiency increases 11-deoxycortisol (the metabolic block is between 11-deoxycortisol and cortisol), 17-OHCS would be elevated in untreated or undertreated CYP11B1 deficiency. Also, CYP11B1 deficiency and CYP21 deficiency are differentiated by elevated plasma 11-deoxycortisol concentrations in the former.

17-Hydroxyprogesterone is metabolically reduced to pregnanetriol, which can be measured in urine. The urinary metabolites of 11-deoxycortisol and cortisol are classified as

POINTS TO REMEMBER

- The adrenal gland is the exclusive source of cortisol and aldosterone and is a minor source of sex hormones.
- Cortisol is synthesized and released by the adrenal gland in response to adrenocorticotropic hormone (ACTH) secreted by the anterior pituitary gland, which in turn is stimulated by corticotropin-releasing hormone (CRH) secreted by the hypothalamus.
- Aldosterone is synthesized and released by the adrenal gland in response to several factors, including angiotensin II and III, ACTH, and increased plasma potassium concentration.
- Overproduction of cortisol results in various symptoms collectively termed *Cushing syndrome.*
- Adrenal insufficiency can precipitate a life-threatening condition known as addisonian crisis, due to inadequate secretion of cortisol and aldosterone.
- Cushing syndrome is most often caused by corticosteroid therapy, but it may also be caused by cortisol-, ACTH-, or CRH-secreting tumors.
- Chronic adrenal insufficiency (Addison disease) can be caused by autoimmune, infectious, and metastatic diseases.
- Provocative tests are often required to determine whether adrenal disease is primary (dysfunction of the adrenal gland) or secondary (dysregulation of adrenocortical-stimulating hormones).

17-OHCS. Analytically, 17-OHCS can be photometrically measured by the reaction of 17,21-dihydroxy-20-oxosteroids with a phenylhydrazine-ethanol-sulfuric acid reagent, producing yellow phenylhydrazones termed *Porter-Silber chromogens.*

Urinary metabolites of 17-hydroxyprogesterone, 11-deoxycortisol, and cortisol are 17-ketogenic steroids (17-KGS). Ketogenic steroids have been measured using the Zimmermann reaction, in which an alkaline solution of meta-dinitrobenzene reacts with methylene groups at carbon-16 of the 17-ketosteroids. In CYP17 deficiency, 17-KGS is not elevated. However, in both CYP21 and CYP11B1 deficiencies, when untreated or undertreated, 17-KGS are elevated.

DHEA and androstenedione metabolites are measured in the same way as 17-ketosteroids because both have ketone groups in the C-17 position. Testosterone has a hydroxyl group at the C-17 position, and therefore is not a 17-ketosteroid. DHEA and androstenedione are elevated in untreated and undertreated CYP21 and CYP11B1 deficiencies. Fig. 41.5 summarizes these urinary steroid metabolites.

Measurement of Renin

Circulating concentrations of prorenin may be as many as 100-fold greater than the concentration of renin (although a 10:1 ratio of prorenin to renin is more common); therefore even minimal cross-reactivity of prorenin with renin antibodies used in immunoassays to measure renin can be problematic. Assays exist to measure PRA (by monitoring the production of angiotensin I) or the mass of renin by immunoassay. Each approach has advantages and disadvantages.

Measuring renin activity. Measuring plasma renin activity (PRA) provides an indication of the biologically active fraction of renin in the specimen because it measures the primary function of the enzyme, which is the conversion of angiotensinogen to angiotensin I. Renin activity measurements, however, are difficult to standardize, and two general approaches are used to measure renin activity. In the classic PRA method, first described in 1975, inhibitors of angiotensinase and ACE are added to prevent the conversion of angiotensin I to angiotensin II (some methods "trap" angiotensin I with an antibody to prevent its conversion to angiotensin II), the specimen is incubated at 37°C, and production of angiotensin I is measured. The rate of the reaction is influenced by pH, incubation time, and, most important, the endogenous angiotensinogen concentration in the specimen (which can be increased in pregnancy, glucocorticoid excess, and estrogen administration). Because the angiotensinogen concentration in blood does not ordinarily exceed the K_m for the renin-angiotensinogen complex, its concentration is rate-limiting. Therefore the classic PRA method produces results that vary significantly, depending on the endogenous concentration of angiotensinogen.

A second approach to measuring renin activity uses exogenous angiotensinogen as substrate and thereby avoids the variability associated with endogenous angiotensinogen concentrations. (This approach is sometimes called *plasma renin concentration assay*, which is a confusing term in that the assay still involves the measurement of activity, rather than concentration. Furthermore, the term is not consistently applied; *plasma renin concentration* has been used in reference to immunoassays that measure the renin concentration, rather than activity.) These renin activity methods use angiotensinogen derived from plasma collected from nephrectomized sheep; it is added at a concentration that is several times the K_m for the renin-angiotensinogen complex, ensuring that the reaction rate is limited by renin activity alone. The advantage of using a consistent source of angiotensinogen is that the activity assays can be calibrated against renin reference materials; an international reference preparation of human renin, validated by bioassay, has been available since 1975.

Care must be taken when measuring and interpreting renin activity in patients who are taking renin inhibitors, a relatively new class of drugs used to treat essential hypertension. Aliskiren was the first member of this class of drugs to be approved by the US Food and Drug Administration. Renin activity will be suppressed in these patients, although renin and prorenin concentrations will be high.

Measuring renin concentration. As an alternative to the PRA assay, the concentration of renin can be measured by immunoassay or mass spectrometry. Immunoassays for renin cross-react with prorenin. A variety of monoclonal immunoradiometric assays (IRMAs) for renin have been developed, some of which measure renin and prorenin.

Immunochemiluminometric (ICL) methods for renin are also available, and the results of these assays have been correlated with IRMA results.

Aldosterone/renin ratio. Primary aldosteronism is characterized by nonsuppressible aldosterone secretion. Normally plasma aldosterone concentration and renin activity are inversely related, but their short half-lives (<20 minutes) make their measurements highly variable. A more useful parameter is the ratio of plasma aldosterone concentration to renin activity (aldosterone-renin ratio; ARR). ARR is used as a screening test for primary aldosteronism. When the units of plasma aldosterone concentration are ng/dL and plasma renin activity is expressed in ng/mL·h, an ARR cutoff of 23.6 has a sensitivity of 97% and specificity of 94% for primary aldosteronism.

REVIEW QUESTIONS

1. Measurement of urinary free cortisol (UFC) is typically used as a screening test for
   a. hyperaldosteronism.
   b. cortisol hypersecretion.
   c. adrenal insufficiency.
   d. congenital adrenal hyperplasia (CAH).
2. Glucocorticoids _____blood glucose concentrations by altering synthesis of gluconeogenic enzymes.
   a. increase
   b. decrease
   c. do not affect
3. The primary mineralocorticoid is
   a. cortisol.
   b. dehydroepiandrosterone (DHEA).
   c. 11-deoxycortisol.
   d. aldosterone.
4. The major site of steroid hormone metabolism that involves the P450 enzyme systems is the
   a. liver.
   b. kidney.
   c. gastrointestinal tract.
   d. adipose tissue.
5. The stimulation test that is used to assess the ability of the adrenal glands to synthesize cortisol is referred to as the
   a. dexamethasone test.
   b. metapyrone test.
   c. cosyntropin test.
   d. salt-loading test.
6. *Primary* adrenal insufficiency is referred to as
   a. Cushing syndrome.
   b. Addison disease.
   c. Conn syndrome.
   d. congenital adrenal hyperplasia (CAH).
7. The adrenocortical disorder that results from loss-of-function mutations in specific adrenocortical enzymes responsible for the synthesis of cortisol is
   a. Cushing syndrome.
   b. Addison disease.
   c. Conn syndrome.
   d. congenital adrenal hyperplasia (CAH).
8. The anterior pituitary hormone that is responsible for increasing the size and number of adrenocortical cells and for eventual cortisol synthesis is
   a. corticotropin.
   b. growth hormone.
   c. corticotropin-releasing hormone (CRH).
   d. aldosterone.
9. One of the main functions of the mineralocorticoids is
   a. increasing blood glucose.
   b. increased fat breakdown.
   c. sodium retention.
   d. glycogen synthesis.
10. Which one of the following disorders is the result of autonomous excessive production of cortisol?
   a. Cushing syndrome
   b. Addison disease
   c. Conn syndrome
   d. Congenital adrenal hyperplasia (CAH)

SUGGESTED READINGS

Alkatib, A. A., Cosma, M., Elamin, M. B., Erickson, D., Swiglo, B. A., Erwin, P. J., et al. (2009). A systematic review and meta-analysis of randomized placebo-controlled trials of DHEA treatment effects on quality of life in women with adrenal insufficiency. *Journal of Clinical Endocrinology and Metabolism*, *94*(10), 3676–3681.

Anagnostis, P., Karagiannis, A., Tziomalos, K., Kakafika, A. I., Athyros, V. G., & Mikhailidis, D. P. (2009). Adrenal incidentaloma: a diagnostic challenge. *Hormones (Athens)*, *8*(3), 163–184.

Bangham, D. R., Robertson, I., Robertson, J. I., Robinson, C. J., & Tree, M. (1975). An international collaborative study of renin assay: establishment of the international reference preparation of human renin. *Clinical Science and Molecular Medicine. Supplement*, *2*, 135a–159s.

Betterle, C., Coco, G., & Zanchetta, R. (2005). Adrenal cortex autoantibodies in subjects with normal adrenal function. *Best Practice Research Clinical Endocrinology Metabolism*, *19*(1), 85–99.

Betterle, C., Scalici, C., Presotto, F., Pedini, B., Moro, L., Rigon, F., et al. (1988). The natural history of adrenal function in autoimmune patients with adrenal autoantibodies. *The Journal of Endocrinology*, *117*(3), 467–475.

Biddie, S. C., & Hager, G. L. (2009). Glucocorticoid receptor dynamics and gene regulation. *Stress*, *12*(3), 193–205.

Bloem, L. M., Storbeck, K. H., Swart, P., du Toit, T., Schloms, L., & Swart, A. C. (2015). Advances in the analytical methodologies: profiling steroids in familiar pathways-challenging dogmas. *The Journal of Steroid Biochemistry and Molecular Biology*, *153*, 80–92.

Bourne, R. S., Webber, S. J., & Hutchinson, S. P. (2003). Adrenal axis testing and corticosteroid replacement therapy in septic shock patients–local and national perspectives. *Anaesthesia*, *58*(6), 591–596.

Contreras, P., & Araya, V. (1995). Cushing's syndrome: review of a national caseload. *Revista médica de Chile*, *123*(3), 350–362.

De, B. K., & Haegeman, G. (2009). Minireview: latest perspectives on antiinflammatory actions of glucocorticoids. *Molecular Endocrinology*, *23*(3), 281–291.

Dorin, R. I., Pai, H. K., Ho, J. T., Lewis, J. G., Torpy, D. J., Urban, F. K., III., et al. (2009). Validation of a simple method of estimating plasma free cortisol: role of cortisol binding to albumin. *Clinical Biochemistry*, *42*(1-2), 64–71.

Havelock, J. C., Auchus, R. J., & Rainey, W. E. (2004). The rise in adrenal androgen biosynthesis: adrenarche. *Seminars Reproductive Medicine*, *22*(4), 337–347.

Jeanneret, F., Tonoli, D., Rossier, M. F., Saugy, M., Boccard, J., & Rudaz, S. (2016). Evaluation of steroidomics by liquid chromatography hyphenated to mass spectrometry as a powerful analytical strategy for measuring human steroid perturbations. *Journal of chromatography A*, *1430*, 97–112.

Manna, P. R., & Stocco, D. M. (2005). Regulation of the steroidogenic acute regulatory protein expression: functional and physiological consequences. Current Drug Targets. Immune. *Endocrine and Metabolic Disorders*, *5*(1), 93–108.

Munver, R., & Volfson, I. A. (2006). Adrenal insufficiency: diagnosis and management. *Current Urology Reports. 2006*, *7*(1), 80–85.

Nieman, L. K., Biller, B. M., Findling, J. W., Newell-Price, J., Savage, M. O., Stewart, P. M., et al. (2008). The diagnosis of Cushing's syndrome: an Endocrine Society Clinical Practice Guideline. *Journal Clinical Endocrinology Metabolism*, *93*(5), 1526–1540.

Pass, K. A., & Neto, E. C. (2009). Update: newborn screening for endocrinopathies. *Endocrinology Metabolism Clinical North America*, *38*(4), 827–837.

Popp, R., Malmstrom, D., Chambers, A. G., Lin, D., Camenzind, A. G., van der Gugten, J. G., et al. (2015). An automated assay for the clinical measurement of plasma renin activity by immuno-MALDI (iMALDI). *Biochimica et Biophysica Acta*, *1854*(6), 547–558.

Rai, R., Cohen, J., & Venkatesh, B. (2004). Assessment of adrenocortical function in the critically ill. *Critical Care and Resuscitation*, *6*(2), 123–129.

Thomson, S., Koren, G., Fraser, L. ,A., Rieder, M., Friedman, T. C., & Van Uum, S. H. (2010). Hair analysis provides a historical record of cortisol levels in Cushing's syndrome. *Experimental and Clinical Endocrinology & Diabetes*, *118*(2), 133–138.

42

Thyroid Disorders

Danielle B. Freedman, David Halsall, William J. Marshall, Christina Ellervik

OBJECTIVES

1. Define the following:
   - Colloid
   - Cretinism
   - Euthyroid
   - Follicle
   - Goiter
   - Iodide
   - Reverse T3 (rT3)
   - Thyroglobulin (Tg)
   - Thyroid storm
   - Thyrotoxicosis
   - Thyrotropin-releasing hormone (TRH)
2. Describe the structure and function of the thyroid gland, including cell types, internal location of key hormone precursors and proteins, regulation of the thyroid gland, and functions of the hormones synthesized and secreted.
3. Describe the metabolism of thyroid hormones, including synthesis, peripheral deiodination, specific effects on target tissues, and catabolism; state the full names of the thyroid hormones.
4. State the effects of increased and decreased thyroid hormones on thyroid-stimulating hormone (TSH), TRH, and target tissues.
5. For the following disorders, state the cause(s), symptoms, laboratory analyses, and lab results used to diagnose:
   - Autoimmune hypothyroidism
   - Graves disease
   - Hashimoto thyroiditis
   - Myxedema
   - Primary and secondary hyperthyroidism
   - Primary and secondary hypothyroidism
   - Sick euthyroid syndrome
   - T3 thyrotoxicosis
6. Name and describe the thyroid autoantibodies associated with thyroid disease, including mechanism of action, specificity of autoantibody to each disease, methods of detection, and interferences.
7. State the methods, specimen requirements, and problems with methods used to assess total and free thyroid hormones, thyroxine-binding globulin (TBG), Tg, and TSH.
8. State the formulae for estimation of free thyroid hormone concentration; given appropriate information, calculate estimates of free hormone concentration.
9. Analyze and solve case studies related to thyroid disease and laboratory analysis of these conditions.

KEY WORDS AND DEFINITIONS

Autoimmune thyroid disease (AITD) Diseases in which the immune system attacks or stimulates the body's own thyroid gland.

Central hypothyroidism Refers to thyroid hormone deficiency due to a disorder of the (1) pituitary, (2) hypothalamus, or (3) hypothalamic-pituitary portal circulation.

Colloid An amorphous material found in the follicular lumen of the thyroid gland. A critical component is *thyroglobulin* (Tg).

Congenital hypothyroidism A pathological condition resulting from severe thyroid insufficiency, which may lead to cretinism or myxedema.

Cretinism An archaic term for the clinical consequences of untreated congenital hypothyroidism caused by a deficiency of thyroid hormone during prenatal development and infancy; characterized in childhood by short stature, developmental delay, dystrophy of the bones, and a low basal metabolism. It is also called *congenital myxedema*, which is also an archaic term.

Euthyroid Having normal thyroid function.

Goiter An enlargement of the thyroid gland that causes a swelling in the front part of the neck.

Graves disease A disorder of the thyroid of autoimmune etiology that causes hyperthyroidism. Characterized by having at least two of the following conditions: hyperthyroidism, goiter, and exophthalmos. It is also known in Europe as *Basedow disease*.

Hashimoto thyroiditis An autoimmune disorder in which the thyroid gland is attacked by a cell-mediated autoimmune process. Also known as *Hashimoto disease* and *chronic lymphocytic thyroiditis*, it is marked by (1) goiter, (2) chronic inflammation of the thyroid (*thyroiditis*), and (3) often hypothyroidism.

Hyperthyroidism A condition caused by excessive production of iodinated thyroid hormones. Symptoms and signs include increased basal metabolic rate, enlargement of the thyroid gland, rapid heart rate, high systolic blood pressure, and a number of secondary symptoms.

Hypothyroidism A condition of deficient thyroid gland activity leading to lethargy, muscle weakness, and intolerance to cold.

Myxedema A severe form of hypothyroidism in which there is accumulation of mucopolysaccharides in the skin and other tissue, leading to a thickening of facial features and a doughy induration of the skin.

Organification A process in the thyroid gland whereby iodide is oxidized and incorporated into tyrosyl residues (tyrosine) of thyroglobulin (Tg). Organification is catalyzed by the enzyme thyroperoxidase (TPO).

Primary hypothyroidism A condition that develops when the thyroid gland fails to produce or secrete as much thyroxine (T4) as the body needs.

Reverse T3 (rT3) A biologically inert metabolite of thyroxine (T4) with three iodine molecules attached in a configuration (l-3,3′,5′-triiodothyronine) different from that of the active thyroid hormone triiodothyronine (T3).

Secondary (central) hypothyroidism Hypothyroidism that arises as a consequence of inadequate secretion of thyroid stimulating hormone (TSH or *thyrotropin*) by the anterior pituitary gland.

Sick euthyroid syndrome Abnormalities of T4, free T4, T3, and TSH concentrations that are seen in people with severe illness. In general, treatment with replacement thyroid hormone is not indicated.

Subclinical hyperthyroidism A biochemical condition with normal concentrations of serum thyroid hormones when the serum TSH concentration is repeatedly low in the absence of hypothalamic or pituitary disease.

T3 thyrotoxicosis A hyperthyroid condition in which T3, but not T4, is elevated.

Thyroglobulin (Tg) An iodine-containing glycoprotein of high molecular weight (663 kDa) present in the colloid of the follicles of the thyroid gland.

Thyroid follicle The secretory unit of the thyroid gland consisting of an outer layer of epithelial cells that enclose an amorphous material called colloid.

Thyroid-stimulating hormone (TSH) A polypeptide hormone synthesized by the anterior pituitary gland that promotes the growth of the thyroid gland and stimulates the synthesis and release of thyroid hormones by the thyroid gland. Also called *thyrotropin*.

Thyroid storm A life-threatening condition that develops in a minority of cases of untreated thyrotoxicosis (hyperthyroidism, or overactive thyroid).

Thyroiditis A condition characterized by inflammation of the thyroid gland.

Thyrotoxicosis A toxic condition resulting from excessive amounts of thyroid hormones in the body.

Thyrotropin-releasing hormone (TRH) A tripeptide produced in the hypothalamus that stimulates the synthesis and release of TSH from the anterior pituitary.

Thyroxine (T4) The major hormone synthesized and released by the thyroid gland that contains four iodine molecules (l-3,5,3′,5′-tetraiodothyronine).

Thyroxine-binding globulin (TBG) One of three major plasma thyroid hormone binding proteins (TBG, transthyretin [TTE], and albumin).

Toxic multinodular goiter A condition in which the thyroid gland contains multiple lumps (nodules) that are overactive and that produce excess thyroid hormones. Also known as *Parry disease* and *Plummer disease*.

Transthyretin (TTR) A protein found in serum and cerebrospinal fluid that binds to and transports thyroxine (T4). TTR complexes with retinol-binding protein (RBP) to prevent its loss through the glomerulus by filtration. It was once called *prealbumin,* because it travels faster than albumin on electrophoresis gels.

Triiodothyronine The biologically active form of thyroid hormone formed primarily outside of the thyroid gland by the peripheral deiodination of thyroxine (T4). It has three iodine molecules attached to its molecular structure (l-3,5,3′-triiodothyronine).

ANATOMY

The thyroid consists of two lobes connected by an isthmus and is like a butterfly in shape with the right lobe being slightly larger than the left. It is in the front of the neck just above the trachea. The gland synthesizes thyroid hormones and calcitonin through two distinct cell types: the epithelial or follicular cells and the parafollicular (or C) cells, respectively. The normal adult thyroid gland weighs 15 to 25 g, but in specific disease states, it can attain a weight of several hundred grams.

The thyroid gland is composed of follicles or acini (Fig. 42.1). In the center of the **follicle** is the lacuna, which contains **colloid** composed predominantly of **thyroglobulin (Tg)**. The parafollicular (or C) cells have been shown to produce the polypeptide hormone *calcitonin.*

Thyroid Morphogenesis and Dysmorphogenesis

Thyroid dysgenesis is a collective designation for thyroid agenesis (or athyreosis; i.e., the complete lack of thyroid tissue), hypoplasia, hemiagenesis (lacking one lobe), and thyroid ectopia (complete or coexisting with normal positioned tissue; i.e., the aberrant location of thyroid tissue along its embryological descent). The developmental defects have

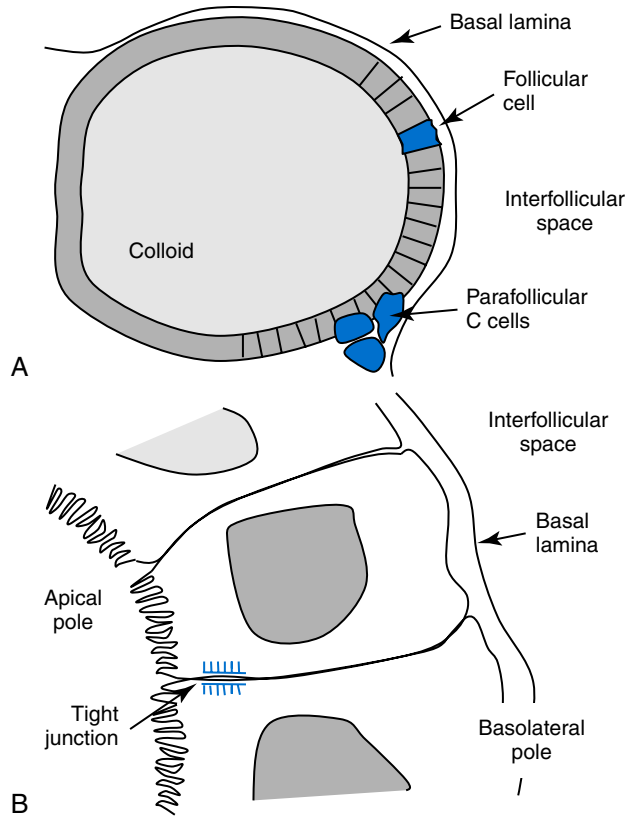


Fig. 42.1 (A) Basic unit of the thyroid gland. The follicle is the basic unit of the thyroid gland. It is composed of thyroid follicular cells surrounding the colloid. Outside the follicular cells is a basal lamina. (B) Apical and basolateral poles of the follicular cells and tight junctions between the follicular cells. Parafollicular (C cells) that secrete calcitonin can be found beneath or outside the basal lamina. Not pictured between the follicles are capillaries and fibroblasts.

TABLE 42.1 Nomenclature and Abbreviations for Thyroid Tests

Name	Abbreviation
Hormones	
Total thyroxine	T4
Total triiodothyronine (3,5,3′-triiodothyronine)	T3
Free thyroxine	FT4
Free triiodothyronine	FT3
Thyrotropin (thyroid-stimulating hormone)	TSH
Reverse T3 (3,3′,5′-triiodothyronine)	rT3
Serum-Binding Proteins	
Thyroid-hormone binding proteins	TBP
Thyroxine-binding globulin	TBG
Transthyretin (thyroxine-binding prealbumin)	TTR
Albumin	Alb
Tests for Autoimmune Thyroid Disease	
Autoimmune thyroid disease	AITD
Thyroglobulin autoantibodies	Anti-Tg
Thyroperoxidase autoantibodies	Anti-TPO
TSH receptor autoantibodies	Anti-TSHR
TSH receptor antibodies	TRAb
TSH receptor blocking antibodies	Anti-TSHRB
TSH receptor stimulating antibodies	Anti-TSHRS
Other Hormones, Thyroid-Related Proteins, and Conditions	
Thyroid peroxidase	TPO
Thyrotropin-releasing hormone	TRH
Thyroglobulin	Tg
Thyroid hormone receptor	THR
Deiodinase 1,2,3	D1, D2, D3
Iodotyrosine dehalogenase 1	DEHAL
Sodium/iodide symporter	NIS
TSH receptor	TSHR
Dual oxidase 1 and 2	DUOX1 and DUOX2
Thyroid hormone receptor-α	THRα
Thyroid hormone receptor-β	THRβ

considerable phenotypic variations. Ectopic tissue may coexist with a normally positioned gland. It is important to localize ectopic thyroid tissue before thyroidectomy.

Fetuses are dependent on thyroid hormones for normal growth and organ development, especially brain development and maturation.

THYROID HORMONES

The thyroid gland secretes two hormones, **thyroxine** (3,5,3′,5′-L-tetraiodothyronine) and **triiodothyronine** (3,5,3′-L-triiodothyronine), which are commonly known as T4 and T3, respectively (Table 42.1). In addition, the thyroid gland secretes very small amounts of biologically inactive 3,3′,5′-L-triiodothyronine (**reverse T3 [rT3]**) and minute quantities of monoiodotyrosine (MIT) and diiodotyrosine (DIT), which are precursors of T3 and T4. The structures of these compounds are shown in Fig. 42.2.

Biological Function

The functions of thyroid hormones include (1) control of the basal metabolic rate and calorigenesis, (2) enhancement of mitochondrial metabolism, (3) stimulation of neural development and normal growth, (4) promotion of sexual maturation, (5) stimulation of adrenergic activity with increased heart rate and myocardial contractility, (6) stimulation of protein synthesis and carbohydrate metabolism, (7) increasing the synthesis and degradation of cholesterol and triglycerides, (8) increasing the requirement for vitamins, (9) increasing calcium and phosphorus metabolism, and (10) enhancing the sensitivity of adrenergic receptors to catecholamines. These effects are typically magnified in patients with either an overactive thyroid gland (e.g., in **hyperthyroidism**), or reduced in patients with reduced thyroid function (e.g., **hypothyroidism**).

Biochemistry

Approximately 40% of secreted T4 is deiodinated in peripheral tissues by enzyme deiodinases to yield T3, and about 45% is deiodinated to yield rT3, a biologically inactive metabolite.

Therefore, with normal T4 production of 100 nmol (80 μg) daily, approximately 40 nmol (26 μg) of T3 and 45 nmol (29 μg) of rT3 are produced by peripheral deiodination.

From the estimated daily production rates for T3 (30 μg) and rT3 (30 μg) in a normal (**euthyroid**) state, at least 85% of T3 production and essentially all rT3 production are accounted for by peripheral deiodination of T4 rather than by direct secretion from the thyroid gland. T3 is at least four to five times more potent in biological systems than T4. Because one-third of all T4 is converted to T3 during the course of its metabolism, T4 may be considered to be a prohormone for T3. The biosynthesis of thyroid hormones occurs by a process termed **organification**. It involves the (1) trapping of circulating iodide by the thyroid gland, (2) incorporation of iodine into Tg tyrosines producing monoiodinated tyrosines and di-iodinated tyrosines, and (3) coupling of two iodinated tyrosyl residues to form the thyronines (T4 and T3) within the protein backbone of the Tg protein in the follicular lumen (Fig. 42.3). Endocytosis followed by proteolytic cleavage of Tg releases the iodothyronines into the circulation.

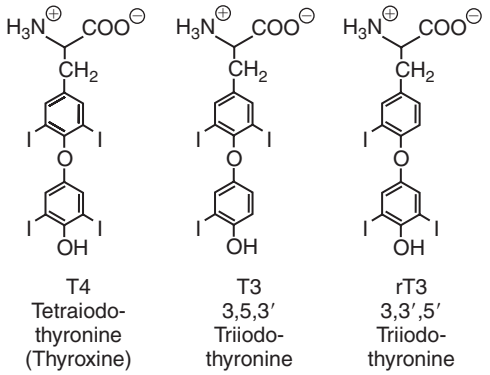


Fig. 42.2 Chemical structures of thyroxine *(T4)*, triiodothyronine *(T3)*, and **reverse T3** ***(rT3)***.

Dietary iodine is the basic element involved in the synthesis of thyroid hormones. It is normally ingested in the form of iodide. Iodide transport to the follicles is the first and rate limiting step in the synthetic process. The follicular cells of the thyroid concentrate iodide to some 30 to 40 times the normal plasma concentration by means of an energy-dependent pump mechanism, the sodium-iodide symporter (approved gene name: solute carrier family 5 [sodium/iodide cotransporter], member 5; approved symbol *SLC5A5*; chromosome 19p13.11).

The synthesis of MIT and DIT (Fig. 42.4), T4 (Fig. 42.5), and T3 (Fig. 42.6), occurs mainly at the follicular cell–colloid interface but also within the colloid (see Fig. 42.3).

Tg is present in highest concentrations within the colloid, where it is stored. The follicular cells engulf colloid globules by endocytosis; these globules then merge with lysosomes in the follicular cell. Lysosomal proteases break the peptide bonds between iodinated residues of Tg, and MIT, DIT, T4,

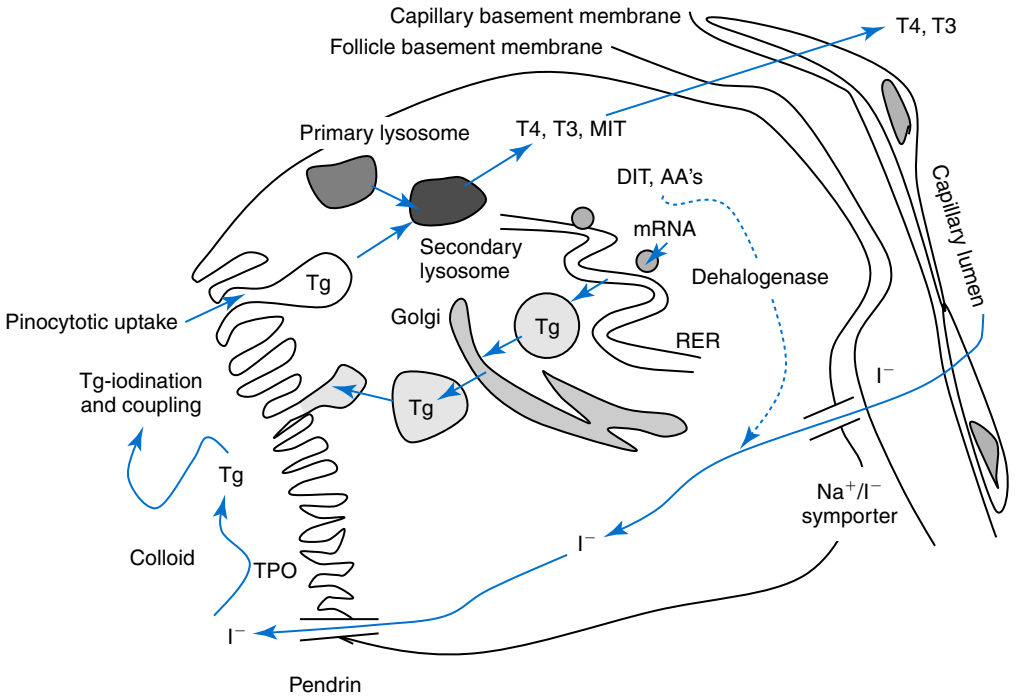


Fig. 42.3 Synthesis of thyroid hormones begins with absorption of iodide by the thyroid follicular cell and the Na^+/I^- symporter. From the cytoplasm, iodide moves into the lacunae via pendrin. Within the lacunae, thyroperoxidase *(TPO)* and the dual oxidases (DUOX [not depicted]) convert iodide to iodine, leading to iodination of tyrosine residues on thyroglobulin *(Tg)*. Tg is synthesized in the cell and exported to the lacunae. TPO is responsible for the coupling of monoiodotyrosine *(MIT)* and di-iodotyrosine *(DIT)* to form T3, and di-iodotyrosine and di-iodotyrosine to form T4. Upon the uptake of iodinated Tg (containing T4 and T3) and the fusion of this phagosome-like vesicle with a primary lysosome, Tg is degraded in a secondary lysosome, releasing T4 and T3 into the circulation and MIT and DIT undergo deiodination via a dehalogenase to recycle the iodine for new thyroid hormone synthesis. *RER,* Rough endoplasmic reticulum.

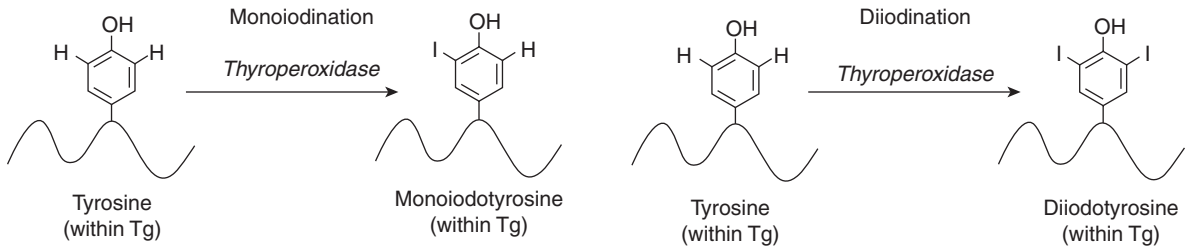


Fig. 42.4 Monoiodination and diiodination of tyrosine. *Tg,* Thyroglobulin.

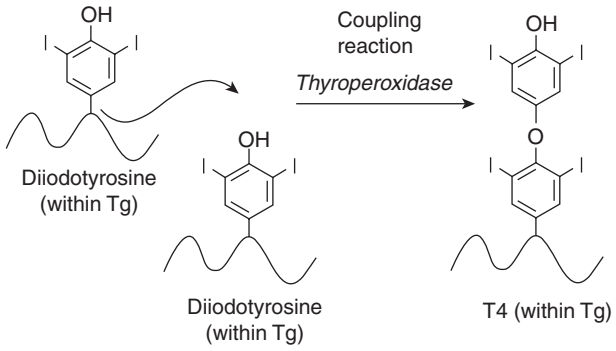


Fig. 42.5 Chemical coupling of two molecules of diiodotyrosines to produce a molecule of thyroxine (T4). The reaction is catalyzed by thyroperoxidase. *Tg,* Thyroglobulin.

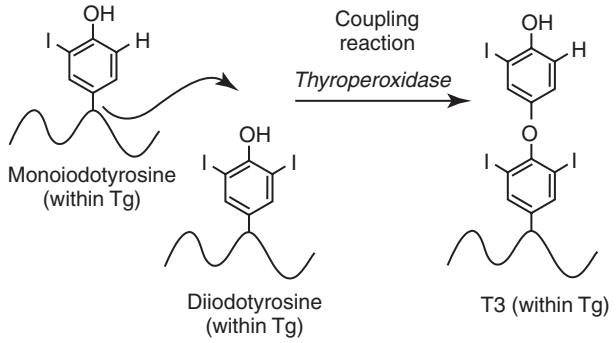


Fig. 42.6 Chemical coupling of one molecule of monoiodotyrosine and one molecule of diiodotyrosine to produce one molecule of T3. The reaction is catalyzed by thyroperoxidase. *Tg,* Thyroglobulin.

and T3, are released into the cytoplasm of the follicular cell. T4 and T3 diffuse into the systemic circulation after their liberation from Tg. DIT and MIT are deiodinated by an intracellular microsomal iodotyrosine dehalogenase (approved gene name iodotyrosine deiodinase, approved symbol *IYD*; chromosome 6q25.1). The freed iodide is then reused for thyroid hormone synthesis.

Each step in the synthesis of thyroid hormones is regulated by pituitary **thyroid-stimulating hormone (TSH)**. TSH (also known as *thyrotropin*) stimulates (1) the "iodide pump," (2) Tg synthesis, and (3) colloidal uptake by follicular cells. TSH also regulates the rate of proteolysis of Tg for the liberation of T4 and T3. In addition, TSH induces an increase in the size and number of the thyroid follicular cells. Prolonged TSH stimulation leads to increased vascularity and eventual hypertrophic enlargement of the thyroid gland (**goiter**). Initially, it was believed that both T4 and T3 entered cells by passive diffusion across the plasma membrane. However, researchers have now shown that thyroid hormones cross plasma membranes using specific transporters. One important thyroid hormone transporter is monocarboxylate transporter 8 (MCT8; gene symbol *SLC16A2*; chromosome Xq13.2). MCT10 (*SLC16A10;* chromosome 6q21–q22) transports iodothyronines and aromatic amino acids across membranes, and MCT10 is likely more preferential than MCT8 in transporting T3 over T4 across plasma membranes. Organic anion transporting polypeptide 1C1 (OATP1C1) also has high affinity for thyroid hormone and may be an important component of the transport of T3 and T4.

Metabolism

Free (unbound) T4 (FT4) is the primary secretory product of the normal thyroid gland. T4 undergoes peripheral deiodination of the outer ring at the 5′ position to yield T3. This deiodination occurs in a number of tissues but primarily in the liver. rT3, produced by removal of one iodine from the inner ring of T4, is metabolically inactive and is an end-product of T4 metabolism. Peripheral deiodination is a rapidly responsive mechanism of control for thyroid hormone balance. Acute or chronic stress or illness causes a shift in the direction of this deiodination, favoring formation of rT3 rather than T3. Various medications also shift peripheral deiodination toward the inactive product rT3.

T4 and T3 in the circulation are bound reversibly and almost completely to carrier proteins. These carrier proteins are (1) **thyroxine-binding globulin (TBG)**, (2) thyroxine-binding prealbumin (TBPA; **transthyretin [TTR]**), and (3) albumin. Collectively, these proteins bind 99.97% of total T4 and 99.7% of total T3. Thus, only a very small fraction of each of these hormones is unbound and free for biological activity. Because a wide variation exists in the concentration of thyroid hormone binding proteins, even under normal circumstances, a wide variation also exists in T4 concentrations among individuals with normal (euthyroid) thyroid function. Total T3 concentrations also vary with alterations in binding proteins, although usually to a lesser degree than T4 concentrations. Circumstances in which thyroid hormone-binding protein concentrations are increased or decreased are shown in Box 42.1.

Regulation and Control

Thyroid hormone synthesis and secretion are controlled by a negative feedback system (Fig. 42.7) involving (1) **thyrotropin-releasing hormone (TRH)** from the hypothalamus, (2) TSH from the pituitary, and (3) thyroid hormones from the follicular cells of the thyroid glands.

BOX 42.1 Alterations in the Concentration or Affinity of Thyroid Hormone-Binding Proteins

Increases In

A. TBG concentration (or affinity)
   1. Genetic (inherited) causes
   2. Nonthyroidal illness (human immunodeficiency virus infection, infectious and chronic active hepatitis, estrogen-producing tumors, acute intermittent porphyria)
   3. Normal physiology (pregnancy, newborn)
   4. Drug use (oral contraceptives, estrogens, tamoxifen, methadone)

B. Prealbumin binding (familial euthyroid thyroxine excess)

C. Albumin binding (familial dysalbuminemic hyperthyroxinemia)

D. Thyroxine binding by autoantibodies (autoimmune thyroid disease, hepatocellular carcinoma)

Decreases In

A. TBG concentration
   1. Genetic (inherited) causes
   2. Nonthyroidal illness (major illness or surgical stress, chronic liver disease, protein-losing enteropathy, nephrotic syndrome)
   3. Drug use (androgens, anabolic steroids, large doses of glucocorticoids)

B. TBG-binding capacity (drugs bound to TBG, such as salicylates and phenytoin)

C. Prealbumin concentration

TBG, Thyroxine-binding globulin.

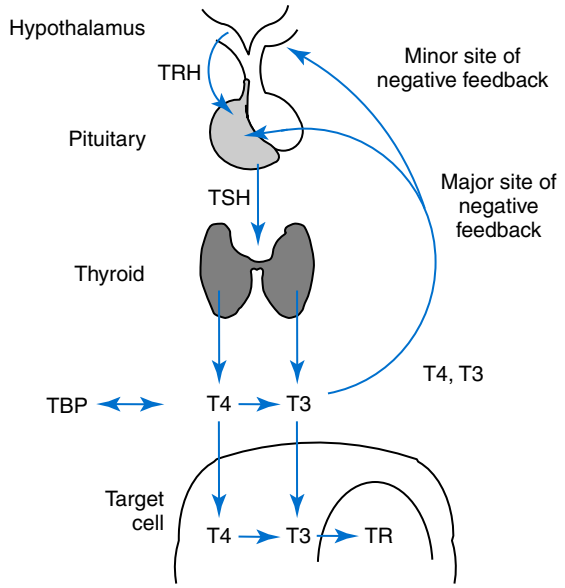
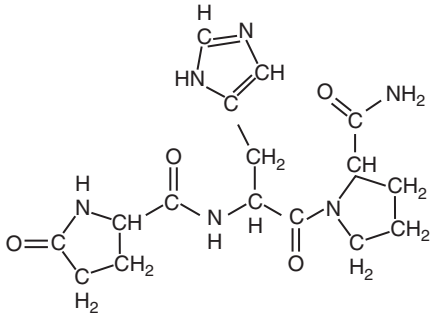


Fig. 42.7 Metabolic control of thyroid hormones. **Thyrotropin-releasing hormone (TRH)** from the hypothalamus enters the hypothalamic-pituitary portal system to release thyroid stimulating hormone (*TSH*; thyrotropin) from anterior pituitary thyrotrophs. TSH stimulates the release of **thyroxine (T4)** and triiodothyronine (T3) from the thyroid gland, although most T3 (~80%) comes from peripheral monodeiodination of T4 to T3. More than 99% of T4 and T3 is bound to various thyroid hormone-binding proteins *(TBP)*. T4 and T3 negatively feedback on the hypothalamus and, more powerfully, the pituitary. T4 and T3 enter into target tissues, where T4 is converted to T3 with T3 binding to the thyroid hormone receptor *(TR)*.



(pyro)Glu-His-Pro(NH_2)

Fig. 42.8 Chemical structure of thyrotropin-releasing hormone *(TRH)*. Note that TRH is a tripeptide (L-pyroglutamyl-L-histidyl-L-prolinamide).

Biochemistry

TRH is the modified tripeptide L-pyroglutamyl-L-histidyl-L-prolinamide secreted by the paraventricular nuclei (PVN) in the hypothalamus (Fig. 42.8). Its concentrations rise in thyroid hormone deficiency with TRH declining when thyroid hormone is in excess. TRH is delivered to the anterior pituitary gland via the hypothalamic-pituitary portal system. TSH (also known as *thyrotropin*) is secreted by anterior pituitary cells known as *thyrotrophs*. In relative terms, TRH has a greater effect on TSH glycosylation than hormone release. However, glycosylation of TSH is necessary for normal TSH bioactivity. When TRH is deficient, TSH may lack potency as the result of insufficient glycosylation, yet nonglycosylated TSH may retain much of its immunoreactivity. Chemically, TSH is a 30 kDa heterodimeric glycoprotein that shares the 14.7 kDa alpha subunit (gene location: chromosome 6q21.1–q23) with (1) luteinizing hormone (LH), (2) follicle-stimulating hormone (FSH), and (3) human chorionic gonadotropin (hCG), whereas each hormone has a unique beta subunit (see Chapters 40 and 44).

Function

The effects of TSH on the thyroid follicular cell are mediated through the TSH receptor. At physiological TSH concentrations, the second messenger system involves G_S proteins, adenylate cyclase, and the formation of cyclic adenosine monophosphate. At supraphysiologic concentrations (100X normal), TSH signals through the inositol-phosphate diacylglycerol cascade. The 743-amino acid glycosylated thyroid-stimulating hormone receptor (TSHR/≈100 kDa) may be in either the "on" conformation, with active signaling, or the "off" conformation, where signaling does not occur.

TSH binding to the TSHR places the TSHR in the "on" conformation, leading to stimulation of the thyroid gland. Within the TSH-stimulated thyroid follicular cell, the (1) activity of the sodium-iodine symporter increases; (2) synthesis of Tg and thyroperoxidase (TPO) increases; and (3) release of T4 and T3 increases. TSH stimulation increases the size and number of thyroid follicular cells.

IODINE AND THYROID DISORDERS

Iodine (or iodide [I^-] in its ionized form) is a trace element and an essential nutrient (see Chapter 27). Iodine is an indispensable component of the thyroid hormones T3 and T4, respectively. These hormones, together with the thyronamines, are the only iodine-containing hormones in vertebrates. Without iodine, there is no biosynthesis of thyroid hormones. Thus, thyroid function requires an adequate supply of iodine to the thyroid gland.

Geography and Iodine Supply

Iodide ions in seawater and coastal seaweed beds are oxidized to elemental iodine, which then volatilizes into the atmosphere and is returned to the soil by rain. This process (the iodine cycle) is slow and incomplete in many regions, thus leaving soils and drinking water iodine depleted.

Most iodine is found in the oceans. The major source of iodine in industrialized countries is iodized salt, and if that is not available, then dairy products.

Cruciferous vegetables, such as cabbage, kale, cauliflower, and broccoli, contain glucosinolates, and their metabolites compete with iodine for thyroidal uptake. Cassava, linseed, and sweet potato contain cyanogenic glucosides, which can be metabolized to thiocyanates and compete with iodine for thyroidal uptake.

Pregnant and lactating women have higher demands for iodine because of an increase in maternal T4 production to maintain maternal euthyroidism and the transfer of thyroid hormone to the fetus early in the first trimester.

Iodine-Containing Products and Goitrogens

Exposure to iodine also can occur from iodine-containing products, such as disinfectants, amiodarone, radiology contrast agents, and topical antiseptics (e.g., the surgical scrub povidone–iodine). The iodine content of these agents is several thousand-fold higher than iodine appearing in naturally occurring foods. Iodine deficiency is the major cause of endemic goiter, but naturally occurring compounds and environmental pollutants may also be goitrogenic. Goitrogens can be either agents acting directly on the thyroid gland or cause a goiter by indirect action. Interestingly, cigarette smoking is associated with high plasma concentrations of thiocyanate, which can compete with iodine for uptake into the thyroid.

Public Health Programs

There are two kinds of public health programs for securing optimal thyroid function. One is newborn screening for thyroid function (see section on congenital hypothyroidism); the other is iodization of salt.

Biochemistry

Dietary iodine controls its own absorption through post-transcriptional regulation of the intestinal Na^+/I^- symporter (NIS). The thyroid gland has autoregulatory mechanisms to handle excess iodine intake involving the sodium iodide symporter. Excess iodine has an inhibitory effect on thyroxine synthesis in intact thyroids. In conditions of adequate dietary iodine supply, the healthy thyroid usually takes up less than 20% of absorbed iodine. However, in chronic iodine deficiency, this fraction can be more than 80%. The thyroid is thus able to adapt to low intakes of dietary iodine by marked modification of its activity. The metabolism of circulating thyroid hormones in peripheral tissues releases iodine that enters the plasma iodine pool, which in turn can be taken up by the thyroid or excreted by the kidney. Approximately 90% of ingested iodine is excreted in the urine and approximately 10% in feces.

Iodine Deficiency

Both deficient and excessive intakes of iodine can impair thyroid function. The clinical presentations of iodine deficiency are largely age dependent. In fetuses, deficiency may cause abortion, stillbirth, and congenital anomalies. In neonates, deficiency may cause infant mortality and endemic **cretinism**. In children and adolescents, deficiency may cause impaired mental function, delayed physical development, and low IQ. In adults, deficiency may cause impaired cognitive function, reduced work productivity, thyroid autoimmunity, toxic nodular goiter, and hypothyroidism. In all ages, goiter is the most visible sign of iodine deficiency but is not specific to it.

Iodine Excess

Thyroid dysfunction may occur in vulnerable patients if exposed to excess iodine. These patients include those with preexisting thyroid disease, older adults, pregnant and lactating women, fetuses, and neonates. Iodine excess can result in adverse thyroidal effects after only a single exposure to iodine-containing substances. Excess dietary iodine intake in pregnant and lactating women in iodine-replete areas may lead to an increase in plasma TSH concentrations and thus to **subclinical hypothyroidism (SCH)**.

EXTRATHYROIDAL FACTORS THAT AFFECT THYROID FUNCTION

Epidemiologic Factors

Concentrations of thyrotropin, thyroid hormones, thyroid antibodies, and thyroid-binding proteins are to varying degrees determined genetically and by epidemiologic factors such as age, gender, ethnicity, body mass index, smoking, pregnancy, nutritional iodine, season, nonthyroidal disease, radiation, and medication.

Drugs

Many drugs, other than those used for treatment of thyroid disorders, interfere with thyroid hormone homeostasis (Table 42.2) through actions on thyroid hormone synthesis, secretion, transport, metabolism, and absorption.

Lithium

Lithium, which is still a widely used drug for the treatment of bipolar disorder, is associated with subclinical and

TABLE 42.2 Effects of Some Drugs on Tests of Thyroid Function

Cause	Drug	Effect
Inhibit TSH secretion	Dopamine L-dopa Glucocorticoids Somatostatin	↓ T4; ↓ T3; ↓ TSH
Inhibit thyroid hormone synthesis or release	Iodine Lithium	↓ T4; ↓ T3; ↑ TSH
Inhibit conversion of T4 to T3	Amiodarone Glucocorticoids Propranolol Propylthiouracil Radiographic contrast agents	↓ T3; ↑ rT3; ↓, ≒, ↑ T4 and FT4; ≒, ↑ TSH
Inhibit binding of T4/T3 to serum proteins	Salicylates Phenytoin Carbamazepine Furosemide NSAIDs Heparin (in vitro effect)	↓ T4; ↓ T3; ≒, ↑ FT4; ≒TSH
Stimulate metabolism of iodothyronines	Phenobarbital Phenytoin Carbamazepine Rifampicin	↓ T4; ↓ FT4; ≒TSH
Inhibit absorption of ingested T4	Aluminum hydroxide Ferrous sulfate Cholestyramine Colestipol Iron sucralfate Soybean preparations Kayexalate	↓ T4; ↓ FT4; ↑ TSH
Increase in concentration of T4-binding proteins	Estrogen Clofibrate Opiates (heroin, methadone) 5-Fluorouracil Perphenazine	↑ T4; ↑ T3; ≒FT4; ≒TSH
Decrease in concentration of T4-binding proteins	Androgens Glucocorticoids	↓ T4; ↓ T3; =FT4; ≒TSH

↓, Reduced serum concentration; ↑, increased serum concentration; ≒ no change; *FT4,* free thyroxine; *NSAID,* nonsteroidal antiinflammatory drug; *rT3,* reverse triiodothyronine; *T3,* triiodothyronine; *T4,* thyroxine; *TSH,* thyroid-stimulating hormone.
Data obtained from Smallridge, R. D. (1995). Thyroid function tests. In K. L. Becker (Ed.), *Principles and practice of endocrinology and metabolism* (7th ed., pp. 299–306). Philadelphia: JB Lippincott.

overt hypothyroidism in up to 34% and 15% of patients, respectively. These may develop even after many years of treatment. Patients should have regular thyroid function tests, at least once or twice per annum. Lithium primarily inhibits thyroid hormone secretion, although it appears also to have effects on iodine trapping, release, and coupling.

Amiodarone

Amiodarone is an antiarrhythmic drug that comprises 37% iodine by weight and has structural similarities with thyroid hormones. The drug has a long half-life, large distribution volume, and wide tissue distribution. The mechanisms of amiodarone injury are multifactorial and involve the accumulation of iodine, the formation of free radicals, and immunologic damage. The effects of amiodarone can be divided into effects occurring in everyone treated with amiodarone, resulting in changes in thyroid function tests ("obligatory effects"), and effects only occurring in some people treated with amiodarone ("facultative effects"), resulting in clinically overt thyrotoxicosis or hypothyroidism. It may take several months for normalization of thyroid function tests after discontinuation of amiodarone treatment.

Amiodarone-induced thyrotoxicosis (AIT) is particularly prevalent (10%) in iodine-deficient regions, in men, and in patients with underlying thyroid disease (e.g., nodular goiter or Graves disease). Amiodarone-induced hypothyroidism occurs primarily in iodine-sufficient regions and in women with preexisting anti-TPO antibodies.

Nonthyroidal Illness

Many disorders are associated with variation in the concentration of thyroid hormones in the absence of definable thyroid disease (Table 42.3). For example, (1) significant nutritional deprivation, (2) acute severe illness, or (3) chronic illness often result in changes in thyroid function. Collectively they are characterized as *nonthyroidal illnesses* (NTI; **sick euthyroid syndrome**). A progressive spectrum of thyroid test result anomalies often accompanies nonthyroidal illnesses in euthyroid patients. The earliest and most common changes that occur are a reduction in the serum concentrations of total and free T3, sometimes to extremely low concentrations, and an elevation in the serum concentration of rT3 (the "low T3 state"). These changes have been ascribed to a block in the 5′-deiodinases that convert T4 to T3 in peripheral tissue. This conversion is inhibited in patients with (1) acute and chronic nutritional problems, (2) poorly controlled diabetes mellitus, and (3) the use of drugs such as hydrocortisone and beta blockers.

Serum TSH concentrations are usually normal in euthyroid sick patients, but may be mildly to moderately depressed with moderate to severe NTI or slightly elevated during recovery from a severe illness (Fig. 42.9). Causes of these transient abnormal TSH concentrations are not fully understood, but may relate to the effects of endogenous or exogenous hormones, such as glucocorticoids or dopamine, which independently suppress secretion of pituitary TSH concentrations. Other possible causes include altered nutrition or altered biological activity of immunoreactive TSH. Decreased T4 levels (both total and free) may result from T4 displacement from thyroid hormone binding proteins. As patients recover from NTIs, many of the thyroid test abnormalities revert to normal. T4 concentrations will be corrected first, followed by a rise in the concentration of T3. Serum TSH may also transiently rebound to high concentrations for several days or weeks before returning to normal.

TABLE 42.3 Nonthyroidal Illness

Hormone	Initial	Midcourse	Prolonged	Resolution
TSH	Normal	Normal to decreased	Decreased	Increases
T4	Normal	Normal to decreased	Decreased	Increases
T3	Decreased (–)	Decreased (––)	Decreased (–––)	Increases
rT3	Increased (+)	Increased (++)	Increased (+)	Decreases

Note: The severity of decreased T3 is proportional to the number of (–) signs. The degree of increase in rT3 is proportional to the number of (+) signs.
T3, Triiodothyronine; *rT3,* reverse T3; *T4,* thyroxine; *TSH,* thyroid-stimulating hormone.

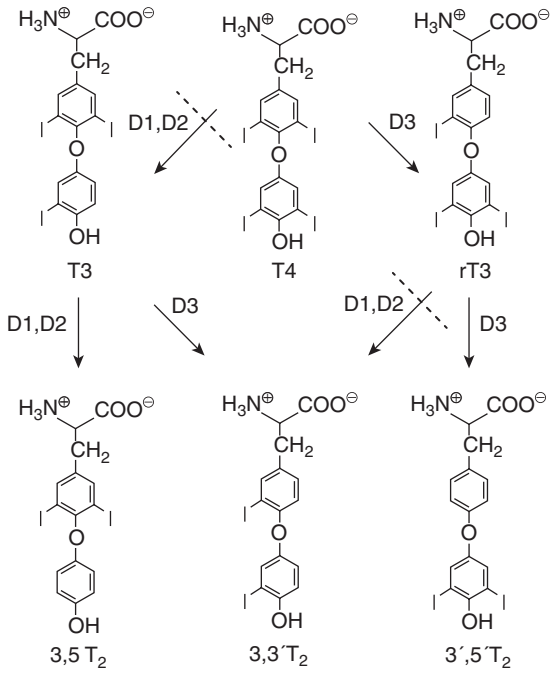


Fig. 42.9 Effects of altered deiodination. In states of nonthyroidal illness (e.g., the sick euthyroid syndrome), thyroid hormone deiodination patterns are altered with reduced T3 concentrations and elevated concentrations of reverse T3 *(rT3).*

EVALUATION OF THYROID FUNCTION

In clinical practice, thyroid function tests are routinely measured to diagnose disorders of the thyroid. For example, in combination with information obtained from (1) history, (2) physical, and (3) laboratory results, patients are classified as (1) hypothyroid, (2) hyperthyroid, or (3) euthyroid (Fig. 42.10). Almost all laboratory tests for thyroid function are commercially available on automated immunoassay instruments. The following is a brief description of tests that are used for the evaluation of thyroid status. Package inserts that accompany commercial products are also a source of additional information. Reference intervals for the analytes, discussed later, are found in the Appendix.

Measurement of Thyroid Stimulating Hormone

Immunoassay is the method of choice for the measurement of serum TSH in the clinical laboratory. Human TSH is a 28 to 30 kDa dimeric glycoprotein consisting of a 92-amino acid alpha subunit (identical to the alpha subunit of hCG, FSH, and LH) and a 112 to 118 amino acid beta subunit. Current third-generation TSH assays distinguish mild to modest TSH suppression (due to, e.g., overtreatment of hypothyroidism) from severely suppressed TSH levels as observed in Graves disease. All assays in clinical practice should be "third generation"; that is, they should have a coefficient of variation (CV) of less than 20% (functional sensitivity) at a concentration of 0.01 mIU/L

Principles of Thyroid-Stimulating Hormone Immunoassays

Most TSH methods currently used in clinical laboratories are two-site "sandwich" heterogeneous immunoassays involving (1) enzyme, (2) fluorometric substrate, or (3) chemiluminescent labels. Heterogeneous immunoassays differ in the method used to separate bound and free fractions before application of the signal antibody. Immunometric assays generate a proportional dose-response calibration curve with higher signals corresponding to increased concentrations of analyte. Because the physiological interval of TSH concentrations is typically less than three orders of magnitude, high-dose hook effects are rarely encountered with TSH immunoassays.

Specimen Collection and Storage

Both serum and plasma are used for TSH measurements. TSH is stable for 5 days at 2°C to 8°C, and for at least 1 month when stored frozen. For newborn screening, whole blood may be collected by heel puncture 48 to 72 hours after birth.

Comments on Thyroid-Stimulating Hormone Measurements

Secretion of TSH is circadian with peak concentrations of TSH occurring between 0200 and 0400 hours, and the lowest between 1700 and 1800 hours. Low-amplitude oscillations occur throughout the day. TSH surges immediately after birth, reaching a peak of 25 to 160 mIU/L within 30 minutes and declining back to cord blood concentrations by postpartum day 3. TSH concentrations stabilize to near adult concentrations within the first few weeks of life. In the first trimester of pregnancy, TSH concentrations decline as hCG stimulates the maternal thyroid gland to produce thyroid hormone, sometimes leading to a TSH concentration that is just below the lower reference limit. Reference intervals for adult TSH concentrations are the same for men and women (generally near 0.4 to 4.5 mIU/L). However, there is active discussion that the upper limit of the reference interval may be as low as

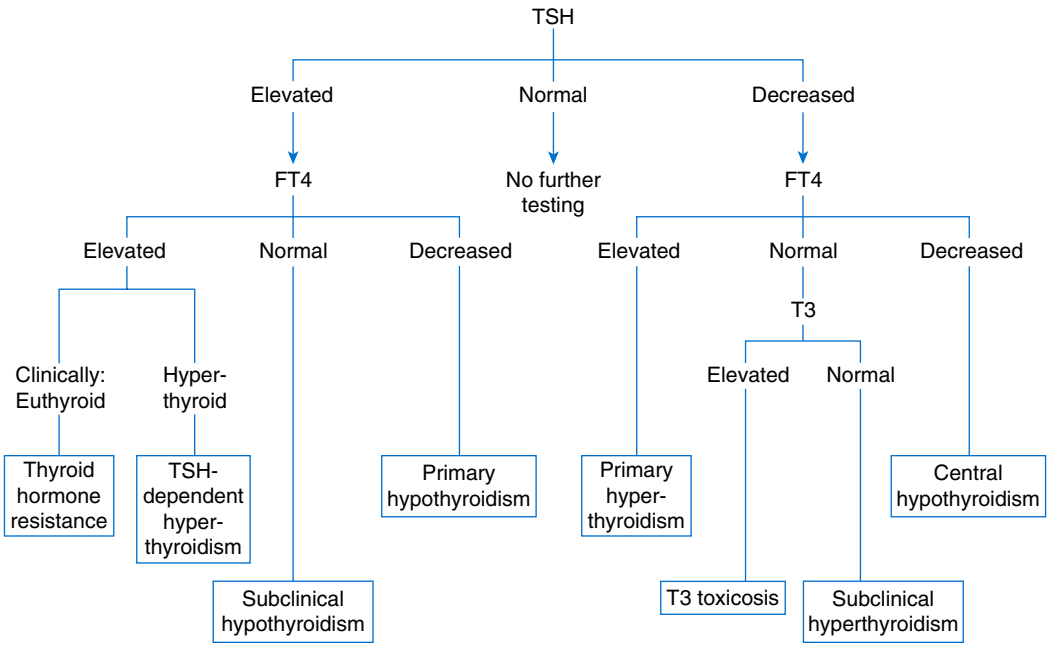


Fig. 42.10 Suggested algorithm for the laboratory evaluation of thyroid function. *TSH,* Thyroid-stimulating hormone.

2.5 to 3.5 mIU/L if people with thyroid autoantibodies and/or abnormal thyroid ultrasound findings are excluded from the reference population.

Measurement of Total Thyroxine

T4 is the principal hormone secreted by the thyroid, and its circulating concentration is under the control of TSH. The hormone is synthesized from tyrosine residues on Tg in the colloid of the thyroid gland. Circulating T4 is highly (>99.9%) protein bound, and the fraction that is bound to protein is biologically inactive. Immunoassays and instrument-based methods are used to measure total T4 in biological fluids, predominantly serum or plasma.

Immunoassays

Many clinical laboratories measure T4 by using an automated competitive immunoassay. Most of these T4 immunoassays use high-affinity polyclonal antibodies produced against an albumin-T4 conjugate. These polyclonal antibodies are highly specific and display minimal cross-reactivity with T3. Tg sometimes has been used as the immunogen, because it contains iodinated tyrosine residues that are the precursors of T4 and T3. Monoclonal antibodies against T4 have also been developed.

Immunoassays of T4 measure both free and protein-bound hormone. Accurate measurement of T4 therefore requires dissociation of the hormone from serum proteins, such as (1) TBG, (2) albumin, and (3) TTR, which bind more than 99.9% of circulating T4. Various blocking agents have been used to dissociate T4 from TBG. The most common of these blocking agents is (1) 8-anilino-1-naphthalene-sulfonic acid (ANS), although (2) salicylate, (3) thimerosal, and (4) phenytoin have also been used for this purpose. Barbital is used to dissociate T4 from TTR.

Nonisotopic assays for T4 are widely available for use on various automated immunoassay systems, or adaptable to automated chemistry platforms. Chemiluminescent and electrochemiluminescent assays also have been adapted for use on automated platforms.

OTHER METHODS

Other instruments used to measure T4 in human serum include (1) electron capture gas chromatography, (2) high-performance liquid chromatography, and (3) isotope dilution tandem mass spectrometry. The latter method has been suggested as a reference method for assay of T4.

Specimen Collection and Storage

Serum is the preferred specimen for the measurement of T4, but plasma with ethylenediaminetetraacetic acid (EDTA) or heparin as an anticoagulant has also been used. If plasma is used, repeated freeze-thaw cycles should be avoided as fibrin clots may form producing potentially spurious results due to changes in specimen viscosity. Gel barrier collection devices do not have any apparent adverse effect on T4 methods. T4 is a stable analyte with no appreciable change in concentration for up to 7 days at room temperature, or 30 days when frozen. Mild to moderate hemolysis and lipemia do not significantly affect most T4 immunoassays; however, grossly hemolyzed specimens should be avoided because of dilutional effects. T4 autoantibodies interfere with some immunoassays and may produce erroneously low or high results, depending on the method. Such interferences should be considered when there is discordance between the clinical and laboratory findings. Heel-stick capillary specimens have been used for T4 measurement, as have dried blood specimens collected on filter paper. Dried blood specimens are stable and easily

transported and are widely used in the United States to screen neonates for **congenital hypothyroidism**. T4 is extracted from a punched-out disk of blood saturated filter paper into a buffer before assay. The filter paper should not be exposed to extreme heat or light.

Comments on Total Thyroxine Measurements

Because it reflects mostly inactive (protein-bound) hormone, the quantity of total T4 alone provides limited clinical information. For this reason, FT4 measurements are preferred over total T4 measurements for routine testing. Cord blood T4 concentrations are lower in preterm than in full-term neonates, and they correlate positively with birth weight in full-term infants. At birth, serum total T4 concentrations are higher in neonates because of the maternal estrogen-induced increase in serum TBG; FT4 concentrations are near adult concentrations. The concentration of total T4 rises abruptly in the first few hours after birth and declines gradually until adolescence.

Measurement of Triiodothyronine

T3 is the principal active thyroid hormone with approximately 20% of circulating T3 secreted by the thyroid gland, and the remainder is produced by deiodination of T4 in the peripheral tissues. Like T4, T3 is highly bound to protein with less than 1% of the total concentration as free active hormone. However, compared with T4, it is less tightly bound to serum proteins by about an order of magnitude. Consequently, the displacement of bound T3 from proteins is essentially complete in the presence of conventional blocking agents, such as ANS. Also, T3 does not ordinarily displace T4 from thyroid hormone-binding proteins.

Immunoassays are the techniques of choice to measure T3 in body fluids, predominantly serum or plasma.

Nonisotopic Immunoassays

Nonisotopic immunoassays for T3 are similar to total T4 methods. Many of the T3 methods have been developed for use on automated immunoassay systems, and some are compatible with chemistry platforms. Many commercial methods use enzyme labels, such as peroxidase or alkaline phosphatase, conjugated to T3 or anti-T3 antibodies. Enzyme activity is determined by using a variety of sensitive (1) photometric, (2) fluorescent, or (3) chemiluminescent substrates. Automated platforms that employ chemiluminescent labels are the analytical systems of choice to measure T3. Both heterogeneous and homogeneous immunoassays for T3 have been described (see Chapter 15).

Specimen Collection and Storage

Serum is the preferred specimen, but plasma with EDTA or heparin as an anticoagulant may be used. Serum specimens should be tested within 24 hours of collection, or stored at 2°C to 8°C if tested beyond 24 hours. Frozen specimens are stable for at least 30 days. Turbid samples may require centrifugation before testing.

Comments on Total Triiodothyronine Measurements

Significant discrepancies have been observed when the results of different T3 immunoassays were compared using reference sera. Interlaboratory quality assurance (proficiency testing) results also demonstrate higher analytical variance for T3 compared with T4 methods. Among the factors that have been suggested to account for the poorer analytical performance of T3 immunoassays are the (1) lower concentration of T3 in serum, (2) greater antibody cross-reactivity, (3) protein interferences, and (4) different limits of detection for T3 methods. Total T3 measurements are useful in the diagnosis and monitoring of hyperthyroid patients with suppressed TSH and normal FT4 concentrations (**T3 thyrotoxicosis**); T3 measurements have only limited value in euthyroid and hypothyroid patients. Thyroid hormone replacement in hypothyroid patients is monitored by TSH and FT4 measurements, and is not monitored by measuring T3.

Measurement of Reverse Triiodothyronine

Several radioimmunoassay (RIA) methods for measuring rT3 have been developed and commercial assays are available. However, none of these assays have been adapted to automated platform because rT3 measurement has limited diagnostic value. The diagnosis of NTI can usually be established without measuring rT3.

Measurement of Free Thyroid Hormones

Numerous methods have been developed for assessing the concentrations of FT4 and FT3 in serum. These methods include (1) direct assays that currently serve as reference methods, and (2) indirect or estimate assays that are more widely available for general laboratory use.

Direct Reference Methods

Direct measurement of FT4 (and FT3) in serum is a technical challenge as the quantity of free hormone in normal serum is exceedingly low—typically less than 100 pmol/L (7.8 ng/dL).

Reliable methods for measuring FT4 and FT3 in serum include those that separate free and bound hormone fractions by (1) direct equilibrium dialysis or (2) ultrafiltration. After separation, the free concentration is measured by a sensitive analytical method, such as immunoassay or mass spectrometry.

Direct equilibrium dialysis. In this procedure isotopically labeled T4 or T3 "tracer" is added to the specimen before dialysis is performed. When equilibrium is established, concentrations of the tracer on both sides of the dialysis membrane are used to calculate the ratio of bound to free hormone, so the FT4 (or FT3) concentration can be calculated based on the total T4 or T3 concentration.

If the amount of tracer added to the specimen is small in comparison with the endogenous hormone concentration, minimal disruption of free/bound hormone is noted.

Ultrafiltration. The ultrafiltration procedure for FT4 determination in serum is significantly less time-consuming than dialysis. In this method, the serum specimen is (1) adjusted to a pH of 7.4, (2) incubated for 20 minutes at 37°C

(to achieve equilibrium of the binding at this temperature), and then (3) applied to an ultrafiltration device for centrifugation for 30 minutes at 37°C and 2000 × *g* (using a fixed-angle rotor). Subsequently the ultrafiltrate is analyzed for T4 by immunoassay.

Free T4 assays based on ultrafiltration measure free hormone without the need for total hormone measurements. Direct equilibrium dialysis and ultrafiltration methods are unaffected by either variation in serum binding proteins or thyroid hormone autoantibodies. Mean values obtained in euthyroid healthy subjects are reported to be slightly higher when using ultrafiltration methods than when using equilibrium dialysis.

Indirect Methods for Estimating Free Thyroid Hormones

Two-step and one-step immunoassays estimate free hormone concentrations by using antibody extraction techniques. They are subdivided into *sequential two-step assays* and *simultaneous one-step ("analogue") assays.* The analogue methods are no longer used, and the sequential two-step assays are preferred. Each procedure involves the direct incubation of serum with a specific anti-T4 or anti-T3 antibody, during which thyroid hormones reach a new equilibrium with all of the binders present. A slight decrease in free hormone concentration occurs, but is insignificant if the antibody sequesters less than 5% of the total amount of hormone present in the specimen. Thus, the amount of immunoextracted T4 or T3 closely approximates the undisturbed free hormone concentration that preexists in serum at equilibrium.

Estimates of FT4 and FT3 generally give reliable results in (1) healthy subjects, (2) hyperthyroid and hypothyroid patients, and (3) patients with only mild binding protein abnormalities. Results are comparable with those of reference methods, such as those that use direct equilibrium dialysis. In certain clinical conditions, free hormone estimate methods may give abnormal results that differ from the generally accepted reference limits. These abnormalities are commonly encountered in euthyroid patients who show significant changes in T4 or T3 binding to serum proteins.

Measurement of Thyroxine-Binding Globulin

Modern TBG methods measure its concentration directly using a variety of immunochemical approaches. For example, one competitive, heterogeneous method measures the competition between endogenous TBG and labeled TBG for binding to an immobilized anti-TBG antibody. Another assay utilizes a solid-phase second antibody; bound conjugate that is measured by chemiluminescence after the addition of Luminol and hydrogen peroxide. A third immunoassay is based on enhanced microparticle turbidimetry, in which the presence of endogenous antigen inhibits the cross-linking of antigen-microparticle complexes by anti-TBG antibody, reducing the turbidity of the reaction mixture.

Measurement of Thyroglobulin

Tg is detectable in the plasma, and as such acts as a marker for the presence of active thyroid tissue. Tg concentrations are determined by thyroid mass, TSH stimulation, and thyroid manipulation including surgery, fine-needle aspiration, or thyroid injury. The primary use of serum Tg measurement is as a tumor marker in patients with differentiated thyroid cancer (DTC) who have undergone thyroidectomy. It is difficult to measure Tg owing to the heterogeneity of the molecule, largely because of different glycoforms, and to the prevalence of endogenous anti-Tg or anti-reagent antibodies that interfere with the immunoassay. Current practice for thyroidectomized thyroid cancer patients is to withdraw thyroxine replacement to stimulate endogenous TSH secretion or to use recombinant human TSH to stimulate any residual or neoplastic thyroid tissue. This can be inconvenient, unpleasant, and expensive. Consequently, a Tg assay with the sensitivity to detect Tg in the absence of TSH stimulation (i.e., suppressed) is desirable. Three methods are in current use, each with its own advantages and limitations.

Competitive immunoassays have lower sensitivity (~5 ng/mL), are more labor intensive, but are robust to antibody interference compared to immunometric assays.

Immunometric assays are very sensitive (0.1 ng/mL or lower), amenable to automation, but prone to antibody interference. Immunometric Tg assays should be reported together with anti-Tg results.

Peptide mass spectrometric assays are limited by the relatively poor analytical sensitivity and low available mass range of mass spectrometric methods. However, they offer the promise of robustness to antibody interference.

Measurement of Thyroid Autoantibodies

Increased circulating concentrations of thyroid autoantibodies are found in a variety of thyroid disorders and in other autoimmune diseases and certain malignancies. These autoantibodies are directed against several thyroid and thyroid hormone antigens, including (1) Tg (TgAb), (2) thyroid peroxidase antibody (TPOAb), (3) the TSH receptor (TRAb), (4) TSH, (5) T4, and (6) T3. Of these autoantibodies, measurement of TPOAb is the test most often used to evaluate thyroid autoimmune diseases. TPOAb and TgAb are detected by a variety of methods, including (1) indirect immunofluorescence, (2) the agar gel diffusion precipitin technique, (3) agglutination (hemagglutination or latex particle agglutination), (4) RIA, (5) complement fixation, (6) ELISA techniques, and (7) chemiluminescence-based immunometric assays.

Measurement of Urinary Iodine Concentration

Because most of the body's iodine is excreted in urine, urinary iodine concentration (UIC) is considered a reliable and valid biomarker of the iodine intake and iodine deficiency of the population. Secondary measurements for estimating iodine deficiency are thyroid size, TSH concentration, and thyroid hormones.

THYROID DISEASE

Thyroiditis

Thyroiditis means inflammation of the thyroid. It encompasses many thyroid disorders, some of which have an

TABLE 42.4 Types of Autoimmune Thyroiditis

	Course	Features
Goitrous (Hashimoto) thyroiditis	Chronic	Goiter, lymphocytic infiltration, fibrosis, thyroid cell hyperplasia
Atrophic thyroiditis (primary myxedema)	Chronic	Atrophy, fibrosis
Juvenile thyroiditis	Chronic (may disappear)	Usually lymphocytic infiltration
Postpartum thyroiditis	Transient: may progress to chronic thyroiditis	Small goiter; some lymphocytic infiltration
Silent (painless) thyroiditis	Transient	Small goiter; some lymphocytic infiltration
Focal thyroiditis	Progressive in some patients	Present in the thyroid glands of 20% of people at autopsy

Reproduced from Weetman, A. P. (2012). Chronic autoimmune thyroiditis. In L. E. Braverman, & D. S. Cooper (Eds.), *Werner & Ingbar's the thyroid* (10th ed., p. 56). Philadelphia: Lippincott, Williams & Wilkins. Copyright Lippincott, Williams & Wilkins.

autoimmune etiology, including Hashimoto thyroiditis, painless sporadic thyroiditis, and painless postpartum thyroiditis (Table 42.4). The prevalence of individuals with high serum concentrations of thyroid antibodies varies with gender, age, and ethnicity; the prevalence is higher in women, in people older than 60 years of age, and in whites. It has been suggested that dietary iodine deficiency may be protective against certain types of thyroiditis because of the geographical variations in incidence seen in Hashimoto thyroiditis, painless postpartum thyroiditis, and painless sporadic thyroiditis.

The different types of thyroiditis may cause thyrotoxicosis, hypothyroidism, or both. All types of thyroiditis may eventually progress to permanent hypothyroidism, and the risk is higher in patients with high serum thyroid antibody concentrations. As thyroid function declines, TSH secretion increases, resulting in SCH (elevated TSH, normal T4 and T3), and later in overt hypothyroidism with clearly elevated TSH and a preferential expression of T3 but low T4, and finally, in thyroid failure, also low T3.

Autoimmune Thyroid Disease

Autoimmune thyroid disease comprises two main diseases, Hashimoto thyroiditis and Graves disease. The prevalence is estimated to be 5%, but the prevalence of thyroid antibodies in the general population is 10% to 20%. **Autoimmune thyroid disease (AITD)** is a T-cell-mediated disorder leading to an immune attack of the thyroid in an individual with genetic susceptibility accompanied by environmental factors. Examples of factors that increase the risk of AITD include ethnicity, pollutants (smoking, radioiodine), dietary (iodine excess, selenium deficiency), endocrine (female gender, parity, postpartum, oral contraceptives), infections, drugs, and trauma. Pathologically, AITD is characterized by lymphocyte infiltrates within the thyroid. AITD is associated with other autoimmune diseases, either organ specific (autoimmune polyglandular syndromes) or systemic (type 1 diabetes, rheumatoid arthritis, systemic lupus erythematosus, Sjögren syndrome, systemic sclerosis, cryoglobulinemia, sarcoidosis, psoriatic arthritis); these associations reflect common genetic susceptibilities and common pathogenic mechanisms. AITD is also associated with papillary thyroid cancer.

Hypothyroidism

Hypothyroidism is defined as a deficiency in thyroid hormone production or secretion producing a variety of clinical signs and symptoms of hypometabolism. The term *myxedema* is used in severe or complicated cases but strictly refers only to the appearance of the skin as it becomes infiltrated with glycosaminoglycans. Hypothyroidism is a common endocrine disorder that is often overlooked but often may present with serious signs and symptoms. The disorder is treatable with a good prognosis. Thyroid hormones affect the function of most of the organs and tissues of the body (Box 42.2). The clinical features of thyroid hormone deficiency are therefore quite diverse and may involve multiple systems. The thyroid gland itself may be enlarged as a goiter; either firm or granular in texture; tender or nontender, or normal, small, or impalpable, depending on the etiology of the hypothyroidism. Importantly, hypothyroidism may be overt or subclinical. Overt hypothyroidism is defined by a high serum TSH concentration and low serum FT4; SCH is defined by a high serum TSH (but usually <10 mIU/L) and a normal serum FT4 concentration.

Epidemiology

Worldwide, the most common cause of hypothyroidism is still iodine deficiency. However, in developed countries where iodine fortification is widespread, the most common cause of **primary hypothyroidism** is Hashimoto thyroiditis. Hypothyroidism is more common in women than men, increases with age, and is higher in whites than in blacks or Hispanics.

Etiology

Hypothyroidism has many etiologies (some of which are described in other sections of this chapter), including thyroiditis, particularly of autoimmune origin, postpartum phase, drugs, previous thyroid injury, hypothalamic or pituitary disorders in **central hypothyroidism**, and iodine deficiency (Boxes 42.3 and 42.4). Congenital hypothyroidism is discussed in a separate section.

Hashimoto thyroiditis. **Hashimoto thyroiditis** was first described in 1912 by Hakaru Hashimoto, a Japanese physician.

BOX 42.2 Clinical Manifestations of Hypothyroidism

Symptoms
Fatigue
Lethargy
Sleepiness
Mental impairment
Depression
Cold intolerance
Hoarseness
Dry skin
Hair loss
Decreased perspiration
Weight gain
Decreased appetite
Constipation
Menstrual disturbances (typically menorrhagia) and infertility
Arthralgia
Paresthesia

Signs
Goiter (may be present)
Slow movements
Slow speech
Hoarseness
Bradycardia
Dry skin
Loss of outer lateral eyebrow
Nonpitting edema (myxedema) (caused by accumulation of glycosaminoglycans in subcutaneous and other interstitial tissue)
Carpal tunnel syndrome
Psychosis
Galactorrhea
Hyporeflexia
Delayed relaxation of reflexes
Myopathy
Congestive cardiac failure (severe hypothyroidism)
Coma (severe hypothyroidism)
Growth failure and mental retardation (undetected congenital hypothyroidism)
Prolonged jaundice (neonatal hypothyroidism) (as a result of immaturity of uridine diphosphate glucuronyltransferase)

BOX 42.3 Causes of Hypothyroidism

Primary Hypothyroidism
Thyroid dysgenesis
Chronic autoimmune thyroiditis: atrophic and goitrous forms
Defective thyroid hormone biosynthesis
Congenital defects in thyroid hormonal biosynthesis
Iodine deficiency

Central Hypothyroidism
Pituitary disease
Hypothalamic disease

Transient Hypothyroidism
Silent (painless) thyroiditis including postpartum thyroiditis
Subacute thyroiditis (De Quervain syndrome)

BOX 42.4 Other Causes of Primary Hypothyroidism

Thyroid ablation
Subtotal and total thyroidectomy
Radioactive iodine therapy or surgery for
- Graves disease
- Thyrotoxicosis
- Toxic nodular goiter

External radiotherapy for
- Hodgkin and non-Hodgkin lymphoma
- Solid cancers of the head and neck
- Aplastic anemia
- Leukemia

Pharmacologic agents
- Lithium carbonate
- Cytokines (interferon-α, interleukin-2)
- Tyrosine kinase inhibitors
- Iodine and iodine-containing drugs (e.g., amiodarone)
- Other drugs: aminoglutethimide, ethionamide, sulfonamides

Infiltrative disorders
- Riedel (invasive fibrous) thyroiditis
- Amyloidosis
- Sarcoidosis
- Hemochromatosis
- Scleroderma
- Cystinosis
- AIDS (including *Pneumocystis* infection)
- Primary thyroid lymphoma

Toxic substances
- Cigarettes
- Industrial and environmental agents

Embryological variants
- Lingual thyroid

Reproduced from Singer, P. A. (2013). Primary hypothyroidism due to other causes. In L. E. Braverman, & D. S. Cooper (Eds.), *Werner & Ingbar's the thyroid* (10th ed., p. 552). Philadelphia: Lippincott, Williams & Wilkins. Copyright Lippincott, Williams & Wilkins.

Also known as chronic lymphocytic thyroiditis, Hashimoto thyroiditis is the most common type of thyroiditis, and in iodine-sufficient areas it is also the most frequent cause of hypothyroidism and goiter. It leads to the destruction of thyroid follicular cells through a T-cell–mediated autoimmune process. Histologically, the gland is infiltrated with lymphocytes and plasma cells, which can lead to the development of secondary lymphoid follicles within the gland, which are similar to the secondary follicles observed in normal lymph nodes. The initial finding is usually a firm, symmetric goiter, but over time, the gland can atrophy, reflecting destruction of the gland. The gland may often be lobulated, which sometimes makes it difficult to distinguish from multinodular goiter. Although atrophy may occur in Hashimoto thyroiditis, atrophic thyroiditis may also occur when autoantibodies against the TSHR bind to the receptor and block the action of endogenous TSH. TSHR-blocking autoantibodies can cross the placenta during pregnancy, causing transient hyperthyrotropinemia in infants (elevated TSH with a normal T4) or even

transient congenital hypothyroidism. Therefore, atrophic thyroiditis falls into the spectrum of AITDs.

The diagnosis of Hashimoto thyroiditis is supported by recognition of autoantibodies against TPO (anti-TPO) or Tg (anti-Tg). Ninety percent of patients with chronic lymphocytic thyroiditis (the histologic description of Hashimoto thyroiditis) have anti-TPO at presentation, and 20% to 50% have anti-Tg autoantibodies, making these autoantibodies excellent markers for the condition. The thyroid is hypoechogenic on ultrasound examination.

If overt hypothyroidism is present or SCH with high serum thyroid antibody concentrations, patients should be treated with levothyroxine with the goal of normalizing TSH. TSH-suppressing doses can be used short term to reduce goiter size. TSH should be used for monitoring. Serum thyroid antibody concentrations do not decrease with treatment.

There is a strong association between Hashimoto thyroiditis and primary B-cell lymphoma of the thyroid, although this is a rare condition. It is thought that prolonged stimulation of the intrathyroid B cells results in the emergence of a malignant clone. In addition, the frequency of papillary carcinoma may be increased in chronic autoimmune thyroiditis, particularly in women. Pain may indicate the presence of lymphoma. Fine-needle aspiration (FNA) biopsy or surgical biopsy may be required to distinguish between these conditions.

Central hypothyroidism. **Central hypothyroidism** is caused by an insufficient stimulation by TSH of an otherwise normal thyroid gland. The prevalence is estimated to be 1 in 20,000 to 1 in 80,000 in the general population, accounting for 1 in 1000 hypothyroid patients. Neonatal screening programs have shown a prevalence of congenital hypothyroidism of central origin of 1 in 160,000. In the Netherlands, the neonatal screening program using a combined TBG, TSH, and T4 strategy has shown that central hypothyroidism is diagnosed earlier with milder forms with an incidence of 1 in 16,000.

Central hypothyroidism may arise from pituitary disorders (secondary) or hypothalamic (tertiary) disorders including the pituitary stalk (Box 42.5). Tertiary hypothyroidism is a result of insufficient TSH stimulation by TRH.

Central hypothyroidism can be classified into invasive or compressive lesions, iatrogenic factors (e.g., cranial surgery or irradiation), injuries (e.g., head traumas), vascular accidents (e.g., pituitary apoplexia, postpartum pituitary [Sheehan] syndrome), autoimmune disease (e.g., polyglandular autoimmune diseases), infiltrative lesions (e.g., iron overload, sarcoidosis), inherited diseases, and infectious diseases (e.g., tuberculosis). Invasive or compressive lesions include pituitary macroadenomas, craniopharyngiomas, meningiomas or gliomas, Rathke cleft cysts, metastases, empty sella syndrome, and carotid aneurysms. The inherited forms are rare and may include TSHβ mutations or TRH receptor mutations or pituitary transcription factor defects (mutations in *POU1F1, PROP1, HESX1, LHX3, LHX4,* or *LEPR*) with combined pituitary hormone deficiencies including TSH deficiency.

Transient or reversible forms of central hypothyroidism can be observed with drugs affecting the neuroendocrine TSH regulation, such as somatostatin analogs, glucocorticoids, or dopaminergic compounds acutely inhibiting TSH secretion; however, the thyroid stimulation may be inadequate to maintain a euthyroid status because the reduction in TSH and thyroid hormone concentrations activates the feedback mechanism. Transient or reversible forms may also occur during recovery from prolonged thyrotoxicosis or severe chronic diseases. Transplacental passage to the fetus of TSH receptor-stimulating antibodies or thyroid hormones from a thyrotoxic mother or corticosteroids or dopamine given to mothers during complicated delivery may also transiently affect neonates' suppression of TSH secretion and central hypothyroidism.

BOX 42.5 Possible Causes of Central Hypothyroidism

Invasive lesion: pituitary macroadenomas, craniopharyngiomas, meningiomas, gliomas, Rathke cleft cysts, metastases, carotid aneurysms
Iatrogenic: cranial surgery or irradiation, drugs
Injury: head traumas, traumatic delivery
Infarction: postpartum necrosis (Sheehan syndrome), pituitary apoplexy, ictus
Immunologic disease: lymphocytic hypophysitis, circulating anti-POU1F1 antibodies
Infiltrative lesion: sarcoidosis, hemochromatosis, histiocytosis X
Infectious disease: tuberculosis, syphilis, mycoses
Inherited: pituitary transcription factor defects (occasionally with childhood onset), TSHβ and TRHR mutations
Idiopathic: Other unknown causes
Empty sella syndrome

TRHR, Thyrotropin-releasing hormone (TRH) receptor (TRHR) gene; *TSHB,* thyroid-stimulating hormone-beta (TSHB) gene.
Reproduced from Persani, L., & Beck-Peccoz, P. (2013). Central hypothyroidism. In L. E. Braverman, & D. S. Cooper (Eds.), *Werner & Ingbar's the thyroid* (10th ed., p. 561). Philadelphia: Lippincott, Williams & Wilkins. Copyright Lippincott, Williams & Wilkins.

The clinical manifestations of central hypothyroidism (caused by either pituitary or hypothalamic disease) are similar to those of primary hypothyroidism but tend to be less severe (unless diagnosed in infancy). If a mass is taking up space in the cranium, headache, visual field disturbances, and hypopituitarism may develop. Nonthyroidal illness may have a similar picture as central hypothyroidism because of suppression of the hypothalamic–pituitary axis. Undiagnosed central hypothyroidism in infancy leads to cretinism.

The laboratory diagnosis is based on demonstration of low, normal, or slightly elevated TSH concentrations combined with low T4 or FT4 concentrations. In neonatal screening programs, central hypothyroidism can only be identified if TSH and T4 are both measured.

Subclinical hypothyroidism. SCH can be classified as mild (TSH between 4 and 9.9 mIU/L) or severe (TSH ≥10 mIU/L); the trimester of pregnancy and age-dependent cutoffs differ and should be applied as appropriate. The elevation in TSH should have persisted for 6 to 12 weeks or longer in the setting of FT4 concentrations that are repeatedly found within the reference interval. The prevalence increases with

TABLE 42.5 Patterns of Thyroid Dysfunction

	TSH	FT4	Comments
Primary hypothyroidism	Increased	Decreased	—
Subclinical primary hypothyroidism	Increased	Normal	—
Primary hyperthyroidism	Decreased	Increased	T3: increased
T3 toxicosis	Decreased	Normal	T3: increased
Subclinical primary hyperthyroidism	Decreased	Normal	T3: normal
Thyroid hormone resistance due to defects in the thyroid hormone receptor	Normal to increased	Increased	T3: increased; rT3: increased
Thyroid hormone resistance due to defects in thyroid hormone metabolism of T4 converted to T3	Mildly increased	Increased	T3: decreased; rT3: increased
Thyroid hormone resistance due to defects in thyroid hormone transport into cells	Normal to increased	Normal to decreased	T3: increased; rT3: decreased

T3, Triiodothyronine; *rT3,* reverse T3; *FT4,* free T4; *T4,* thyroxine; *TSH,* thyroid-stimulating hormone.

age and is higher in women than in men. Population prevalence differs; in the United States, it is estimated to be 4% to 9% in individuals without known thyroid disease.

Myxedematous coma. Patients with severe hypothyroidism may present with myxedematous (hypothyroid) coma, a syndrome of decreased consciousness, hypothermia, and other features of hypothyroidism. This condition has a high mortality and requires aggressive treatment in a critical care facility with facilities for mechanical ventilation if required.

Laboratory Diagnosis of Hypothyroidism

Serum TSH is the first-line measurement with an age-dependent reference interval usually between 0.4 and 4.2 mIU/L (see Appendix for reference intervals). *Overt hypothyroidism* is defined as increased TSH and low FT4 (Table 42.5). Autoantibodies should also be measured, indicating an autoimmune etiology. The British Thyroid Association recommends the combination of TSH and T4 (FT4 or total T4) as the first-line measurement.

Patients diagnosed with SCH (defined as increased TSH but FT4 within the reference interval) should be followed up with repeat measurements together with the measurement of TPO antibodies preferably after a 2- to 3-month interval. The presence of antibodies indicates increased risk of progression to overt hypothyroidism. Other conditions in which TSH is elevated but FT4 is normal encompass the recent institution of thyroid hormone replacement therapy (FT4 returns to normal before TSH declines), poor compliance with treatment in primary hypothyroidism, recovery from nonthyroidal illness, positively interfering heterophilic antibodies, (e.g., human antimouse antibodies) in double-antibody immunoassays, and thyroid hormone resistance.

In central hypothyroidism, the laboratory diagnosis is based on low, normal, or slightly elevated TSH combined with low total or free T4 concentrations.

Treatment

Levothyroxine sodium (thyroxine) is the treatment of choice for hypothyroidism with few adverse events if used appropriately.

Thyrotoxicosis

The term **thyrotoxicosis** refers to a condition with excess thyroid hormone. The term *hyperthyroidism* refers to a sustained increase in thyroid hormone biosynthesis and secretion by the thyroid gland. The term *thyrotoxicosis* relates to its clinical manifestations: a syndrome of hypermetabolism and hyperactivity resulting from an elevation of plasma T4 or T3 concentration (most usually both). The terms *thyrotoxicosis* and *hyperthyroidism* are not entirely synonymous. For example, thyrotoxicosis can occur as a result of excessive hormone release from the thyroid in the absence of increased synthesis, as may occur in thyroiditis. Excessive intake of thyroid hormones can also cause thyrotoxicosis but not hyperthyroidism. Box 42.6 summarizes the clinical manifestations of thyrotoxicosis.

Thyrotoxicosis can affect any physiological system in the body with the frequency and severity of signs and symptoms varying considerably among patients. Some of the causes produce characteristic clinical signs; for example, orbital and cutaneous manifestations, *which is unique to* Graves disease. The age of the patient and presence of concomitant disturbances may have an impact on the clinical features of hyperthyroidism, either exaggerating or diminishing them. For example, older patients may have less marked evidence of sympathetic activation, such as anxiety, hyperactivity, or tremor, and less weight loss but marked features of cardiovascular dysfunction such as congestive cardiac failure and atrial fibrillation.

Graves disease—the most common cause of hyperthyroidism—affects approximately 0.4% of the US population and occurs more often in women than in men (5:1). It is frequently associated with other autoimmune disorders. There is a genetic susceptibility to Graves disease, as shown by a sibling occurrence risk of 11.6% and a heritability of 75% in twin studies; the rest is environmental with smoking being a significant risk factor.

The causes of thyrotoxicosis are listed in the "Points to Remember" section. Thyrotoxicosis is usually associated with hyperthyroidism but not always. Common causes for

BOX 42.6 Clinical Manifestations of Thyrotoxicosis

Symptoms
Nervousness, stroke, agitation, or irritability
Fatigue, lethargy
Weakness
Increased perspiration
Heat intolerance
Tremor
Hyperactivity
Palpitation
Appetite change (usually increase)
Weight change (usually weight loss)
Increased bowel movement
Menstrual disturbances

Signs
Hyperactivity
Tachycardia or atrial arrhythmia
Systolic hypertension
Warm, moist, smooth skin
Stare and eyelid retraction
Tremor
Hyperreflexia
Muscle weakness
Goiter
Thyroid bruits (with Graves disease, exophthalmos, pretibial myxedema, onycholysis, thyroid acropachy)
Digital clubbing, swelling of digits and toes
Periosteal reaction at extremities of bones

thyrotoxicosis *not* associated with hyperthyroidism include (1) *thyroiditis* (silent painless thyroiditis, postpartum thyroiditis, subacute thyroiditis), and (2) *excess exogenous thyroid hormone* (iatrogenic or factitious).

The prevalence of the causes varies with iodine intake, such that in iodine-sufficient areas, Graves disease is the most common cause of thyrotoxicosis, accounting for 60% to 90% of cases. But in iodine-deficient areas, thyroidal autonomy is more common. Thyroiditis accounts for about 10% of all causes of thyrotoxicosis. Iodine fortification induces a temporary, modest increase in the incidence of hyperthyroidism in mild to moderate iodine-deficient regions.

Thyroid (Thyrotoxic) Storm

Thyroid storm (also known as thyrotoxic crisis) is a rare, severe, exaggerated, and life-threatening condition of thyrotoxicosis, which is triggered by precipitating factors (e.g., intercurrent illness or surgery). Thyroid storm is most often seen in the context of underlying Graves hyperthyroidism but can complicate thyrotoxicosis of any etiology.

Graves Disease

Graves disease results from an autoantibody that binds to and activates the TSHR, producing excessive release of thyroid hormone and clinical hyperthyroidism. In patients with Graves disease, the thyroid gland is no longer under the control of pituitary TSH but is constantly stimulated by the circulating antibodies with TSH-like activity. Both B and T lymphocytes are known to be directed at three well-characterized thyroid autoantigens, namely Tg, TPO, and TSHR. However, the majority of the evidence suggests that TSHR is the primary autoantigen of Graves disease and that the immune response to the other two thyroid antigens is reflective of the resulting thyroiditis.

Graves ophthalmopathy has an autoimmune pathogenesis, with important genetic and environmental influences, particularly smoking. Orbital muscle, connective tissue, and adipose tissue are infiltrated by lymphocytes and macrophages. The extracellular compartment of extraocular muscles and orbital fibroadipose tissue becomes edematous, secondary to the deposition of hydrophilic glycosaminoglycans.

The diagnosis of Graves disease is based on laboratory demonstration of thyrotoxicosis, clinical features (particularly Graves disease-specific extrathyroidal manifestations, including ophthalmopathy and dermopathy), and the presence of a moderate, diffuse, and soft goiter over which a vascular bruit may be detectable. Ophthalmopathy, thyroid dermopathy, and acropachy occur in 25%, 1.5%, and 0.3 % of patients with Graves disease, respectively. Patients with ophthalmopathy who are smokers have a higher risk of developing or worsening of the condition (6% in nonsmokers vs. 23% in smokers).

Treatment of Graves disease includes antithyroid drugs (thionamides), radioiodine therapy, and thyroidectomy.

Toxic Adenomas and Toxic Multinodular Goiter

Toxic adenomas and **toxic multinodular goiter** are conditions in which thyrocytes function and produce thyroid hormones independently of thyrotropin (TSH) and TSHR-stimulating antibody. This autonomous secretion of thyroid hormones leads to TSH suppression and ultimately results in thyrotoxicosis. Such thyroid autonomy is a common finding in iodine-deficient areas, where it accounts for up to 60% of cases of thyrotoxicosis. However, thyroid autonomy is rare in regions with sufficient iodine supply (3% to 10% of cases with thyrotoxicosis).

Gain-of-Function Mutations of the Thyroid-Stimulating Hormone Receptor

A familial autosomal dominant form of hyperthyroidism has been described that is caused by gain-of-function mutations in the TSH receptor. The gain-of-function mutation places the TSHR in the "on" position in the absence of ligand (TSH) binding. In infants homozygous for such mutations, neonatal thyrotoxicosis, so severe as to require emergency thyroidectomy, has been observed. Certain heterozygous mutations have been reported to cause infantile hyperthyroidism.

Central Hyperthyroidism

Central hyperthyroidism is rare but is most frequently caused by nearly always benign thyroid-stimulating hormone-secreting pituitary adenomas. A total of 75% of the tumors are macroadenomas, having a diameter larger than 10 mm at the time of diagnosis, but microadenomas (diameter <10 mm) are increasingly recognized owing to earlier

diagnosis with improved imaging techniques. Patients with TSH-secreting tumors present with signs and symptoms of thyrotoxicosis, but extrathyroidal manifestations (i.e., ophthalmopathy, pretibial **myxedema**, and acropachy) are absent. Goiter is a common finding as a consequence of chronic TSH hyperstimulation. Patients may also have symptoms related to the mass effect of the pituitary adenoma such as visual field defects, headache, or loss of other anterior pituitary function (menstrual disorders, galactorrhea, acromegaly).

The laboratory diagnosis is based on a nonsuppressed TSH in the presence of high levels of free thyroid hormones (FT3 and FT4). Other diagnostic criteria include evidence of a pituitary mass on computed tomography or magnetic resonance imaging. There are both clinical situations (in particular) and possible laboratory artifacts that may cause a biochemical profile that is characteristic of patients with TSH-secreting tumors; these include thyroid hormone resistance, binding protein abnormalities, falsely high FT4 results, and falsely high TSH concentrations. The molar ratio of α-subunit to TSH may serve as a useful tumor marker in the differential diagnosis; in case of a TSH-secreting pituitary adenoma, the ratio of α-subunit to TSH concentration is typically larger than 1 ng/mL; in thyroid hormone resistance syndrome, the ratio is less than or equal to 1 ng/mL.

Subclinical Hyperthyroidism

The laboratory diagnosis of subclinical hyperthyroidism is defined by a low plasma TSH with normal concentrations of T4 and T3. The condition is classified as mild if TSH is in the range 0.1 to 0.4 mIU/L.

Persistent subclinical hyperthyroidism may be caused by exogenous iatrogenic overdose of levothyroxine or by endogenous causes as in primary hyperthyroidism such as Graves disease, toxic multinodular goiter, or solitary autonomous nodule. Exogenous subclinical hyperthyroidism is the most common and is reversible by reduction of levothyroxine dose.

Transitory subclinical hyperthyroidism may be caused by treatment with radioiodine or antithyroid drugs in patients previously with overt hyperthyroidism or as part of thyroiditis.

Diagnosis and Differential Diagnosis of Thyrotoxicosis

The diagnosis of Graves disease may be obvious if the patient has other clinical signs, such as exophthalmos. The laboratory diagnosis of thyrotoxicosis is based on a suppressed or low TSH and increased FT4 or increased FT3 or total T3. TSH is the most sensitive biomarker. A total of 2% to 4% of patients with hyperthyroidism have increased concentration of FT3 or total T3 but normal concentration of FT4 (T3 thyrotoxicosis). However, low TSH can also be caused by other conditions relevant for the diagnosis. Subclinical hyperthyroidism is defined as low TSH with normal circulating concentrations of T4 and T3. Low TSH may also be drug induced (glucocorticoids and dopamine) or caused by nonthyroidal illness.

In pituitary adenoma (secondary hyperthyroidism), TSH can be inappropriately normal or high for an elevated concentration of peripheral thyroid hormones.

If the etiology of thyrotoxicosis is uncertain, a radioactive iodine uptake test should be performed; iodine uptake of the thyroid gland is low in thyroiditis and high in Graves disease and autonomous nodules (single or multiple).

The presence of TRAb (TSHR-stimulating antibodies) effectively confirms a diagnosis of Graves disease. The likelihood of Graves disease is 1000- to 3000-fold greater if TRAb are present. Sensitivity and specificity for third-generation TRAb assays are high. A total of 75% of patients with Graves disease have TPO antibodies, an observation that may help differentiate this disease from toxic nodular hyperthyroidism if necessary.

Due to the side-effects of antithyroid drugs, patients with Graves disease should have a baseline complete blood count, including a white blood cell (WBC) count with differential, and a liver profile, including bilirubin and transaminases, before antithyroid drug therapy is initiated.

POINTS TO REMEMBER: CAUSES OF HYPERTHYROIDISM

Endogenous Thyroid Disorders
- Autoimmune thyroid disease
  - Graves disease
  - Hashitoxicosis
  - Postpartum thyroiditis
- Toxic multinodular goiter
- Toxic adenoma
- Struma ovarii
- hCG-induced hyperthyroidism
  - Gestational hyperthyroidism
  - hCG-secreting tumors (trophoblastic tumor)
- Atopic thyroid tissue
- Secondary hyperthyroidism (pituitary tumor secreting TSH)

Exogenous Disorders
- Thyroid destruction from viral or bacterial thyroiditis (e.g., de Quervain)
- Iodine-induced hyperthyroidism (e.g., miodarone)
- Thyroid hormone ingestion (thyrotoxicosis factitia)

hCG, Human chorionic gonadotropin; *TSH,* thyroid-stimulating hormone.

THYROID DISORDERS IN PREGNANCY AND POSTPARTUM

It is estimated that approximately 4% of pregnant women have a history of thyroid disease, develop thyroid disease during pregnancy, or are diagnosed for the first time with thyroid disease 5 years after a pregnancy.

Physiological Changes

Plasma T3 and T4 concentrations increase during pregnancy owing to an increase in TBG concentration. This increase is caused by enhanced hepatic synthesis and reduced metabolism (a result of increased estrogen levels) early in pregnancy, resulting in a 1.5-fold increase in TBG by 6 to 8 weeks of gestation. TBG remains elevated throughout pregnancy.

Placental hCG shares the same α subunit with TSH but has a unique β subunit and acts in early pregnancy as a TSH agonist by binding to TSH receptors on the thyroid gland. The physiological consequences of the mild hCG stimulation of the thyroid in early pregnancy leads to a physiological rise in T4 and T3, which, by the hypothalamic-pituitary-thyroid (HPT) axis feedback mechanism, inhibits TSH secretion, which causes TSH to fall. Serum TSH decreases in the first trimester and then rises during the second and third trimesters in response to falling hCG concentrations, but does not exceed pre-pregnancy values. There is transient rise in FT4 during the first trimester owing to the relatively high circulating concentration of hCG, and FT4 gradually falls in the second and third trimesters. Changes in FT3 concentrations broadly parallel those of FT4.

Concentrations of hCG are higher in a twin pregnancy, which causes a more pronounced transient TSH suppression in the first trimester relative to a singleton pregnancy.

Hypothyroidism in Pregnancy

A total of 2% to 3% of all iodine-sufficient pregnant women have undiagnosed hypothyroidism, mostly SCH. Overt hypothyroidism is estimated to occur in 0.5% of all pregnant women. Worldwide, the most common cause is endemic iodine deficiency. The main cause of hypothyroidism in iodine-sufficient populations is chronic autoimmune thyroiditis. Two percent have isolated hypothyroxinemia (e.g., low thyroid hormone without plasma TSH elevation and without the presence of autoantibodies). About 10% to 20% of women in the childbearing years have detectable autoantibodies (TPO or Tg autoantibodies).

The diagnosis of hypothyroidism in pregnancy is based, as in nonpregnant subjects, on the finding of an elevated serum TSH concentration with low concentrations of FT4, using trimester-specific reference intervals (see Appendix). Untreated overt maternal hypothyroidism is associated with adverse maternal and fetal outcomes. There is an increased risk of miscarriage, preterm delivery, and preeclampsia in the mother. In the newborn, there is an increased risk of neonatal mortality caused by preterm delivery, risk of low-for-gestational-age birth weight, and decreased IQ. The complications are similar in SCH, but occur at a lower frequency. Also, in euthyroid pregnant women, positive for thyroid peroxidase and Tg antibodies, the risks of miscarriage, preterm delivery, and postpartum thyroiditis are increased.

Thyrotoxicosis in Pregnancy

Graves disease occurs in 0.1% to 1% of all pregnancies. Gestational transient hyperthyroidism usually occurs in the first trimester with a prevalence of 2% to 3% in Europeans.

The causes of thyrotoxicosis in pregnancy are the same as for thyrotoxicosis generally; however, gestational transient hyperthyroidism is pregnancy specific. Thyrotoxicosis may be present before pregnancy or be diagnosed in pregnancy or postpartum.

The diagnosis of thyrotoxicosis in pregnancy is made by finding a low plasma TSH concentration and elevated concentrations of FT3 and/or FT4, compared to trimester-specific reference intervals. Subclinical hyperthyroidism in pregnancy is defined as a low plasma TSH concentration with normal concentrations of FT4 or FT3. In patients with Graves disease, TSH and thyroid hormones should be measured every 4 to 6 weeks during pregnancy.

The measurement of TSH receptor stimulating antibodies (TRAb) should be reserved for patients with Graves disease who become pregnant or if Graves disease is suspected during pregnancy. In the former, TRAb should be measured at diagnosis and at 24 to 28 weeks' gestation because these antibodies can cross the placenta, starting in the late second trimester. Testing for other thyroid autoantibodies is not required, although these are typically present in high titers.

GRAVES DISEASE IN PREGNANCY

Uncontrolled, untreated Graves disease is associated with adverse pregnancy outcomes during and following pregnancy. Risks for the mother include fetal loss, preeclampsia, miscarriage, premature labor, thyroid storm, and congestive heart failure. Risks for the fetus and neonate include hyperthyroidism caused by TSH receptor stimulating antibodies (TRAb) crossing the placenta (fetal tachycardia, accelerated bone maturation, fetal goiter, intrauterine growth restriction, and signs of congestive heart failure, low birthweight for gestational age, poor Apgar scores, and respiratory distress syndrome), risk of hypothyroidism caused by treatment with antithyroid drugs, and congenital abnormalities caused by hyperthyroidism and the teratogenic effects of antithyroid drugs. Neonatal hyperthyroidism is infrequent and occurs in 1 in 50,000 neonates born to mothers with Graves disease. However, if it is not recognized and treated properly, the mortality rate can be as high at 30%.

Gestational Transient Hyperthyroidism and Hyperemesis Gravidarum

Gestational transient thyrotoxicosis is a nonautoimmune hyperthyroidism occurring in pregnant women with a spectrum ranging from no emesis, to emesis, to hyperemesis gravidarum (when dehydration could be so severe that intravenous fluid replacement may be required). Gestational transient thyrotoxicosis occurs in 2% to 3% of all pregnancies and results from activation of TSH receptors by hCG. Serum hCG concentrations are positively correlated with the severity of nausea and vomiting. The thyrotoxicosis of hyperemesis gravidarum usually resolves spontaneously within several weeks as the vomiting disappears. The degree of hyperthyroidism is typically mild.

Postpartum Thyroiditis

Postpartum thyroiditis may be difficult to differentiate from Graves disease. The differences between the two include the presence of goiter, ophthalmopathy, and TRAb in Graves disease; these are usually not present in thyroiditis. Approximately

4% to 9% of unselected postpartum women develop postpartum thyroiditis, although the incidence varies with geographical location. Postpartum thyroiditis often recurs in subsequent pregnancies, and 50% of women eventually develop hypothyroidism.

Thyroid Function Testing in Pregnancy

TSH is a reliable indicator of thyroid function during pregnancy in most cases. Trimester-specific reference intervals for TSH and thyroid hormones should be applied (see Appendix). International guidelines recommend that trimester- and assay-specific reference intervals should be established or verified in each laboratory. If that is not possible, then reference intervals from the American Thyroid Association (ATA) should be applied.

Serum FT4 and FT3 are in the picomolar range, and to be valid, they are technically difficult to measure because they must be free from interference by the much higher total hormone concentrations (in the nanomolar range). The reliability of immunoassays for the measurement of FT3 and FT4 is decreased in pregnancy owing to higher TBG but lower albumin concentrations. Liquid chromatography–tandem mass spectrometry combined with equilibrium dialysis or ultrafiltration methods are more reliable for both total and free hormone measurements during pregnancy.

THYROID NEOPLASIA

The prevalence of palpable thyroid nodules in adults is about 5%, the prevalence of nodules found on ultrasonography is 13% to 30%, and the prevalence of nodules found on autopsy is 49% to 57%. The prevalence of cancer in single nodules has been estimated to be 5% and may be less frequent in multinodular goiter. Cancer risk depends on age, sex, radiation exposure, history, family history, and other factors.

There are four main types of thyroid cancer (listed from the most common to the least common): DTC, including (1) papillary, and (2) follicular thyroid cancer, accounting for more than 90% of thyroid cancers in United States; (3) medullary thyroid cancer (MTC) (<5%); and (4) anaplastic thyroid cancer, accounting for about 2%. Parafollicular or C cells secrete calcitonin and give rise to MTC, with an intermediate prognosis. About 80% of MTC is sporadic (i.e., not genetically inherited and occurs randomly), and the rest is hereditary.

The main role for the clinical biochemist is the monitoring of TSH suppression therapy, the determination of cancer ablation, the detection of recurrence in patients given definitive treatment, such as thyroidectomy, and the prognosis.

Diagnosis

Differentiated Thyroid Cancer

Patients with DTC have a favorable prognosis. Tg is not a diagnostic marker, but it is a useful marker for disease recurrence in thyroidectomized patients because it should be undetectable. Tg should not be used as a tumor marker in the presence of anti-Tg antibodies owing to measurement interference. Thyroid stimulation either by temporarily ceasing thyroid hormone replacement or by the administration of recombinant TSH greatly increases the sensitivity of Tg measurement, but this may not be necessary using more sensitive Tg assays. The optimal cut-off for Tg for predicting recurrence is not known. There is a general agreement that a stimulated Tg less than 1 ng/mL, with no other radiologic or clinical evidence of disease and in the absence of autoantibodies, suggests no evidence of disease, although recurrence has been described below this cut point. Newer Tg assays have a functional sensitivity of 0.1 ng/mL or less (compared to older assays with 1 ng/mL); this allows for earlier identification of recurrence and helps avoid the use of TSH stimulation. As described earlier, immunometric assays for Tg and TgAbs are prone to interferences.

Medullary Thyroid Cancer

Calcitonin, produced by C cells, is a valuable blood tumor marker for MTC. Calcitonin is a 32-amino-acid monomeric peptide—the result of cleavage and the posttranslational processing of procalcitonin, which is itself a product of preprocalcitonin. Serum calcitonin can be used as a screening test in patients with a family history of MTC, who are at risk of developing the disease. The use of calcitonin as a screening marker in patients with nodular goiter and with no family history is debatable. Calcitonin also is used diagnostically as an immunohistochemical marker. Basal serum calcitonin and carcinoembryonic antigen (CEA) should be measured concurrently. CEA is not a specific biomarker for MTC and is not useful in the early diagnosis of MTC. However, CEA is useful for evaluating disease progression in patients with clinically evident MTC and for monitoring patients after thyroidectomy. All patients diagnosed with MTC should be tested for the genetic mutations associated with hereditary forms of the disease. To predict outcome and plan long-term follow-up in patients treated by thyroidectomy for MTC, tumor, node, metastasis classification, the number of lymph node metastases, and postoperative serum calcitonin and CEA should be used. Genetic status also determines the prognosis in patients with hereditary MTC.

Measurements of calcitonin also may be used to monitor for persistent or recurrent disease after surgery because the concentrations correlate with tumor burden. The MTC growth rate can be determined by measuring serum levels of calcitonin or CEA over multiple time points to determine the rate at which each marker's value doubles. Furthermore, serum TSH and serum calcium should be measured postoperatively. The goals are to get patients euthyroid and to prevent hypocalcemia.

Blood samples should be drawn in the fasting state. Calcitonin has a low stability in serum at room temperature, so the sample must be immediately separated and then frozen and transported on ice to the laboratory. The measurement of calcitonin is by immunoassay (e.g., immunochemiluminometry), which is prone to interferences. Because of inter-method differences, concentrations pre- and post-thyroidectomy and during treatment should be measured using the same assay and instrument in the same laboratory. Calcitonin values are

higher in men than in women, which is likely because of a higher C-cell mass. Calcitonin may also be high during the first week of life and in low-birthweight children and premature infants. Besides gender and age, pregnancy, lactation, and growth in children may also influence circulating concentrations of calcitonin. Calcitonin may be increased in nonthyroidal cancer, inflammation and sepsis, acute and chronic renal failure, hypercalcemia, pulmonary disease, and hypergastrinemia.

THYROID DISEASE IN CHILDREN

Thyroid disease can occur at any age; however, some characteristics in newborns, children, and adolescents are worth mentioning. The consequences of a hypofunctioning thyroid in developing and maturing children may be long lasting if not diagnosed and treated early. Neonatal hyperthyroidism may be caused by the transplacental passage of thyroid-stimulating maternal immunoglobulins (due to active maternal Graves disease) or by activating TSH receptor mutations (see subsequent text). In older children and adolescents, the most common cause of hyperthyroidism is Graves disease, and the most common cause of hypothyroidism in iodine-replete areas is Hashimoto thyroiditis.

Congenital Hypothyroidism

Before the introduction of newborn screening programs, the incidence of congenital hypothyroidism was estimated to be 1 in 7000 newborns. However, recent surveys estimate the incidence to be 1 in 2000 to 1 in 4000 newborns; the incidence is higher in Asians and Hispanics than in whites and blacks. The incidence is higher in older compared with younger mothers, and in preterm infants versus term infants. Incidence rates are dependent on TSH screening cut-offs.

Congenital hypothyroidism can be categorized into *permanent* and *transient* forms, which again can be divided into *primary* (thyroid disorder), *secondary* (i.e., central hypothyroidism; see earlier discussion), and *peripheral*. Permanent congenital hypothyroidism (75% to 86%) needs lifelong replacement treatment, but transient forms resolve within weeks to months after birth. Permanent primary congenital hypothyroidism may be caused by defects in thyroid development, thyroid dysgenesis (85%), or dyshormonogenesis (15%)—a biosynthesis defect of thyroid hormone production in a structurally normal gland. Thyroid dysgenesis may consist of either thyroid agenesis, failure of the gland to descend normally during embryologic development with or without ectopy, or hypoplasia of a normal localized gland. Central hypothyroidism occurs in 1 in 25,000 to 1 in 50,000 newborns.

Transient congenital hypothyroidism is most commonly caused by inadequate maternal iodine intake in areas of endemic iodine deficiency. It may also be caused by maternal antithyroid medication during pregnancy, transfer of maternal blocking antibodies, maternal iodine exposure (e.g., amiodarone), liver hemangiomas causing increased production of deiodinase 3, and genetic defects.

Unless born with thyroid agenesis, most newborns with congenital hypothyroidism have some thyroid function. Many newborns, even those with thyroid agenesis, do not present with classic symptoms and signs of hypothyroidism owing to transplacental passage of maternal thyroid hormones. Early symptoms and signs of congenital hypothyroidism include a lethargic infant with increased sleep, prolonged jaundice, myxedematous facies, large fontanels, macroglossia, distended abdomen, hypothermia, and hypotonia. Later symptoms and signs include poor sucking effort leading to feeding difficulties, constipation, developmental delay with cognitive and growth retardation, myxedema, umbilical hernia, and decreased activity. Ten percent of children born with congenital hypothyroidism have other congenital birth defects, and 50% of them have congenital heart defects.

Levothyroxine is the treatment of choice with treatment goals to raise serum T4 and normalize serum TSH. Levothyroxine treatment can prevent mental retardation in the majority of children (>90%) if commenced within the first 2 weeks of life.

Laboratory Diagnosis of Congenital Hypothyroidism

Newborn screening for congenital hypothyroidism is a successful public health program for secondary prevention of mental retardation. Worldwide it is estimated that 25% of the newborn population undergoes screening for congenital hypothyroidism.

Screening strategies differ between countries, with some countries measuring TSH initially with a reflex T4 if TSH is abnormal. Others measure T4 as a first-line test, and if T4 is below a certain cut-off, they reflex test for TSH or measure a combination of TSH, T4, and Tg to differentiate between primary and secondary causes. The disadvantage of screening programs only measuring TSH with a reflex T4, is the inability to detect central hypothyroidism. The initial screening occurs on the second to fifth day of life; children discharged from the hospital on the first day of life may have a sample taken at this time. Some programs routinely obtain a second specimen at 2 to 6 weeks of age. Some programs use cord blood at birth, and others use a heel prick on filter paper after birth. Many programs now use the initial TSH measurement from heel prick blood on filter cards. Filter cards are mailed to a central laboratory. Each program has its own cut-offs for test results. Thyroid hormone levels and TSH are higher in the first days of life but have usually fallen to concentrations typically seen in infancy within the 2 to 4 weeks.

An abnormal result on screening should lead to a confirmatory test in a serum sample, but this should not delay treatment. Confirmatory testing includes TSH and free or total T4, and should be compared with appropriate age-dependent reference intervals (see Appendix for reference intervals). Further diagnostic tests may include radionuclide thyroid uptake and scanning, thyroid sonography, and serum Tg to determine the subtype of congenital hypothyroidism, but again these investigations should not delay the initiation of treatment.

False-positive elevations in TSH within the first 2 days of life may be revealed after repeated confirmatory testing. Transplacental transfer of heterophile antibodies is well described as a false-positive interference in blood spot TSH, and maternal thyroid function tests need to be checked in this context. Preterm infants with an immature HPT axis and acutely ill term infants may have a late rise in TSH and may not show elevated TSH on the first screening test; many programs have a second screening test for these babies. Dopamine used in the treatment of ill premature neonates can also attenuate TSH release. Seasonal variations in TSH occur with an increased false-positive rate of congenital hypothyroidism in the winter (0.9%) compared with the summer (0.6%); this is in accordance with globally conducted previous studies that have identified an increased prevalence of suspected and confirmed cases of congenital hypothyroidism in the winter months.

GENETICS

Evidence from twin studies has shown that about 65% of baseline TSH and thyroid hormone concentrations are genetically determined, with about 20% of the variability coming from common genetic variation at the population level. This suggests a genetic basis for narrow intraindividual variation in these hormone concentrations. Several loci have been identified in genome-wide association studies (GWAS) for the circulating concentrations of TSH, thyroid hormones, thyroid autoantibodies, and deiodinases.

Autoimmune susceptibility against the thyroid gland is estimated to affect 5% of the general population, and about 80% of this susceptibility can be explained by genetic factors.

Genetics in Congenital Hypothyroidism

Thyroid dyshormonogenesis is inherited in an autosomal recessive pattern. Thyroid dysgenesis, on the other hand, is inherited in only approximately 2% of cases; the rest are considered to be sporadic.

Thyroid-Stimulating Receptor Mutations and Resistance to Thyroid-Stimulating Hormone

Cases of familial thyrotoxicosis with no evidence of autoimmunity and children with persistent isolated neonatal hyperthyroidism should be evaluated for familial nonautoimmune autosomal dominant hyperthyroidism (FNAH) and persistent sporadic congenital nonautoimmune hyperthyroidism (PSNAH) caused by rare germline mutations in the *TSHR* gene.

Thyroid Hormone Receptor Mutations and Resistance to Thyroid Hormone

Two different thyroid hormone receptors are known; thyroid hormone receptor α (THRα) and β (THRβ), which are encoded by the *THRA* and *THRB* genes, respectively. Patients with mutations in *THRB* present with resistance to thyroid hormone β (RTHβ) characterized by raised levels of thyroid hormone, normal or elevated levels of TSH, and goiter, suggesting a critical role for *THRB* in negative-feedback regulation.

Only a few patients with resistance to thyroid hormone α (RTHα) have been described.

REVIEW QUESTIONS

1. What is the most common type of laboratory assay that is widely used to assess thyroid hormone concentrations?
   a. Ultrafiltration
   b. Immunoassay
   c. Potentiometry
   d. Estimation from using a formula
2. What stimulates the uptake of iodide by the thyroid gland for thyroid hormone synthesis?
   a. Thyroid-stimulating hormone (TSH)
   b. Thyroxine (T4)
   c. Tyrosine
   d. Thyroiditis
3. What causes primary hypothyroidism?
   a. The absence or dysfunction of the thyroid gland
   b. Increased TSH
   c. A pituitary disorder
   d. A hypothalamus disorder
4. Hyperthyroidism is also referred to as
   a. Athyreosis
   b. Myxedema
   c. Thyrotoxicosis
   d. Exophthalmos
5. What is the function of thyroid hormones?
   a. Inhibit the secretion of growth hormone
   b. Solely regulate reproductive processes in male and females
   c. Maintain water homeostasis
   d. Regulate carbohydrate, lipid, and protein metabolism within cells
6. *Secondary (or central) hypothyroidism* is indicated by which of the following sets of lab results?
   a. Increased T4, and decreased TSH
   b. Increased T4, and increased TSH
   c. Decreased T4, and increased TSH
   d. Decreased T4, and decreased TSH
7. The thyroid gland produces all of the following hormones *except* which of the following hormones?
   a. TSH
   b. Calcitonin
   c. Thyroxine (T4)
   d. Triiodothyronine (T3)
8. What is the amino acid precursor of T4?
   a. Threonine
   b. Tyrosine
   c. Thyronine

d. Alanine

9. What is the technique of choice to measure T3 in body fluids?
a. Nephelometry
b. Potentiometry
c. Immunoassay
d. Atomic absorption

10. Which of the following analytes is most sensitive for the early detection of primary thyroid gland failure?
a. Thyroxine (T4)
b. TSH
c. Thyroglobulin (Tg)
d. Thyroid-binding globulin (TBO)

SUGGESTED READINGS

Bailey, D., Colantonio, D., Kyriakopoulou, L., et al. (2013). Marked biological variance in endocrine and biochemical markers in childhood: establishment of pediatric reference intervals using healthy community children from the CALIPER cohort. *Clinical Chemistry*, *59*, 1393–1405.

Cooper, D. S., & Laurberg, P. (2013). Hyperthyroidism in pregnancy. *Lancet Diabetes Endocrinol*, *1*, 238–249.

Cooper, D. S., & Biondi, B. (2012). Subclinical thyroid disease. *Lancet*, *379*, 1142–1154.

Elmlinger, M. W., Kuhnel, W., Lambrecht, H. G., et al. Reference intervals from birth to adulthood for serum thyroxine (T4), triiodothyronine (T3), free T3, free T4, thyroxine binding globulin (TBG) and thyrotropin (TSH). *Clinical Chemistry and Laboratory Medicine, 39*, 973–979.

Franklyn, J. A., & Boelaert, K. (2012). Thyrotoxicosis. *Lancet*, *379*, 1155–1166.

Gruters, A., Krude H. Detection and treatment of congenital hypothyroidism. *Nature Reviews Endocrinology*, *8*, 104–113.

Hoofnagle, A. N., Becker, J. O., Wener, M. H., et al. Quantification of thyroglobulin, a low-abundance serum protein, by immunoaffinity peptide enrichment and tandem mass spectrometry. *Clinical Chemistry*, *54*, 1796–1804.

Midgley, J. E. (2001). Direct and indirect free thyroxine assay methods: theory and practice. *Clinical Chemistry*, *47*, 1353–1363.

Pearce, E. N., Farwell, A. P., & Braverman, L. E. (2003). Thyroiditis. *New England Journal of Medicine*, *348*, 2646–2655.

Roberts, C. G., & Ladenson, P. W. (2004). Hypothyroidism. *Lancet*, *363*, 793–803.

Stagnaro-Green, A., & Pearce, E. (2012). Thyroid disorders in pregnancy. *Nature Reviews Endocrinology*, *8*, 650–658.

Taylor, P. N., Porcu, E., Chew, S., et al. (2015). Whole-genome sequence-based analysis of thyroid function. *Nature Communications*, *6*, 5681.

Thienpont, L. M., Van Uytfanghe, K., Beastall, G., et al. (2010). Report of the IFCC working group for standardization of thyroid function tests; part 1: thyroid-stimulating hormone. *Clinical Chemistry*, *56*, 902–911.

Thienpont, L. M., Van Uytfanghe, K., Beastall, G., et al. (2010). Report of the IFCC Working Group for Standardization of Thyroid Function Tests; part 3: total thyroxine and total triiodothyronine. *Clinical Chemistry*, *56*, 921–929.

Thienpont, L. M., Van Uytfanghe, K., & Van Houcke, S. (2010). Tests IWGfSoTF. Standardization activities in the field of thyroid function tests: a status report. *Clinical Chemistry and Laboratory Medicine*, *48*, 1577–1583.

Thienpont, L. M., Van Uytfanghe, K., Van Houcke, S., et al. (2014). A progress report of the IFCC committee for standardization of thyroid function tests. *European Thyroid Journal*, *3*, 109–116.

Vestergaard, P., Rejnmark, L., Weeke, J., et al. (2002). Smoking as a risk factor for Graves' disease, toxic nodular goiter, and autoimmune hypothyroidism. *Thyroid*, *12*, 69–75.

Zimmermann, M. B., & Boelaert, K. (2015). Iodine deficiency and thyroid disorders. *The Lancet Diabetes Endocrinology*, *3*, 286–295.

43

Reproduction-Related Disorders

Robert D. Nerenz, Emily Jungheim, Ann Gronowski

OBJECTIVES

1. Identify hormonal factors that determine male and female reproductive function.
2. Describe clinical symptoms associated with states of reproductive dysfunction.
3. Explain pathologic mechanisms responsible for infertility.
4. Describe laboratory tests used in the evaluation of reproductive disorders.

KEY WORDS AND DEFINITIONS

Amenorrhea The absence of menstruation.
Dehydroepiandrosterone (DHEA) The most abundant adrenal androgen.
Estradiol The predominant female sex steroid.
Follicle A pouch-like sac on the surface of the ovary that contains the maturing ovum (egg).
Follicle-stimulating hormone (FSH) A heterodimeric glycoprotein secreted by the pituitary gland that stimulates ovarian follicle maturation and sperm production.
Gonad A gamete-producing gland.
Gonadotropin-releasing hormone (GnRH) A hypothalamic polypeptide hormone that stimulates LH and FSH release by the pituitary gland.
Hypothalamus The portion of the brain just above the pituitary gland that regulates its endocrine functions.
Luteinizing hormone (LH) A heterodimeric glycoprotein secreted by the pituitary gland that stimulates gonadal sex steroid production and triggers ovulation.
Menopause Cessation of menstruation, which usually occurs around the age of 50.
Menses The monthly flow of blood from the female genital tract.
Pituitary The master regulatory endocrine gland located at the base of the brain.
Polycystic ovary syndrome (PCOS) A female condition that is characterized by multiple ovarian follicles, ovulatory dysfunction, and increased androgen production.
Progesterone A steroid hormone secreted by the corpus luteum and placenta with a primary role in maintaining pregnancy.
Ovary Female gonad.
Ovulation The process of ovarian egg release.
Sex hormone–binding globulin A circulating plasma protein that binds estradiol, testosterone, and dihydrotestosterone.
Testis Male gonad.
Testosterone The predominant male sex steroid.
Virilization The induction or development of male secondary sex characteristics; especially the induction of such changes in females.

BACKGROUND

The field of reproductive endocrinology encompasses the hormones of the hypothalamic-pituitary-gonadal axis and the adrenal glands; these hormones are crucial for reproductive function. Hypothalamic **gonadotropin-releasing hormone (GnRH)** directs the **pituitary** to synthesize and release **follicle-stimulating hormone (FSH)** and **luteinizing hormone (LH)**, which in turn stimulate gonadal synthesis of the sex steroids that govern the development and maintenance of secondary sex characteristics. In states of reproductive health, serum concentrations of these hormones rise and fall in a tightly regulated and well-characterized pattern. In states of reproductive dysfunction, measurement of these hormones in the clinical laboratory often provides the necessary information to identify the underlying abnormality and guide appropriate treatment.

MALE REPRODUCTIVE BIOLOGY

The mature **testes** synthesize both sperm and androgens. The testes contain a structured network of tightly packed seminiferous tubules, the lumina of which are lined by maturing germ cells and Sertoli cells. Sertoli cells play a crucial role in sperm maturation and secrete *inhibin,* a 32-kDa glycoprotein

that inhibits the pituitary secretion of FSH. Surrounding the seminiferous tubules are the interstitial Leydig cells, the primary site of androgen production. The principal androgen in humans is **testosterone**, which serves a central role in reproductive physiology. Testosterone is required for sexual differentiation, spermatogenesis, and the promotion and maintenance of sexual maturity at puberty. At the cellular level, these effects are mediated by binding of testosterone or its more potent metabolite dihydrotestosterone (DHT) to the androgen receptor or via aromatization to estradiol and subsequent binding to the estrogen receptor. Testicular function is under the control of the hypothalamic-pituitary-gonadal axis.

HYPOTHALAMIC-PITUITARY-GONADAL AXIS

GnRH is a decapeptide synthesized in the **hypothalamus** and transported to the anterior pituitary gland, where it stimulates the release of both FSH and LH.

In adult men, GnRH and thus LH and FSH are secreted in pulsatile patterns. A circadian rhythm is present, with higher concentrations found in the early-morning hours and lower concentrations in the late evening. LH acts on Leydig cells to stimulate the conversion of cholesterol to pregnenolone. FSH acts on Sertoli cells and spermatocytes and is central to the initiation (in puberty) and maintenance (in adulthood) of spermatogenesis. Sex steroids and inhibin provide negative feedback control of LH and FSH secretion, respectively.

Androgens

Androgens are a group of C-19 steroids (Fig. 43.1) responsible for masculinization of the genital tract and development and maintenance of male secondary sex characteristics. Testosterone is the principal androgen secreted in men.

Biosynthesis of Testosterone

Testosterone is synthesized primarily by the Leydig cells of the testes (95%) and, to a lesser extent (≈5%), via peripheral conversion from the precursors **dehydroepiandrosterone (DHEA)** and androstenedione, which are synthesized in the zona reticularis of the adrenal glands (see Fig. 43.1).

Androgen Transport in Blood

Testosterone and DHT circulate in plasma freely (≈2% to 3%) or bound to plasma proteins. Binding proteins include the specific **sex hormone–binding globulin** (SHBG) and nonspecific proteins such as albumin. SHBG is an α-globulin that has low capacity for steroids but binds with very high affinity ($K_a = 1 \times 10^8$ to 1×10^9), whereas albumin has high capacity but low affinity ($K_a = 1 \times 10^4$ to 1×10^6). SHBG has the highest affinity for DHT and the lowest for estradiol.

The biologically active fraction includes free testosterone; some have suggested that albumin-bound testosterone may also be available for tissue uptake. Therefore the bioavailable testosterone is equal to approximately 35% of the total quantity (free + albumin-bound).

Testosterone and SHBG exhibit rhythmic variation in their circulating concentrations. Testosterone concentrations peak at approximately 0400 to 0800 hours, and nadir concentrations occur at between 1600 and 2000 hours. Daily variations in SHBG concentrations are similar to those of other proteins and albumin in serum, with major changes related to posture. Concentrations of SHBG are elevated with hyperthyroidism and in hypogonadal men.

Metabolism of Testosterone

Circulating testosterone serves as a precursor for the formation of two additional active metabolites: DHT and estradiol. In one pathway, 5α-reductase converts 6% to 8% of testosterone to DHT. Both testosterone and DHT bind the androgen receptor, but DHT binds with higher affinity. In an alternative pathway, testosterone and androstenedione are converted to estrogens (≈0.3%) through aromatase (CYP19). DHT is formed in androgen target tissues such as the skin and prostate, whereas aromatization occurs in many tissues, especially the liver and adipose tissue. Peripheral aromatization occurs primarily in adipose tissue (of both men and women) because of the high concentration of aromatase in this tissue. The rate of extraglandular aromatization therefore increases with body fat.

The main excretory metabolites of androstenedione, testosterone, and DHEA are shown in Fig. 43.2. Except for epitestosterone, these catabolites constitute a group of steroids known as *17-ketosteroids* (17-KS); they are excreted primarily in the urine.

Testosterone Concentrations

Testosterone is required for proper sexual development and function throughout all stages of life: fetal, pubertal, and adult (Fig. 43.3). Fetal testes produce testosterone around the seventh week of gestation, with peak serum concentrations of approximately 250 ng/dL (8.7 nmol/L) observed at the beginning of the second trimester, and with concentrations gradually returning to baseline by birth. Shortly after birth, the concentration of testosterone begins to increase, peaking again at approximately 250 ng/dL (8.7 nmol/L) at 2 to 3 months of age, and then falling to baseline again by 6 to 12 months. The concentration of testosterone remains low (<50 ng/dL, 1.7 nmol/L) until puberty, when the concentration of testosterone rises to 500 to 700 ng/dL (17.3 to 24.3 nmol/L). Testosterone remains elevated through adulthood until around the third to fourth decade.

Men beyond 30 to 40 years of age experience an age-dependent decrease in circulating testosterone concentration, estimated at 0.5% to 2% per year from about the fourth decade onward. This decline in testosterone is thought to be due to (1) a decrease in Leydig cell numbers, (2) decreased GnRH pulse amplitude, and (3) increases in SHBG. In the past, these decreases in circulating concentrations of testosterone were viewed as a normal part of the aging process. Now, however, these decreases, when accompanied by symptoms of decreased libido, sexual dysfunction, decreased energy levels, and decreased muscle mass, are regarded as a syndrome

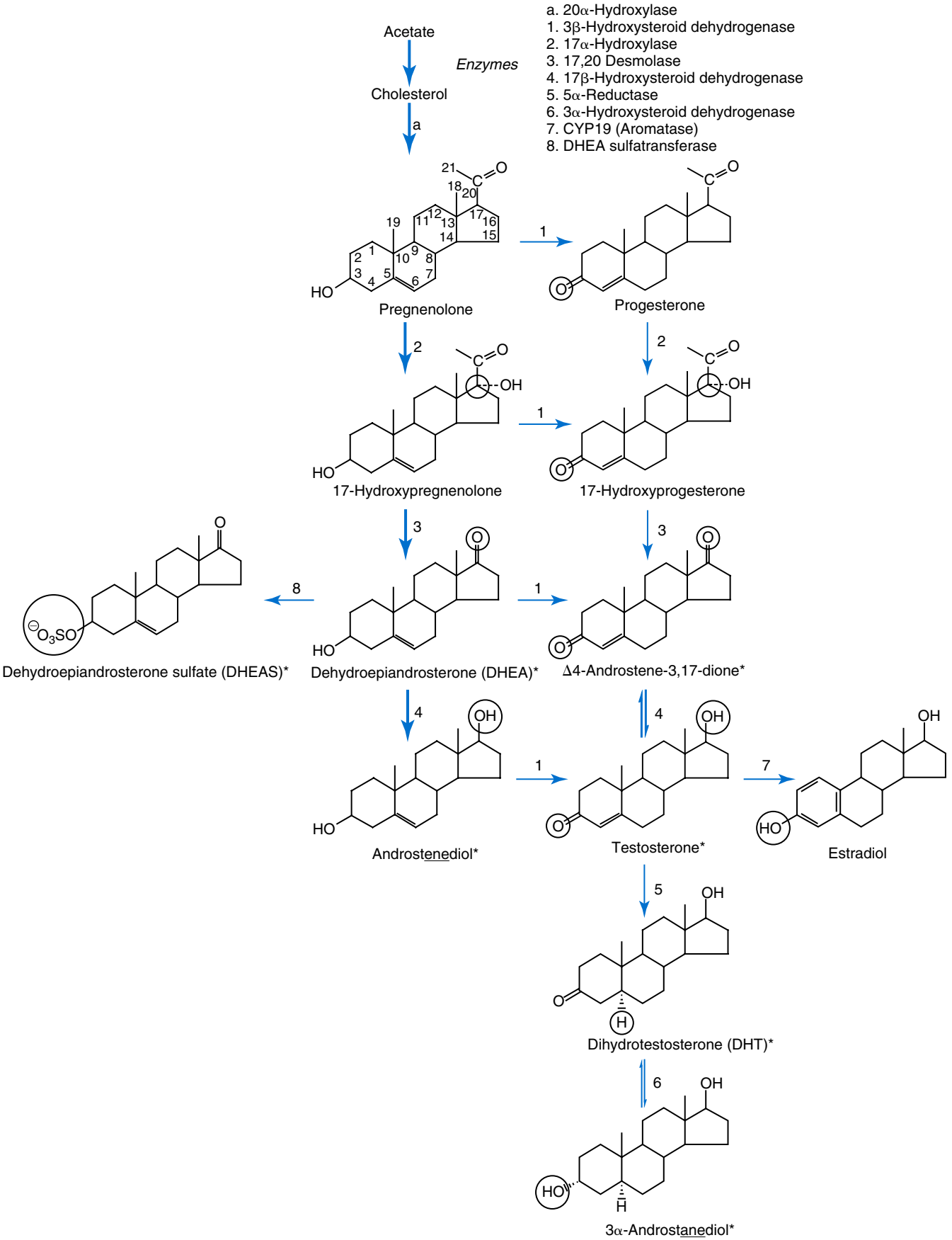


Fig. 43.1 Biosynthesis of androgens (adrenal glands and testis). The heavy *arrows* indicate the preferred pathway. The *circled area* represents the site of chemical change. *Denotes androgens.

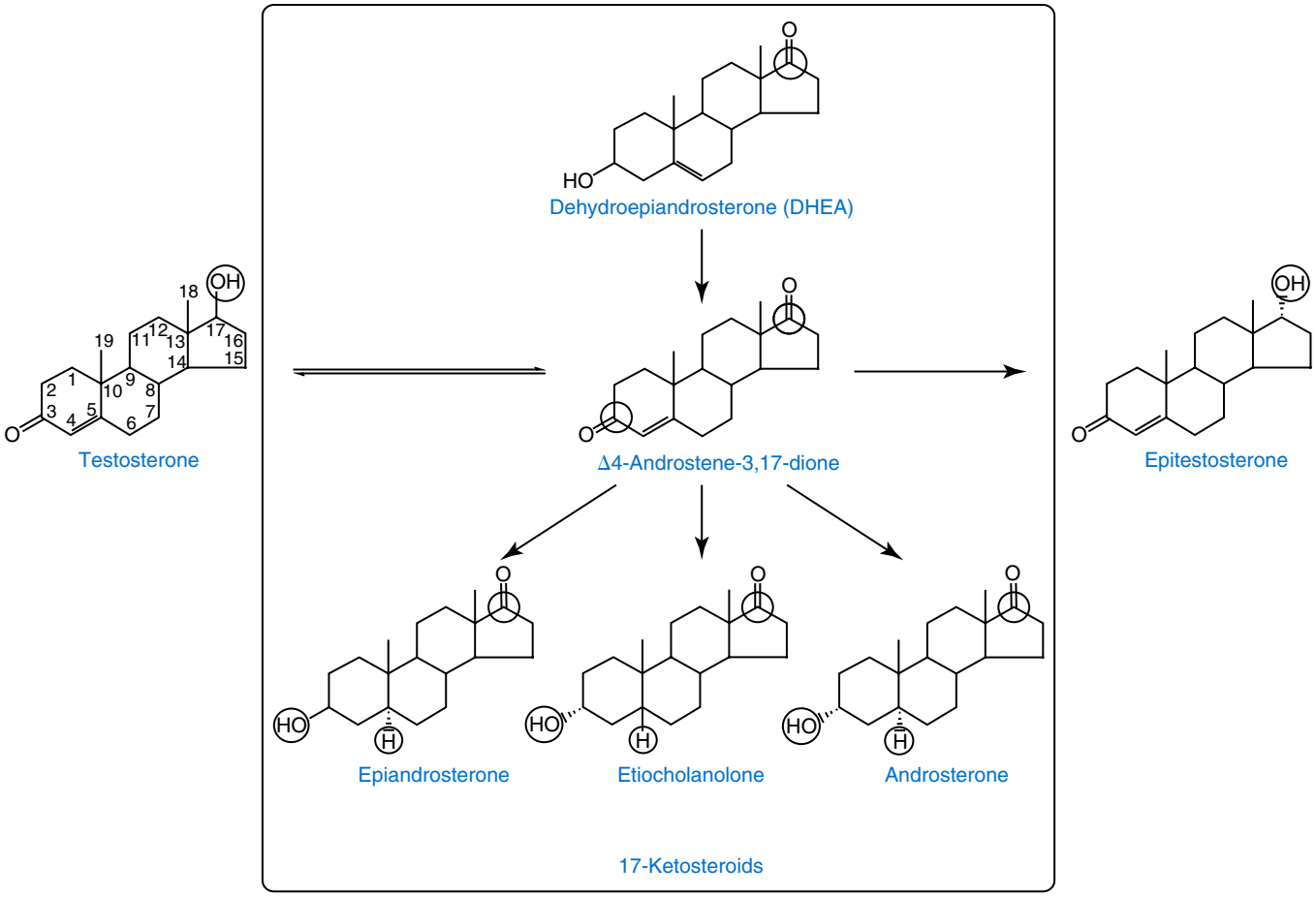


Fig. 43.2 Catabolism of C19O2 androgens. The *circled area* represents the site of chemical change.

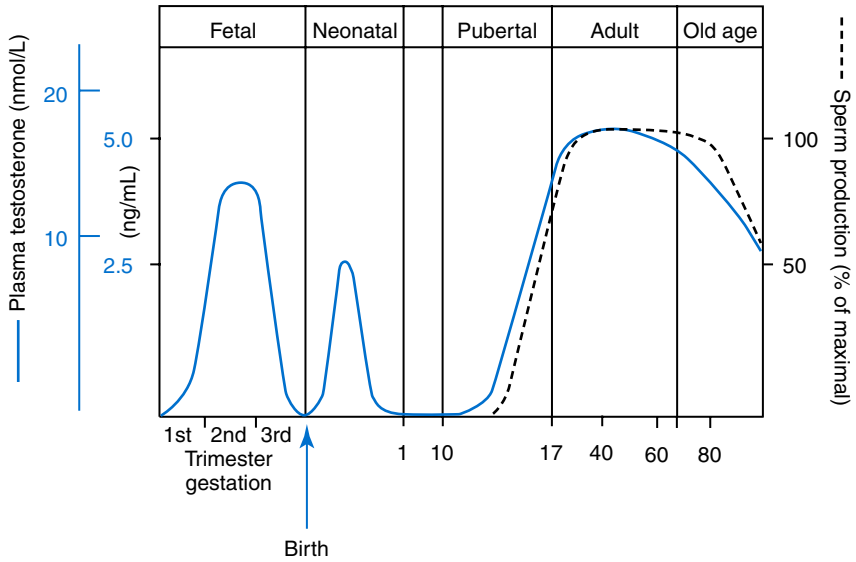


Fig. 43.3 Schematic diagram of different phases of male sexual function during life as indicated by mean plasma testosterone concentration and sperm production at different ages. (From Griffin, J. E., & Wilson, J. D. [1980]. The testis. In P. K. Bondy, & L. E. Rosenberg [Eds.], *Metabolic control and disease* [8th ed., pp. 1535–1578]. Philadelphia: WB Saunders.)

with a variety of names, such as androgen deficiency in the aging male (ADAM), partial androgen deficiency of the aging male (PADAM), late-onset hypogonadism (LOH), and, erroneously, andropause. The name *andropause* is inaccurate and misleading, given that in contrast to menopause in women, concentrations of sex steroids in men do not decrease sharply with secondary cessation of reproductive function. A name put forward more recently is *testosterone deficiency syndrome (TDS),* highlighting a specific deficit in testosterone as part of the clinical picture.

Diagnosis of LOH (TDS) should be based on both clinical and laboratory assessment. Clinically, patients should exhibit symptoms suggestive of testosterone deficiency, such as decreased libido, erectile dysfunction, decreased muscle mass and strength, decreased bone mineral density, and changes in mood. Patients should exhibit one to three of these symptoms with a concomitant low concentration of serum testosterone to fit various diagnostic criteria. Total serum testosterone is the most widely used biochemical parameter for assessment of hypogonadism, but measurement of free or bioavailable testosterone should be considered when total testosterone is not diagnostic despite the clinical presentation of hypogonadism.

Male Reproductive Abnormalities

A wide variety of abnormalities can affect the male reproductive system before birth, in childhood, or in adulthood. For the purposes of this chapter, they have been divided into categories of (1) hypogonadotropic hypogonadism, (2) hypergonadotropic hypogonadism, (3) defects in androgen action (Box 43.1), (4) erectile dysfunction, and (5) gynecomastia.

Hypogonadotropic Hypogonadism

Hypogonadotropic hypogonadism occurs when defects in the hypothalamus or pituitary prevent normal gonadal stimulation. Causative factors include congenital or acquired panhypopituitarism, hypothalamic syndromes, GnRH deficiency, hyperprolactinemia, malnutrition or anorexia, and iatrogenic causes. All of these abnormalities are associated with decreased testosterone and gonadotropin concentrations.

Kallmann syndrome, the most common form of hypogonadotropic hypogonadism, results from a deficiency of GnRH in the hypothalamus during embryonic development. It is characterized by hypogonadism and anosmia (loss of the sense of smell) in male or female patients and is inherited as an autosomal dominant trait with variable penetrance.

Hypergonadotropic Hypogonadism

Hypergonadotropic hypogonadism results from a primary gonadal disorder. Patients with primary testicular failure have increased concentrations of LH and FSH and decreased concentrations of testosterone. Causes of primary hypogonadism are categorized as (1) acquired causes (irradiation, castration, mumps orchitis, or cytotoxic drugs), (2) chromosomal defects (Klinefelter syndrome), (3) defective androgen synthesis (20α-hydroxylase deficiency), (4) testicular agenesis, (5) seminiferous tubular disease, and (6) other miscellaneous causes. Aging is associated with gonadal failure, specifically, decreased Leydig cell mass and reserve capacity with reduction in pulsatile secretion of GnRH by the hypothalamus, leading to decreased testosterone secretion.

BOX 43.1 Male Reproductive Abnormalities

Hypogonadotropic Hypogonadism
- Panhypopituitarism (congenital or acquired)
- Hypothalamic syndrome (acquired or congenital)
  - Structural defects (neoplastic, inflammatory, and infiltrative)
  - Prader-Willi syndrome
  - Laurence-Moon-Biedl syndrome
  - GnRH deficiency (Kallmann syndrome)
  - Hyperprolactinemia (prolactinoma or drugs)
  - Malnutrition and anorexia nervosa
  - Drug-induced suppression of luteinizing hormone (androgens, estrogens, tranquilizers, antidepressants, antihypertensives, barbiturates, cimetidine, GnRH analogs, and opiates)

Hypergonadotropic Hypogonadism
- Acquired (irradiation, mumps orchitis, castration, and cytotoxic drugs)
- Chromosome defects
  - Klinefelter syndrome (47,XXY) and mosaics
  - Autosomal and sex chromosomes, polyploidies
  - True hermaphroditism
- Defective androgen biosynthesis
  - 20α-Hydroxylase (cholesterol 20,22-desmolase) deficiency
  - 17,20-Lyase deficiency
  - 3β-Hydroxysteroid dehydrogenase deficiency
  - 17α-Hydroxylase deficiency
  - 17β-Hydroxysteroid dehydrogenase deficiency
- Testicular agenesis
- Selective seminiferous tubular disease
- Miscellaneous
  - Noonan syndrome (short stature, pulmonary valve stenosis, hypertelorism, and ptosis)
  - Streak gonads
  - Myotonia dystrophica
- Acute and chronic disease

Defects in Androgen Action
- Complete androgen insensitivity *(testicular feminization)*
- Partial androgen sensitivity
  - 5α-Reductase deficiency

GnRH, Gonadotropin-releasing hormone.

Defects in Androgen Action

The most common and severe defect in androgen action is *androgen insensitivity syndrome* (AIS), a disorder arising from mutations in the androgen receptor gene *(AR)*. AIS may be classified as complete (CAIS) or partial (PAIS), depending on the amount of residual receptor function. Individuals with complete AIS (formerly known as testicular feminization) have a male karyotype (46,XY) with female external genitalia (labia, clitoris, and vaginal opening). The testes are present intra-abdominally,

and because they produce anti-Müllerian hormone (AMH), no uterus, fallopian tubules, or proximal vagina is present. The circulating concentration of testosterone in these patients is greater than or equal to that of a healthy male. Concentrations of LH are increased, presumably because of resistance of the hypothalamic-pituitary system to androgen inhibition.

Males with *5α-reductase deficiency* (5-ARD) do not convert testosterone to the more potent DHT. Because DHT leads to masculinization of external genitalia in utero, males are born with ambiguous genitalia. High ratios of the circulating concentrations of testosterone to DHT are indicative of 5-ARD. Moreover, evidence indicates that DHT formation is deficient in the tissues of the urogenital tract in these patients.

In patients with cryptorchidism or ambiguous genitalia, identification of abdominal gonads is essential for proper diagnosis and treatment. The presence of testicular tissue has traditionally been detected by the measurement of Leydig cell testosterone production after stimulation with human chorionic gonadotropin (hCG). A growing appreciation of the assessment of Sertoli cell function has been noted. Inhibin and AMH reflect Sertoli cell function and may offer a noninvasive evaluation of the integrity of the seminiferous tubules.

Erectile Dysfunction

Erectile dysfunction (formerly referred to as *impotence*) is the persistent inability to develop or maintain a penile erection that is sufficient for intercourse and ejaculation in 50% or more of attempts. A wide variety of organic and psychologic abnormalities may cause changes in sexual drive and in the ability to have an erection or to ejaculate. Psychogenic erectile dysfunction is the most common diagnosis. Other causes include vascular disease, diabetes mellitus, hypertension, uremia, neurologic disease, hypogonadism, hyperthyroidism and hypothyroidism, neoplasms, and drugs. If no obvious explanation for erectile dysfunction can be found, measurement of morning serum testosterone, LH, and thyroid-stimulating hormone (TSH) concentrations has been suggested. Elevated gonadotropin concentrations indicate primary hypogonadism.

Gynecomastia

Gynecomastia, the benign growth of glandular breast tissue in men, is a common finding among males of various ages. This condition, which is related to an increase in the estrogen/androgen ratio, is commonly associated with three distinct periods of life. First, transient gynecomastia can be found in 60% to 90% of all newborns because of the high estrogen concentrations that cross the placenta. The second peak occurs during puberty in 50% to 70% of normal boys. It is usually self-limited and may be due to low serum testosterone, low DHT, or a high estrogen/androgen ratio. The last peak is found in the adult population, most frequently among men aged 50 to 80 years. Gynecomastia may be due to testicular failure, resulting in an increased estrogen/androgen ratio, or to increased body fat, resulting in increased peripheral aromatization of testosterone to estradiol. It is important to note that prolactin plays an important role in *galactorrhea* (milk production) but only an indirect role in gynecomastia.

POINTS TO REMEMBER

- GnRH stimulates LH and FSH release, which increase testosterone and sperm production, respectively.
- LH secretion is inhibited by testosterone and FSH secretion is inhibited by inhibin.
- Testosterone is transported in blood tightly bound to SHBG and loosely bound to albumin.
- Free testosterone is biologically active and represents 2% to 3% of total testosterone.
- Pituitary defects result in hypogonadotropic hypogonadism.
- Primary gonadal defects result in hypergonadotropic hypogonadism.

FEMALE REPRODUCTIVE BIOLOGY

The ovaries produce ova and secrete the sex hormones progesterone and estrogen. Every healthy female neonate possesses approximately 400,000 primordial follicles, each containing an immature ovum. During the reproductive life span of an adult woman, 300 to 400 follicles will reach maturity. A single mature **follicle** is produced during each normal menstrual cycle at approximately day 14. Surrounding the oocyte of the mature follicle are three distinct cell layers: *theca externa, theca interna,* and *granulosa cells.* The theca interna cells are the primary source of androgens, which are transported to adjacent granulosa cells, where they are aromatized to estrogens.

The mature follicle undergoes ovulation by the process of rupture, thereby releasing the oocyte into the proximity of the fallopian tubes. The follicle then fills with blood to form the corpus hemorrhagicum. The granulosa and theca cells of the follicle lining quickly proliferate to form lipid-rich luteal cells, replacing the clotted blood and forming the *corpus luteum* (yellow body). The luteal cells produce estrogen and progesterone. If fertilization and pregnancy occur, the corpus luteum persists and continues to produce estrogen and progesterone. If no pregnancy occurs, the corpus luteum regresses, and the next menstrual cycle begins.

The uterine cavity is lined by the endometrium. The endometrium undergoes cyclic changes in preparation for implantation and pregnancy in response to cyclic changes in estrogen and progesterone. During the follicular phase, the endometrial lining increases in thickness and vascularity in response to increasing circulating concentrations of estrogen; after regression of the corpus luteum, menstruation begins, and the endometrium is shed in response to the withdrawal of progesterone.

Hypothalamic-Pituitary-Gonadal Axis

In adult women, a tightly coordinated feedback system exists between the hypothalamus, anterior pituitary, and ovaries to orchestrate menstruation. FSH serves to stimulate follicular growth, and LH stimulates ovulation and progesterone secretion from the developing corpus luteum.

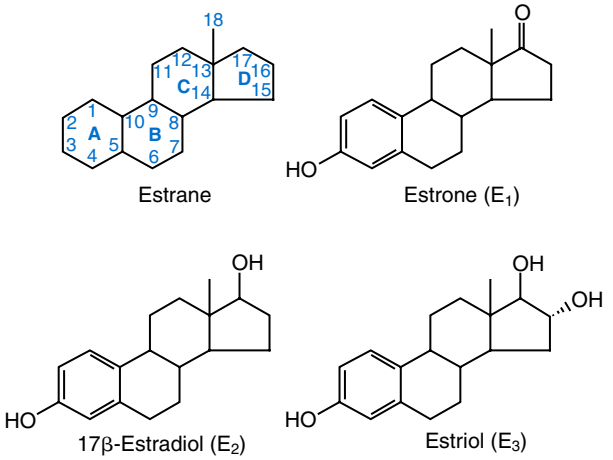


Fig. 43.4 Biologically active estrogens.

Estrogens

Estrogens are responsible for the development and maintenance of female sex organs and female secondary sex characteristics. In conjunction with progesterone, they participate in regulation of the menstrual cycle and of breast and uterine growth as well as in the maintenance of pregnancy.

Estrogens affect calcium homeostasis and have a beneficial effect on bone mass. They decrease bone resorption and accelerate linear bone growth in prepubertal girls, resulting in epiphyseal closure. Long-term estrogen depletion is associated with loss of bone mineral content, an increase in stress fractures, and postmenopausal osteoporosis.

Estrogens also have well-established effects on plasma proteins that influence endocrine testing. They increase concentrations of SHBG, corticosteroid-binding globulin, and thyroxine-binding globulin. Hence boys and girls have comparable concentrations of SHBG, but adult men have SHBG concentrations that are about half those of adult women. Concentrations of plasma proteins that bind copper and iron are also elevated in response to estrogen, as are those of high-density and very high-density lipoproteins. In addition, estrogens are believed to play a preventive role in coronary heart disease.

Chemistry

The three most biologically active estrogens in order of potency are estradiol (E_2), estrone (E_1), and estriol (E_3) (Fig. 43.4).

Biosynthesis

The biochemical pathway illustrating aromatization of testosterone to estradiol and androstenedione to estrone is shown in Fig. 43.5.

Estrogens are secreted primarily in healthy women by the ovarian follicles and the corpus luteum and during pregnancy by the placenta. The adrenal glands and testes (in men) are also believed to secrete minute quantities of estrogens. The **ovary** synthesizes estrogens via aromatization of androgens. Synthesis of estrogens begins in the theca interna cells with the enzymatic synthesis of androstenedione from cholesterol. Androstenedione is then transported to the granulosa cells, where it is further metabolized directly to estrone (androstenedione → estrone), or first to testosterone and then to estradiol (androstenedione → testosterone → estradiol). These conversions are catalyzed by the enzyme aromatase. The healthy human ovary produces all three classes of sex steroids: estrogens, progestagens, and androgens; however, estradiol and progesterone are its primary secretory products. The most potent estrogen secreted by the ovary is 17β-estradiol. Because it is derived almost exclusively from the ovaries, its measurement is often considered sufficient for the evaluation of ovarian function.

Estrogens are also produced by the peripheral aromatization of androgens, primarily androstenedione. In healthy men and women, approximately 1% of secreted androstenedione is converted to estrone. Although the ovaries of postmenopausal women do not secrete estrogens, these women have significant blood concentrations of estrone originating from the peripheral conversion of adrenal androstenedione. Because a major site of this conversion is adipose tissue, estrone is increased in obese postmenopausal women, sometimes yielding enough estrogen to produce bleeding.

Biosynthesis During Pregnancy

Estrogen biosynthesis differs qualitatively and quantitatively in pregnant women compared with nonpregnant ones. In pregnant women the major source of estrogens is the placenta, whereas in nonpregnant women the ovaries are the main site of synthesis. In contrast to the microgram quantities secreted by nonpregnant women, the quantity of estrogens secreted during pregnancy increases to milligram amounts. The major estrogen secreted by the ovary is **estradiol** (E_2), whereas the major product secreted by the placenta is estriol (E_3). Estriol is formed in the placenta by sequential desulfation and aromatization of plasma dehydroepiandrosterone sulfate (DHEA-S). Except during pregnancy, measurements of E_3 have little clinical value.

Transport in Blood

More than 97% of circulating E_2 is bound to plasma proteins. It is bound specifically and with high affinity to SHBG and nonspecifically to albumin. SHBG concentrations are increased by estrogens and therefore are higher in women than in men. They are also increased during pregnancy, oral contraceptive use, hyperthyroidism, and administration of certain antiepileptic drugs such as phenytoin (Dilantin). SHBG concentrations may decrease in hypothyroidism, obesity, or androgen excess. Only 2% to 3% of total E_2 circulates in free form. In contrast, estrone and estrone sulfate circulate bound almost exclusively to albumin. As with testosterone, both free and albumin-bound fractions of E_2 are thought to be biologically available, but measurement of this fraction has not been shown to be clinically important.

Metabolism

The metabolism of E_2 is chiefly an oxidative process dominated by three pathways, of which the fastest is oxidation of

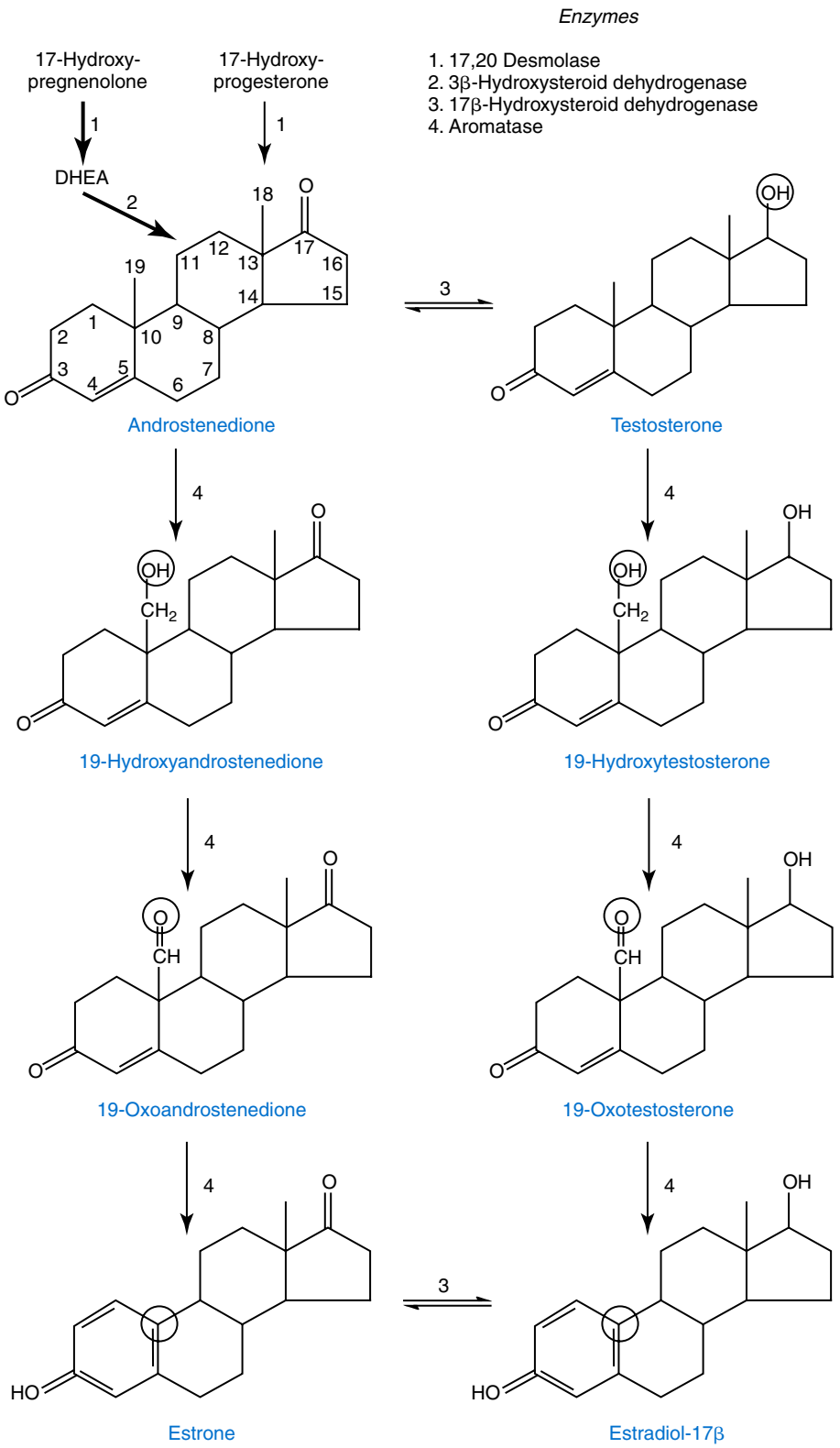


Fig. 43.5 Biosynthesis of estrogens. *Heavy arrows* indicate the Δ5-3β-hydroxy pathway. The *circled area* represents the site of chemical change. See Fig. 43.1 for early synthetic steps.

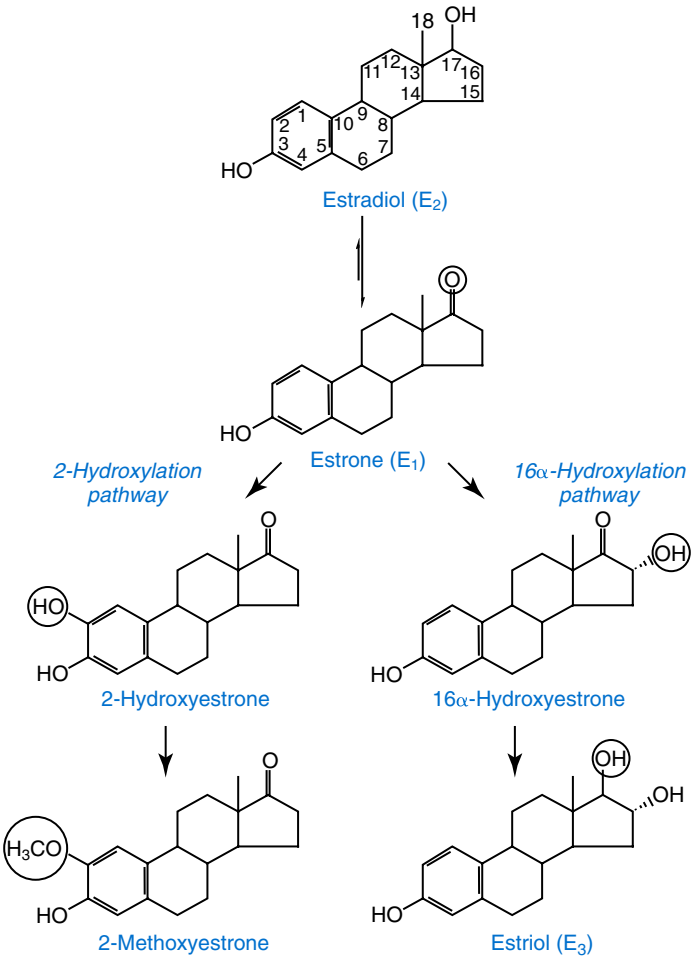


Fig. 43.6 Main pathways of estradiol metabolism in humans. The *circled area* represents the site of chemical change.

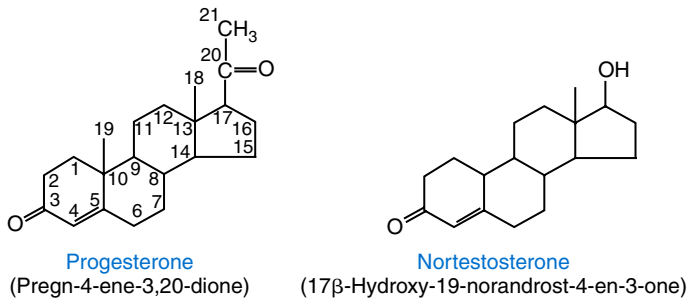


Fig. 43.7 Structural formulas of progesterone and 19-nortestosterone.

the β-hydroxy group at C-17 to a ketone (estradiol → estrone). This process is reversible; however, equilibrium favors the estrone species (Fig. 43.6).

Progesterone

Progesterone, similar to the estrogens, is a female sex hormone (Fig. 43.7). In conjunction with estrogens, it helps to regulate the accessory organs during the menstrual cycle. This hormone is especially important in preparing the uterus for implantation of the blastocyst and in maintaining pregnancy. In nonpregnant women, progesterone is secreted mainly by the corpus luteum. During pregnancy, the placenta becomes the major source of this hormone. Minor sources are the adrenal cortex in both sexes and the testes in men.

Biosynthesis

Biosynthesis of progesterone in ovarian tissues follows the same path, from acetate to cholesterol through pregnenolone, as it does in the adrenal cortex (see Fig. 43.1). In luteal tissue, however, low-density lipoprotein cholesterol is thought to serve as the preferred precursor despite the potential of the corpus luteum to synthesize progesterone de novo from acetate. Initiation and control of luteal secretion of progesterone are regulated by LH and FSH.

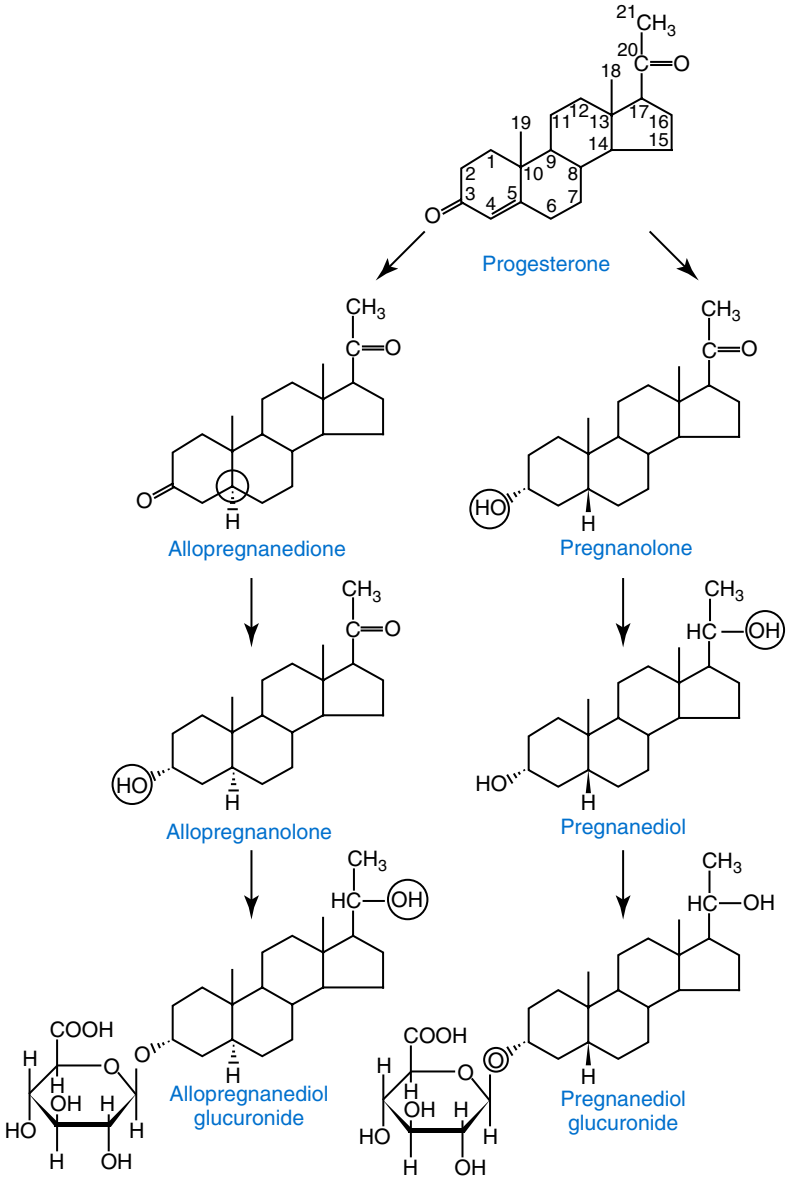


Fig. 43.8 Metabolism of progesterone. The *circled area* represents the site of chemical change.

Transport in Blood

Progesterone does not have a specific plasma-binding protein but is primarily bound to albumin with a smaller fraction bound to corticosteroid-binding globulin. Reported concentrations for plasma free progesterone vary from 2% to 10% of total concentration, and the percentage of unbound progesterone remains constant throughout the normal menstrual cycle. The production rate of progesterone during the luteal phase reaches as high as 30 mg/day (95 μmol/day), whereas the production rate of progesterone by the placenta during the third trimester of pregnancy is approximately 300 mg/day (950 μmol/day).

Metabolism

The important metabolic events leading to inactivation of progesterone are reduction and conjugation. The main metabolic pathway for the metabolism of progesterone is outlined in Fig. 43.8. Reduced metabolites are eventually conjugated with glucuronic acid and excreted as water-soluble glucuronides.

Female Reproductive Development

Reproductive development begins with anatomy during the fetal period, followed by a postnatal period of adaptation to reduced maternal sex steroids, and finishes with sexual maturation during puberty. Normal females remain fertile and menstruating until menopause.

Fetal

In the genotypic female, lack of testosterone and AMH causes regression of the Wolffian ducts and maintenance of the Müllerian ducts, thus forming the female reproductive tract. Gonadotropin activity in utero is suppressed because of high concentrations of circulating estrogens derived from the mother.

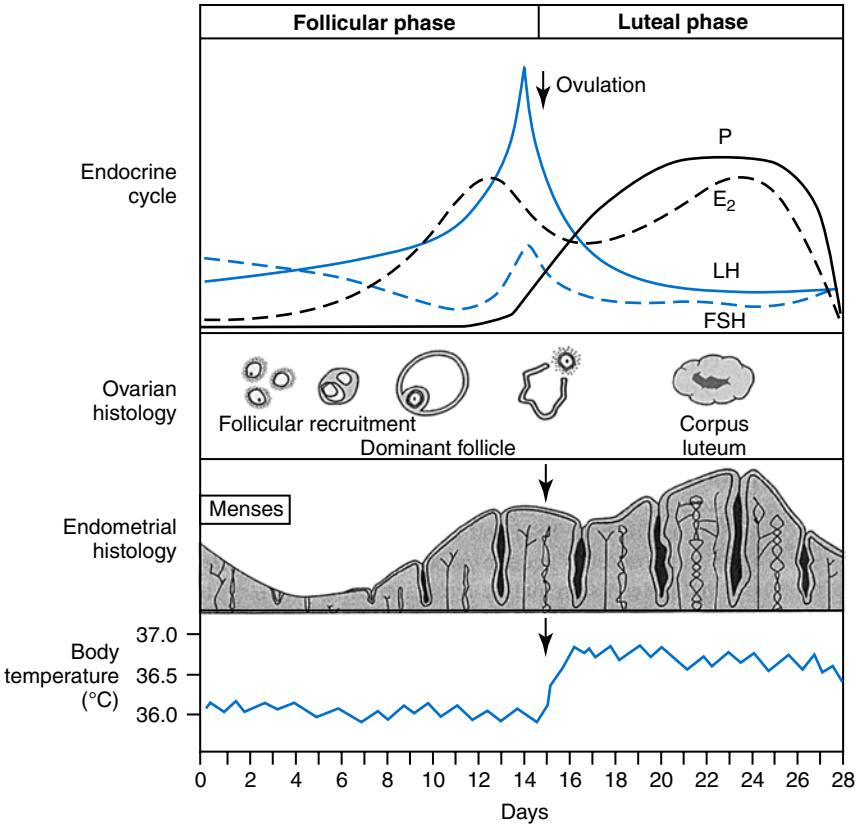


Fig. 43.9 Hormonal, ovarian, endometrial, and basal body temperature changes throughout the normal menstrual cycle. *E_2*, Estradiol; *FSH*, follicle-stimulating hormone; *LH*, luteinizing hormone. (From Carr, B. R., & Bradshaw, K. D. [2001]. Disorders of the ovary and female reproductive tract. In E. Braunwald, A. Fauci, D. Kasper, S. L. Hauser, D. L. Longo, & J. J. Jameson [Eds.], *Harrison's principles of internal medicine* [15th ed., p. 2158]. New York: McGraw-Hill; used with permission.)

Postnatal

When the placenta separates, concentrations of fetal sex steroids drop abruptly. Serum E_2 in neonates is decreased to basal concentrations within 5 to 7 days after birth and persists at this concentration until puberty. The negative feedback action of steroids is now removed and gonadotropins are released. Postnatal peaks of LH and FSH are measurable for a few months after birth, peaking at 2 to 5 months and then dropping to basal concentrations. During childhood, circulating concentrations of sex steroids and gonadotropins are low and are similar for both sexes. However, in patients with hypogonadism (Turner syndrome), LH and FSH concentrations are higher than in healthy children.

Puberty

The transition from sexual immaturity appears to begin with diminished sensitivity of the pituitary gland, hypothalamus, or both to the negative feedback effect of sex steroids. The mechanism for this change is unclear. As puberty approaches, nocturnal secretion of gonadotropins occurs. Concentrations of LH, FSH, and gonadal steroids rise gradually over several years before stabilizing at adult concentrations when full sexual maturity is reached. In girls, puberty is considered precocious if the onset of pubertal development (secondary sex characteristics) occurs before age 8 years and is considered delayed if no development has occurred by age 13 years or menarche has not occurred by age 16.5 years.

Adrenarche precedes puberty by a few years. In girls, the rise in adrenal androgen concentrations (DHEA, DHEA-S, and androstenedione) begins at age 6 to 7 years. This rise lasts until late puberty. In girls, puberty is associated with elevations in estrogen secretion by the ovary in response to gonadotropin concentrations that increase in response to GnRH. Estrogen secretion by the ovary increases, causing enlargement of the uterus and breasts. In the breast, estrogen enhances the growth of ducts; progesterone augments this effect. As the breast develops, estrogen increases adipose tissue around the lactiferous duct system, contributing to the further enlargement of breast tissue. These physiologic and physical processes associated with puberty in girls culminate in *menarche*—the beginning of menstrual function and the first menstrual period.

Normal Menstrual Cycle

During a normal menstrual cycle, a closely coordinated interplay of feedback effects occurs between the hypothalamus, anterior lobe of the pituitary gland, and ovaries. In addition, cyclic hormonal changes lead to functional and structural changes in the ovaries (follicle maturation, ovulation, and corpus luteum development), uterus (preparation of the endometrium for possible implantation of the fertilized ovum), cervix (to permit transport of sperm), and vagina (Fig. 43.9).

Phases. The menstrual cycle is measured beginning on day 1 as the first day of menstrual bleeding. Each cycle consists of a follicular phase followed by ovulation and then a luteal phase.

Follicular phase. The *follicular phase*—that is, the selection and growth of the dominant follicle—begins during the last few days of the previous luteal phase and terminates at ovulation (see Fig. 43.9). During the early part of the follicular phase, concentrations of FSH rise and then decline until ovulation (see Fig. 43.9). LH secretion begins to increase around the middle of the follicular phase. Just before ovulation, estrogen secretion by the follicle increases dramatically; this positively stimulates the hypothalamus and triggers the LH surge, which is a reliable predictor of ovulation, with onset of the surge for 90% of women occurring 16 to 58 hours before ovulation and the peak occurring 3 to 36 hours before ovulation. Ovulation occurs around day 14 in a 28-day menstrual cycle.

Luteal phase. The *luteal phase,* the last half of the cycle, is characterized by the increasing production of progesterone and estrogen from the corpus luteum with consequent gradual lowering of LH and FSH concentrations. The concentration of progesterone reaches a peak at about 8 days postovulation. If ovulation does not occur, the corpus luteum fails to form and the cyclic rise in progesterone is subnormal. If ovulation and pregnancy occur, hCG maintains the corpus luteum and progesterone continues to rise. In the absence of conception, the corpus luteum resolves, resulting in a decrease in estrogen and progesterone concentrations and a breakdown of the endometrium. The average duration of menstrual flow is 4 to 6 days and average menstrual blood loss is 30 mL.

Cyclic variation. Healthy women display considerable variation in cycle length ranging from 26 to 34 days (28 days on average). Much of the cyclic variation can be attributed to variation in the length of the follicular phase, while the length of the luteal phase remains relatively constant.

Gonadotropin-releasing hormone. GnRH triggers the surge of LH that precedes ovulation. There appear to be two separate feedback centers in the hypothalamus: a tonic negative feedback center in the basal medial hypothalamus and a cyclic positive feedback center in the anterior regions of the hypothalamus. Low concentrations of E_2, such as those that are present during the follicular phase, affect the negative feedback center, whereas high concentrations of E_2, such as those seen just before the midcycle LH peak, trigger the positive feedback center. Progesterone in combination with estrogen affects the negative feedback center in the luteal phase. GnRH is released in a pulsatile fashion and has a self-priming effect; the first dose potentiates the effects of subsequent doses. The magnitude of the LH response to GnRH increases steadily through the follicular phase and is greatest at the time of the preovulatory surge of LH, after which it declines again.

Follicle-stimulating hormone. A few days before day 1 of the cycle, FSH begins to rise (see Fig. 43.9), probably triggered by a fall in E_2 concentration that briefly eliminates the negative feedback effect. This rise in FSH initiates the growth of a cohort of ovarian follicles. LH and FSH release is pulsatile throughout the cycle; therefore the values shown in Fig. 43.9 represent integrated concentrations. As estrogen is released from the growing follicles, FSH concentrations fall again and remain low through the follicular phase. By days 5 to 7, a single dominant follicle is selected for further growth and maturation. A rise in FSH at midcycle is triggered by progesterone. During the luteal phase, FSH is suppressed by negative feedback from E_2 until a lesser FSH peak, occurring near the end of the cycle, starts off the follicular recruitment for the next cycle.

Luteinizing hormone. LH secretion is suppressed in the follicular phase by negative feedback from E_2. As E_2 production by the developing follicle increases, the effect of E_2 on the positive feedback center becomes important. Increasing release of GnRH from the hypothalamus and increasing sensitivity of the anterior lobe of the pituitary gland to GnRH lead to the midcycle surge of LH. Ovarian follicle receptors for LH, sensitized by FSH and E_2, transmit the stimulus to enhance differentiation of the theca cells and production of progesterone by the developing corpus luteum. During the luteal phase, LH production is suppressed by negative feedback from progesterone combined with E_2, but a low concentration of LH is probably necessary to prolong corpus luteum function.

Estradiol. Estradiol production by the ovary decreases near the end of a cycle but begins to increase again under the influence of FSH (see Fig. 43.9). Estradiol enhances the FSH effect on a maturing follicle through changes in FSH receptors of the follicular cells, but it suppresses pituitary FSH and LH release during the follicular phase through negative feedback. Before the mid-follicular phase, estrogen concentrations are less than 50 pg/mL (183.5 pmol/L), but they increase rapidly as the follicle matures. Estradiol production increases, reaching a midcycle peak at between 250 and 500 pg/mL (917.5–1835 pmol/L). Estradiol concentrations decrease abruptly after ovulation but increase again as the corpus luteum is formed, reaching concentrations of approximately 125 pg/mL (458.8 pmol/L) during the luteal phase. Progesterone produced by the corpus luteum combined with E_2, exerts a negative effect on the hypothalamus and anterior lobe of the pituitary gland. As a result, the secretion of LH and FSH is again suppressed during the luteal phase. Estradiol is essential for the development of proliferative endometrium and is synergistic with progesterone for the development of changes in the endometrium that initiate shedding; the decrease in negative feedback from E_2 on the anterior lobe of the pituitary gland triggers the FSH surge that begins the development of an ovarian follicle for the next cycle.

Progesterone. Progesterone is not produced in significant amounts until the midcycle LH surge and ovulation. LH enhances theca cell differentiation and progesterone production, which increase by a factor of 10 to 20 to a maximum about 8 days after the midcycle peak of LH. Progesterone is thought to stimulate the ovulatory peak of FSH and to promote the growth of secretory endometrium, which is necessary for implantation of the fertilized ovum.

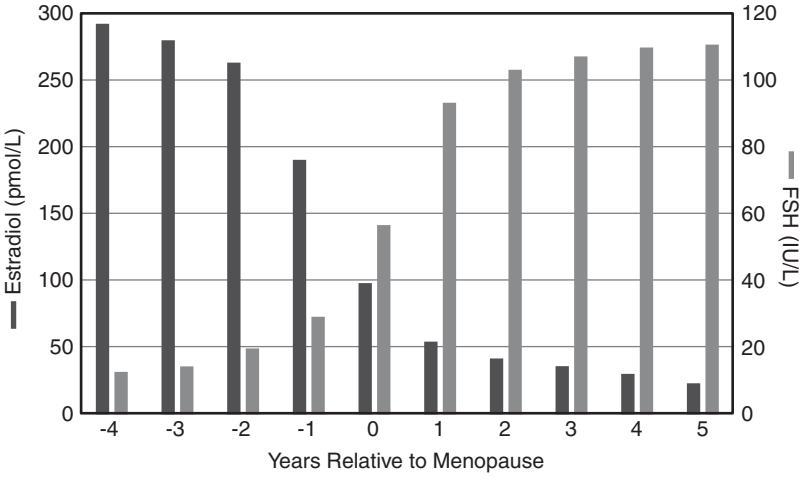


Fig. 43.10 Geometric means for follicle-stimulating hormone *(FSH)* and estradiol in relation to the final menstrual period (FMP). The horizontal axis represents time (years) with respect to the FMP (0); negative (positive) numbers represent time before (after) the FMP. (Modified from Burger, H. G., Dudley, E. C., Hopper, J. L., Groome, N., Guthrie, J. R., Green, A., et al. [1999]. Prospectively measured levels of serum follicle-stimulating hormone, estradiol, and the dimeric inhibins during the menopausal transition in a population-based cohort of women. *J Clin Endocrinol Metab, 84*, 4025–4030.)

Ovulation. An intricate interplay of endocrine events contributes to follicular maturation. Growth of ovarian follicles appears to be continuous. How an individual follicle is singled out for each menstrual cycle is not known; however, the late-cycle peak in FSH concentration is likely important in this process. Once a follicle has been stimulated, E_2 production causes that specific follicle to be more receptive to the effects of FSH. The high concentration of E_2 just before midcycle is responsible for triggering positive feedback in the hypothalamus, which leads to the midcycle LH surge. After ovulation, LH is suppressed by progesterone and E_2, but the effect of LH on the corpus luteum is increased. In the event of successful fertilization and implantation, corpus luteum function is sustained by hCG produced by trophoblastic cells of the developing embryo with high molecular homology to LH and is capable of binding and stimulating LH receptors. Otherwise the declining concentration of E_2 leads to regression of the corpus luteum and the late-cycle FSH peak that starts the process again.

Menopause

Menopause is defined as the permanent cessation of menstruation resulting from loss of ovarian follicular activity. It begins with the ovaries failing to produce adequate amounts of estrogen and inhibin; as a result, gonadotropin production is increased in a continued attempt to stimulate the ovary (Fig. 43.10). The mean age of menopause in the United States is 51 years, but it varies considerably. Ovarian failure may occur at any age, but menopause before age 40 years is considered premature.

Hormonal changes begin about 5 years before the actual menopause as the response of the ovary to gonadotropins begins to decrease and menstrual cycles become increasingly irregular. At this time, FSH concentrations increase and E_2 concentrations decrease, whereas LH and progesterone concentrations remain unchanged, indicating that menstrual cycles remain ovulatory. The decrease in estrogen concentrations gives rise to vasomotor instability and "hot flashes."

After menopause, the ovary continues to produce androgens, particularly testosterone and androstenedione, as a result of increased LH concentrations. In addition, the adrenal gland continues to secrete androgens. The resulting decrease in the estrogen/androgen ratio is the cause of the hirsutism seen in some postmenopausal women. In general, menopause may be diagnosed in women over the age of 45 on the basis of menstrual history and age without relying on laboratory test results, although serum FSH values may be helpful. In women below age 40 with menopausal symptoms and irregular **menses**, other causes of menstrual irregularities should be ruled out prior to making the diagnosis of primary ovarian insufficiency (POI).

It is important to note that perimenopausal and postmenopausal women secrete pituitary hCG. Serum concentrations are generally low (<13 IU/L), but positive hCG results often cause confusion and can delay important diagnostic tests or treatments. Pituitary versus placental hCG can be confirmed by measuring serum FSH (concentrations of FSH >45 IU/L are consistent with menopause and make pregnancy unlikely) or by 2 weeks of hormone replacement therapy (hormone replacement therapy should decrease LH, FSH, and hCG concentrations).

Female Pseudohermaphroditism

In pseudohermaphroditism, the gonadal sex varies from the genital sex. The female pseudohermaphrodite is an individual who is genetically female but whose phenotypic characteristics are, to varying degrees, male. In neonates with a 46,XX karyotype and ambiguous genitalia, *congenital adrenal hyperplasia* (CAH) should be considered. In female fetuses, exposure to androgens before the 12th week of gestation causes

ambiguous genitalia; after 13 weeks, it results in clitoral enlargement. Because androgen excess occurs before the 12th week of gestation in those with CAH, ambiguous genitalia are almost always present. Only deficiencies of 21-hydroxylase and 11β-hydroxylase are predominantly virilizing disorders. Deficiency of 3β-hydroxysteroid dehydrogenase is rare, but when it occurs, affected girls may exhibit virilization.

Diagnosis of *21-hydroxylase deficiency* is made in infants and children with excess excretion of urinary 17-KS and pregnanetriol (a metabolite of 17-hydroxyprogesterone) and elevated concentrations of plasma 17-hydroxyprogesterone and androstenedione. However, sick and premature infants may have elevated concentrations of 17-hydroxyprogesterone and androstenedione. Additionally, molecular diagnostic testing is now available for detection of the mutations that account for most cases (80% to 90%) of 21-hydroxylase deficiency.

An *11β-hydroxylase deficiency* is confirmed by finding elevated plasma concentrations of 11-deoxycortisol and deoxycorticosterone, increased concentrations of their metabolites in urine, and their suppression by glucocorticoid therapy. Plasma renin activity and aldosterone concentrations are low in this deficiency.

Elevated plasma concentrations of 17-hydroxypregnenolone, DHEA, and DHEA-S are found in patients with *3β-hydroxysteroid dehydrogenase deficiency*. Plasma concentrations of 17-hydroxyprogesterone may be elevated as a result of peripheral conversion of 17-hydroxypregnenolone. The ratio of 17-hydroxypregnenolone to 17-hydroxyprogesterone is strikingly elevated in these patients.

Precocious Puberty

Precocious puberty is the development of secondary sexual characteristics in girls younger than 8 years old and boys younger than 9 years old. Early puberty is manifested by the appearance of secondary sexual characteristics such as premature thelarche (breast development) for girls or premature testicular enlargement for boys. When presented as isolated cases, these secondary sexual characteristics are not necessarily considered to be pathologic. However, if a child demonstrates progressive development of the characteristics and/or increased rates of bone growth and maturation, causes of true precocious puberty must be considered.

GnRH-dependent precocious puberty (also called *central precocious puberty*) is due to precocious activation of the hypothalamic-pituitary-gonadal axis. In girls, the cause is most commonly idiopathic (90%); however, idiopathic cases account for less than 10% of central precocious puberty in boys. Central nervous system tumors have also been known to cause central precocious puberty, the most common being hypothalamic hamartoma. Neurofibromatosis has been documented to lead to GnRH-dependent precocious puberty.

GnRH-independent precocious puberty (also called *pseudoprecocious puberty*) refers to precocious sex steroid secretion that is independent of pituitary gonadotropin release. CAH is a common cause of pseudoprecocious puberty. Classic forms of CAH present with virilization, growth acceleration, and accelerated bone maturation. Nonclassic or late-onset forms usually present in childhood or adolescence with premature adrenarche and acne. In fact, 5% to 10% of children who present with premature adrenarche have late-onset adrenal hyperplasia. Tumors of the adrenal gland, ovaries, and testes that secrete androgens or estrogens may result in GnRH-independent precocious puberty.

Diagnosis of precocious puberty is based on clinical presentation, a thorough pubertal history, bone age determinations, and laboratory tests performed to assess gonadotropin concentrations and response to exogenous GnRH. The GnRH stimulation test is the gold standard for diagnosis of GnRH-dependent precocious puberty. Response to exogenous GnRH is suppressed and LH and FSH concentrations are low in individuals with GnRH-independent precocious puberty.

Irregular Menses

Amenorrhea, the absence of menstrual bleeding, is traditionally categorized as primary (women who have never menstruated) or secondary (women in whom menstruation is present for a variable time and then ceases). Amenorrhea is a relatively common disorder, with an estimated prevalence of 5% in the general population and as high as 8.5% in an unselected adolescent postpubescent population.

Primary amenorrhea. Primary amenorrhea is defined as failure to establish spontaneous periodic menstruation by the age of 16 years regardless of whether secondary sex characteristics have developed. About 40% of phenotypic females who have primary amenorrhea (nearly always associated with absence of development of secondary sex characteristics) have *Turner syndrome* (45,X karyotype) or *pure gonadal dysgenesis* (46,XX or XY karyotype). *Müllerian duct agenesis* or *dysgenesis* with absence of the vagina or uterus is the second most common manifestation, and the third most common is *androgen insensitivity syndrome* (androgen receptor deficiency and normal or elevated plasma testosterone concentrations if the patient is past puberty and is karyotype XY but has female sex characteristics).

Evaluation of primary amenorrhea. When puberty is delayed in a girl, serum gonadotropins should be measured. Low concentrations may indicate pituitary failure, whereas concentrations elevated into the postmenopausal interval indicate definite gonadal failure.

Secondary amenorrhea. Secondary amenorrhea is defined as absence of periodic menstruation for at least 6 months in women who have previously experienced menses. *Oligomenorrhea* is infrequent menstruation that occurs fewer than eight times per year. With few exceptions, the causes of primary and secondary amenorrhea overlap (Box 43.2). Pregnancy, the most common cause of secondary amenorrhea, must be considered first and ruled out. Elevated concentrations of prolactin—iatrogenic or induced by a prolactin-secreting tumor—have been found to result in oligomenorrhea or amenorrhea. About one-third of women with no obvious cause of amenorrhea have elevated prolactin concentrations. It is thought that hyperprolactinemia interferes with GnRH pulsatility, resulting in impaired release of LH and FSH.

BOX 43.2 Causes of Amenorrhea

Primary Amenorrhea
- Lower tract defects
  - Vaginal aplasia
  - Imperforate hymen
  - Congenital vaginal atresia
- Uterine disorders
  - Congenital absence of the uterus
  - Endometritis
  - Müllerian agenesis (Mayer-Rokitansky-Kuster-Hauser syndrome)
- Ovarian disorders
  - XO gonadal and X dysgenesis and variants
  - XX gonadal dysgenesis
  - Turner syndrome
  - 17-Hydroxylase deficiency of the ovaries and adrenal glands
  - Autoimmune oophoritis
  - Resistant ovary syndrome
  - Polycystic ovary syndrome
- Adrenal disorders (congenital adrenal hyperplasia)
- Thyroid disorders (hypothyroidism)
- Pituitary-hypothalamic disorders
  - Hypopituitarism
  - Constitutional delay in the onset of menses (physiologic)
  - Nutritional disorders
  - Kallmann syndrome

Secondary Amenorrhea
- Pregnancy/lactation
- Uterine disorders
  - Posttraumatic uterine synechiae (Asherman syndrome)
  - Progestational agents
- Ovarian disorders
  - Polycystic ovary syndrome (hypothalamic)
  - Ovarian tumor
  - Primary ovarian insufficiency (idiopathic, autoimmune, chemotherapy, radiation, injury)
  - Antimetabolite therapy
- Adrenal disorders
  - Late-onset adrenal hyperplasia
  - Cushing syndrome
  - Virilizing adrenal tumors
  - Adrenocorticoid insufficiency
- Thyroid disorders
  - Hypothyroidism
  - Hyperthyroidism
- Pituitary disorders
  - Acquired hypopituitarism (trauma, tumor, Sheehan syndrome, lymphocytic hypophysitis)
  - Physiologic or pathologic hyperprolactinemia
- Hypothalamic disorders
  - Tumor and infiltrative disease
  - Nutritional disorders
  - Hypophysitis
  - Excessive exercise
  - Stress
- Iatrogenic
  - Antipsychotics (phenothiazines, haloperidol, clozapine, pimozide)
  - Antidepressants (tricyclics, monoamine oxidase inhibitors)
  - Antihypertensives (calcium channel blockers, methyldopa, reserpine)
  - Drugs with estrogenic activity (digitalis, flavonoids, marijuana, oral contraceptives)
  - Drugs with ovarian toxicity (busulfan, chlorambucil, cisplatin, cyclophosphamide, fluorouracil)

Disorders of the ovary, such as *POI* and loss of ovarian function, have been known to cause amenorrhea. POI has been defined as failure of ovarian estrogen production occurring in a hypergonadotropic state at any age between menarche and 40 years. If the patient is younger than 25 years or shorter than 5 feet tall, karyotyping or chromosomal microarray should be performed to rule out the presence of a variety of chromosomal abnormalities involving duplications or absence of the X chromosome or the presence of a Y chromosome. Screening for the fragile X premutation *(FMR1)* should also be performed. Autoimmune disorders have been associated with 20% to 40% of cases of POI that result in destruction of the ovary and in amenorrhea. Other causes for ovarian failure include oophorectomy, cystic degeneration, trauma, infection, galactosemia, interference with blood supply, radiotherapy treatment, and treatment with cytotoxic chemotherapeutic agents. In rare patients, ovarian resistance to gonadotropins may be evident.

Secondary amenorrhea may also result from an issue with the outflow tract. The patient with a uterine problem is normal hormonally but does not menstruate. Endometrial damage may occur in response to a dilatation and curettage and to infection of the endometrium. Pituitary dysfunction will also cause secondary amenorrhea. This is most often due to intrinsic pituitary tumors. However, Sheehan syndrome and pituitary apoplexy can result in panhypopituitarism. Empty sella syndrome has been reported in 4% to 16% of patients with amenorrhea and galactorrhea.

Evaluation of secondary amenorrhea. Evaluation of women with secondary amenorrhea should begin with a careful history that includes a complete description of menstrual patterns. In addition, the patient should be evaluated for galactorrhea, hot flashes, symptoms of hypothyroidism, hirsutism, prior abdominal surgery, pelvic or uterine trauma, medications prescribed, nutritional history, patterns of exercise, previous contraceptive use, weight changes, stress, and chronic disease. Serum or urine β-hCG should be measured to rule out pregnancy (Table 43.1). Because both hypothyroidism and hyperprolactinemia have been known to cause amenorrhea, they are easily excluded by measuring concentrations of serum TSH and prolactin.

A 24-hour urine sample for cortisol measurement or an overnight dexamethasone suppression test is performed in those patients suspected of having Cushing syndrome. On the basis of the preliminary assessment, magnetic resonance imaging (MRI) of the sella turcica should be performed in patients with evidence of pituitary or hypothalamic disease, or clinical hypoestrogenism. A GnRH stimulation test with

TABLE 43.1 Differential Diagnosis of Secondary Amenorrhea

Causes	FSH	LH	E_2	Uterine Bleeding After Progesterone
Hypothalamic				
CNS—hypothalamic dysfunction				
Idiopathic	↓ or N	↓ or N	↓ or N	±
Secondary to medications	↓ or N	↓ or N	↓ or N	±
Secondary to stress	↓ or N	↓ or N	↓ or N	±
CNS—hypothalamic dysfunction or failure due to exercise	↓ or N	↓ or N	↓ or N	±
CNS—hypothalamic dysfunction or failure due to weight loss				
Simple weight loss	↓ or N	↓ or N	↓ or N	±
Anorexia nervosa	↓	↓	↓	—
CNS—hypothalamic failure				
Lesions	↓	↓	↓	—
Idiopathic	↓	↓	↓	—
CNS—hypothalamic–adreno-ovarian dysfunction (polycystic ovary syndrome) or hyperandrogenic chronic anovulation	N	↑*	N	+
Pituitary				
Destructive lesions (Sheehan syndrome)	↓	↓	↓	—
Tumor	↓	↓	↓	—
Ovarian				
Premature ovarian failure	↑	↑	↓	—
Loss of ovarian function (oophorectomy, infection, cystic degeneration, chemotherapy, radiation)	↑	↑	↓	—
Uterine				
Uterine synechiae (Asherman syndrome)	N	N	N	—

CNS, Central nervous system; *E_2,* estrogen; *FSH,* follicle-stimulating hormone; *LH,* luteinizing hormone; *N,* value within normal reference interval; *↓,* value below normal reference interval; *↑,* value above normal reference interval; *↑**, >25 IU/L, less than menopausal concentration; ±, positive or negative bleeding response to progesterone.

From Davajan, V., & Kletzky, O. A. (1991). Amenorrhea. In D. R., Mishell, V. Davajan, & R. A. Lobo (Eds.), *Infertility, contraception and reproductive endocrinology* (3rd ed., p. 373). Boston: Blackwell Scientific Publications.

measurement of LH and FSH concentrations in those patients with gonadotropin deficiency assists in differentiating hypothalamic disease from pituitary disease.

Progesterone challenge for evaluating amenorrhea. When the cause of amenorrhea is unclear after the initial assessment, relative estrogen status should be determined. Serum E_2 can be measured, but results must be interpreted with caution because serum E_2 fluctuates throughout the menstrual cycle. A *progesterone challenge* may be performed as a functional assessment of relative estrogen status. Women with an estrogen-primed uterus have withdrawal vaginal bleeding after treatment with oral progestin (medroxyprogesterone acetate; Provera). If estrogen concentrations are adequate and the outflow tract is intact, menstrual bleeding should occur within a week of treatment. In patients with withdrawal bleeding, the plasma E_2 concentration is usually greater than 40 pg/mL (146.8 pmol/L). However, the progesterone challenge test is subject to false positives because up to 20% of normoestrogenic women with oligomenorrhea do not experience withdrawal bleeding. It is also subject to false negatives because up to 40% of women with oligomenorrhea and reduced plasma E_2 experience withdrawal bleeding.

If the patient demonstrates withdrawal bleeding after the progestin challenge test, this indicates that the ovaries are producing sufficient estrogen to cause endometrial proliferation and no anatomic obstruction is present. Most of these women have a history of progestin-containing contraceptive use (the progestin-dominant state can thin the endometrium), stress, weight loss, or excessive exercise.

If bleeding fails to occur after progestin challenge, then additional laboratory tests are indicated, including measurement of FSH to localize the problem to the follicle, pituitary, or hypothalamus. High FSH concentrations indicate that the ovarian follicle is not responding to gonadotropin stimulation. A single measurement of FSH greater than 50 IU/L is suggestive of ovarian failure or POI in women younger than 40 years of age.

Androgen excess. Amenorrhea due to androgen excess can stem from adult-onset CAH, corticotropin-dependent Cushing syndrome, or **polycystic ovary syndrome (PCOS)**. Patients with androgen excess will often present with acne, obesity, and variable degrees of excess hair on the face, chest, abdomen, and thighs. Some individuals with 21-hydroxylase deficiency do not manifest any developmental abnormalities or salt wasting but present with signs of androgen excess. This clinical syndrome,

TABLE 43.2 Clinical Features of the Polycystic Ovary Syndrome[a]

Clinical Feature	Frequency (%)
Hirsutism	65
Acne	25
Obesity	35
Infertility	50
Amenorrhea	35
Oligomenorrhea	40
Regular menstrual cycle	20

[a]Data were compiled from three studies. Two used ultrasonography as the primary method of diagnosis, one used ovarian histology. Total $n = 1935$.
Modified from Franks, S. (1995). Polycystic ovary syndrome. *N Engl J Med, 333,* 853.

referred to as *nonclassic adult-onset* or *late-onset CAH,* may be clinically indistinguishable from PCOS.

PCOS is characterized by infertility, hirsutism, obesity (in approximately half of those affected), and various menstrual disturbances ranging from amenorrhea to irregular vaginal bleeding (Table 43.2). Women with PCOS have an increased prevalence of diabetes, along with increased risk for coronary heart disease and endometrial cancer. PCOS patients have substantial estrogen production because of the peripheral conversion of androgens to estrogens. Abnormal bleeding patterns seen in PCOS are due to chronic anovulation and lack of progesterone stimulation and withdrawal. Chronic estrogen exposure without progesterone may predispose patients to endometrial cancer. Although this syndrome is associated with polycystic ovaries, the name is actually a misnomer in that the ovaries are covered with follicles, not cysts.

PCOS occurs in 4% to 10% of premenopausal women, and the prevalence varies depending on the diagnostic criteria used. The Rotterdam Consensus Criteria define hyperandrogenism, oligomenorrhea or amenorrhea, and polycystic ovaries by ultrasound as the characteristic signs and symptoms and require the presence of two of the three for diagnosis of PCOS. This definition has been criticized by the Androgen Excess Society (AES), which requires the presence of hyperandrogenism in conjunction with either oligomenorrhea/amenorrhea or polycystic ovaries. Despite these concerns, current Endocrine Society guidelines recommend the use of the Rotterdam Consensus Criteria.

Because the pathophysiologic mechanism is unknown, PCOS remains a diagnosis of exclusion, clinically defined by hyperandrogenism in women with chronic anovulation and no other cause. Relatively low FSH and disproportionately high LH concentrations are common in PCOS, although this ratio should not be used in a routine PCOS workup. Studies have reported higher serum AMH concentrations in women with PCOS relative to healthy women, and ongoing work is evaluating the potential utility of including AMH measurement in PCOS diagnostic criteria.

Ovarian hyperthecosis, a nonneoplastic lesion of the ovary characterized by the presence of islands of luteinized thecal cells in the ovarian stroma, is sometimes confused with PCOS. Features that distinguish it from PCOS include higher concentrations of testosterone, androstenedione, and DHT derived from ovarian secretion.

BOX 43.3 Causes of Hirsutism

- Ovarian
  - Severe insulin resistance
  - Hyperthecosis, hilus cell or stromal cell hyperplasia
  - Androgen-producing ovarian tumor
  - Menopause
- Adrenal
  - Classic congenital hyperplasia
  - 21-Hydroxylase deficiency
  - 11-Hydroxylase deficiency
  - 3β-Hydroxysteroid dehydrogenase deficiency
  - Adult or attenuated adrenal hyperplasia
  - Androgen-producing adrenal tumor
- Familial Hirsutism
- Endocrine Disorders
  - Polycystic ovary syndrome
  - Hyperprolactinemia
  - Acromegaly
  - Cushing syndrome
- Idiopathic Hirsutism (includes increased skin sensitivity to androgens)
- Iatrogenic
  - Androgens
  - Phenytoin
  - Diazoxide
  - Minoxidil
  - Streptomycin
  - Cyclosporine
  - Danazol
  - Metyrapone
  - Phenothiazides
  - Progestagens (19-nonsteroid derivatives)

Hirsutism and virilization. Hirsutism is defined as excessive growth of terminal hair in women and children in a distribution similar to that occurring in postpubertal men. True hirsutism, which is androgen responsive, has to be distinguished from hypertrichosis, which consists of excessive growth of vellus, or non–androgen-responsive hair. Women with androgen-dependent hirsutism may have exposure to excess androgens or may have heightened sensitivity to normal circulating concentrations of androgen (Box 43.3).

Virilization is characterized by clitoral hypertrophy, deepening of the voice, temporal hair recession, baldness, increased libido, decreased body fat, and menstrual irregularities or amenorrhea. Hirsutism is usually associated with normal or slightly elevated serum androgens, whereas virilization is associated with marked increases in ovarian or adrenal androgen production and is an indication for more intensive investigation.

Laboratory evaluation of hirsutism/virilization. The two most important screening tests used in the evaluation of women for hirsutism and virilization are serum total or free testosterone and DHEA-S. Elevation of DHEA-S concentration suggests an adrenal origin of androgens, whereas elevations in testosterone indicate an adrenal or ovarian source. Neoplastic disease is unlikely if the serum testosterone concentration is less than 200 ng/dL (6.9 nmol/L), the DHEA-S

concentration is less than 700 μg/dL (19 μmol/L), or 17-KS concentrations are less than 30 mg/dL (1 mmol/L).

Other factors. Hypothalamic dysfunction consists of those disorders that disrupt the frequency or amplitude of GnRH. Rarely is this due to a lesion or tumor. However, most commonly, disruption occurs in response to psychologic stress, depression, severe weight loss, anorexia nervosa, or strenuous exercise. A syndrome known as the *female athletic triad* has been described. This syndrome is prevalent in women who exercise vigorously, and it is associated with amenorrhea, disordered eating, and osteoporosis. Competitive long-distance runners, gymnasts, and professional ballet dancers appear to be at highest risk. Although the mechanism for the disturbance is unclear, symptoms and laboratory profiles are similar to those of other forms of hypothalamic amenorrhea. LH and FSH concentrations are normal or low, and E_2 concentrations are low.

Several hormone-producing tumors of the ovary, pituitary gland, and adrenal glands occur in combination with amenorrhea. This amenorrhea may be confused with pregnancy if the tumors produce hCG. Choriocarcinoma of the uterus or ovary may produce large amounts of hCG, which cause hyperthyroidism because of the slight thyrotropic action of hCG. Granulosa–theca cell tumors are usually associated with estrogen secretion resulting in amenorrhea and irregular menses and, rarely, excessive androgen with associated virilization.

Many drugs produce amenorrhea (see Box 43.2), particularly phenothiazines and other psychotropic drugs such as haloperidol, pimozide, and clozapine. Phenothiazine-induced amenorrhea is usually associated with hyperprolactinemia and galactorrhea. Drugs that affect the normal pathway of dopamine secretion will produce amenorrhea by decreasing the secretion of norepinephrine. Because norepinephrine is important in controlling the synthesis and secretion of GnRH, any alteration in its synthesis or secretion will result in menstrual abnormalities. Normal menses should resume shortly after the cessation of combined oral contraceptive pills. If menses has not resumed after ceasing the oral contraceptive pill, evaluation is warranted.

POINTS TO REMEMBER

- GnRH stimulates LH and FSH release, which control oocyte maturation and ovulation and increase estradiol and progesterone production.
- Estradiol is transported in blood tightly bound to SHBG and loosely bound to albumin, whereas progesterone is bound to corticosteroid-binding globulin.
- Estradiol and progesterone both inhibit and stimulate LH and FSH release, with the type of regulation dependent on the stage of the menstrual cycle and the serum steroid hormone concentration.
- PCOS is defined by hyperandrogenism and oligo/anovulation and remains a diagnosis of exclusion based primarily on clinical evaluation.
- Pituitary secretion of hCG is often seen in postmenopausal women, and the pituitary origin can be confirmed by the presence of concurrently elevated serum FSH.

BOX 43.4 Male Infertility Factors

Endocrine Disorders
Hypothalamic dysfunction (Kallmann syndrome)
Pituitary failure (tumor, radiation, surgery)
Hyperprolactinemia (drug, tumor)
Androgen insensitivity syndrome (AIS)
Exogenous androgens
Thyroid disorders
Adrenal hyperplasia
Testicular failure

Anatomic
Congenital absence of vas deferens
Obstructed vas deferens
Congenital abnormalities of ejaculatory system
Varicocele
Retrograde ejaculation

Abnormal Spermatogenesis
Unexplained azoospermia
Chromosomal abnormalities
Mumps orchitis
Cryptorchidism
Chemical or radiation exposure

Abnormal Motility
Absent cilia (Kartagener syndrome)
Antibody formation

Psychosocial
Unexplained impotence
Decreased libido

Modified from Morell, V. (1997). Basic infertility assessment. *Primary Care, 24*, 195–204.

INFERTILITY

Infertility is defined as the inability to conceive after 1 year of unprotected intercourse. It has been estimated that 93% of healthy couples practicing unprotected intercourse should expect to conceive within 1 year. A specific cause of infertility is identified in approximately 80% of couples: one-third are due to female factors alone, one-third to male factors alone, and one-third to a combination of problems.

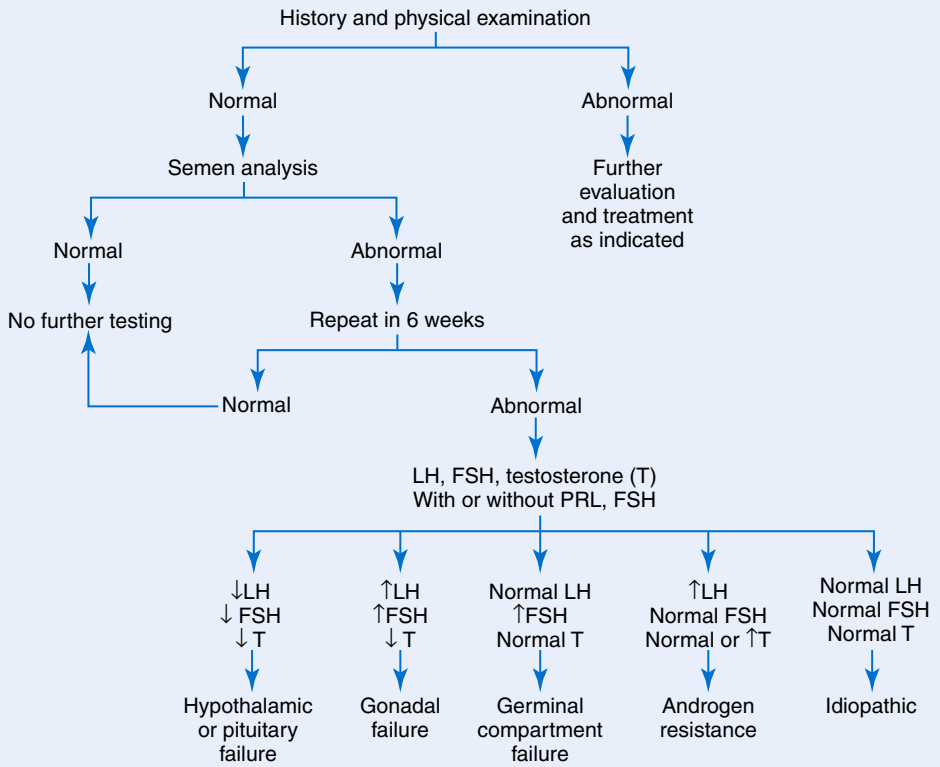
Primary infertility refers to couples or patients who have had no previous successful pregnancies. Secondary infertility encompasses patients who have previously conceived but are currently unable to conceive. These types of infertility generally share common causes.

Infertility problems often arise as a result of hormonal dysfunction of the hypothalamic-pituitary-gonadal axis. This section focuses on hormonal and biochemical aspects of evaluating infertility.

Male Infertility

A list of the most common male infertility factors is given in Box 43.4. One algorithm for the evaluation of male infertility is shown in the At a Glance box. Initial evaluation of male infertility should include a detailed history and physical

AT A GLANCE: EVALUATING MALE INFERTILITY



Algorithm for the evaluation of male infertility. *FSH,* Follicle-stimulating hormone; *LH,* luteinizing hormone; *PRL,* prolactin; *T,* testosterone.

examination. The physical examination should pay particular attention to (1) the external genitalia—for evidence of proper androgenization; (2) hair pattern—degree of virilization; (3) breast abnormalities—gynecomastia and/or discharge; and (4) neurologic findings—sense of smell and visual impairments. The history must include (1) reproductive history, including living children and any pregnancies that resulted in miscarriage; (2) prescribed medications; (3) use of recreational and performance-enhancing drugs and alcohol; (4) systemic illness; and (5) potential toxin exposure. Sexual history should include sexual technique, frequency of intercourse, and use of any lubricants. Issues of potency must be distinguished from those of infertility or subfertility. Testosterone should be measured, especially when the patient's history or physical examination suggests deficient development of secondary sex characteristics. Laboratory evaluation of male infertility should begin with evaluation of the semen (Table 43.3), which should be followed by evaluation of endocrine parameters.

Evaluation of Semen

Semen analysis measures ejaculate volume, pH, sperm count, motility, forward progression, and morphology. Although semen analysis is not a test for infertility, it is considered the most important laboratory test in the evaluation of male

TABLE 43.3 Normal Seminal Fluid Values

Parameter	Value
Ejaculate volume	>1.5 mL (1.4–1.7)[a]
Sperm density	>15 million/mL (12–16)[a]
Total sperm count	>39 million/ejaculate (33–46)[a]
Motility	>32% progressive motility (31–34); >40% total (38–42)[a]
Morphology	>4.0% normal (3–4)[a]
pH	7.2–8.0[a]
Color	Gray-white-yellow
Liquefaction	Within 40 min
Fructose	>1200 μg/mL
Acid phosphatase	100–300 μg/mL
Citric acid	>3 mg/mL
Inositol	>1 mg/mL
Zinc	>75 μg/mL
Magnesium	>70 μg/mL
Prostaglandins (PGE_1 + PGE_2)	30–200 μg/mL
Glycerylphosphorylcholine	>650 μg/mL
Carnitine	>250 μg/mL
Glucosidase	>20 mU per ejaculate

[a]Values from World Health Organization. Cooper, T. G., Noonan, E., von Eckardstein, S., Auger, J., Baker, H. W., Behre, H. M., et al. (2010). World Health Organization reference values for human semen characteristics. *Hum Reprod Update, 3*, 231.

fertility. If semen analysis is normal, it is unlikely that other laboratory testing will be useful.

Evaluation of Obstruction

Testosterone produced after administration of hCG causes the seminal vesicles, epididymis, and prostate to increase the volume of ejaculate. An appropriate increase in serum testosterone without change in the ejaculate volume may indicate mechanical blockage. Absence of or a decrease in specific biochemical markers such as acid phosphatase and citric acid (from prostate), fructose, and prostaglandins (from seminal vesicles) can assist in determining the location of blockage.

Evaluation of Endocrine Parameters

If severe oligospermia or azoospermia is found, measurement of serum testosterone, LH, and FSH concentrations is warranted, with or without measurement of prolactin and TSH. Hyperprolactinemia is a cause of secondary testicular dysfunction. Prolactin excess likely causes hypogonadism by impairing GnRH release. It also leads to under androgenization and erectile dysfunction (see earlier section titled Erectile Dysfunction). If hyperprolactinemia is found, it is imperative to check for hypothyroidism because elevated TRH concentrations can result in hyperprolactinemia.

Patients with borderline or suppressed testosterone concentrations can be evaluated with an *hCG stimulation test.* With this test, an injection of 5000 IU hCG is administered intramuscularly following collection of a basal, early-morning testosterone sample. Serum testosterone is measured 72 hours later. Hypogonadal men show a depressed rise in testosterone concentration in response to this challenge. Doubling of testosterone concentration over baseline is consistent with normal Leydig cell function. As testosterone is essential for normal sperm development, any disorder that results in hypogonadism (and thus low testosterone concentrations) results in infertility.

Hypergonadotropic hypogonadism. Measurement of the concentration of FSH is indicated in men with a sperm count lower than 5 to 10 million/mL. Elevated concentrations of FSH indicate Sertoli cell dysfunction and, in azoospermic men, primary germinal cell failure, Sertoli cell–only syndrome, or genetic conditions such as Klinefelter syndrome. Radiotherapy and gonadotoxic chemotherapy can also lead to testicular failure with azoospermia. Elevated FSH (>120 IU/L) in the setting of decreased testosterone (<200 ng/dL, 7 nmol/L) and oligospermia indicate primary testicular failure.

Hypogonadotropic hypogonadism. Decreased concentrations of testosterone (<200 ng/dL, 7 nmol/L) and decreased concentrations of FSH (<10 IU/L) are suggestive of hypogonadotropic hypogonadism. The administration of GnRH may help to distinguish between gonadal insufficiencies caused by pituitary versus hypothalamic failure. Because the pituitary is sensitive to sex steroids for appropriate gonadotropin secretion, patients with long-standing hypogonadism should be given exogenous testosterone for 1 week before the GnRH stimulation test is administered.

BOX 43.5 Female Infertility Factors

Ovarian or Hormonal Factors
- Metabolic disease
  - Thyroid
  - Liver
  - Obesity
  - Androgen excess
  - Polycystic ovarian syndrome
- Hypergonadotropic hypogonadism
  - Menopause
  - Luteal phase deficiency
  - Gonadal dysgenesis
  - Primary ovarian insufficiency (autoimmune, cytotoxic, chemotherapy, radiation, tumor)
  - Resistant ovary syndrome
- Hypogonadotropic hypogonadism
  - Hyperprolactinemia (tumor, drugs)
  - Hypothalamic insufficiency (Kallmann syndrome)
  - Pituitary insufficiency (tumor, necrosis, thrombosis, stress, exercise, anorexia)

Tubal Factors
- Occlusion or scarring
- Salpingitis isthmica nodosa
- Infectious salpingitis

Cervical Factors
- Stenosis
- Inflammation or infection
- Abnormal mucous viscosity

Uterine Factors
- Leiomyomata
- Congenital malformation
- Adhesions
- Endometritis or abnormal endometrium

Psychosocial Factors
- Decreased libido
- Anorgasmia

Iatrogenic

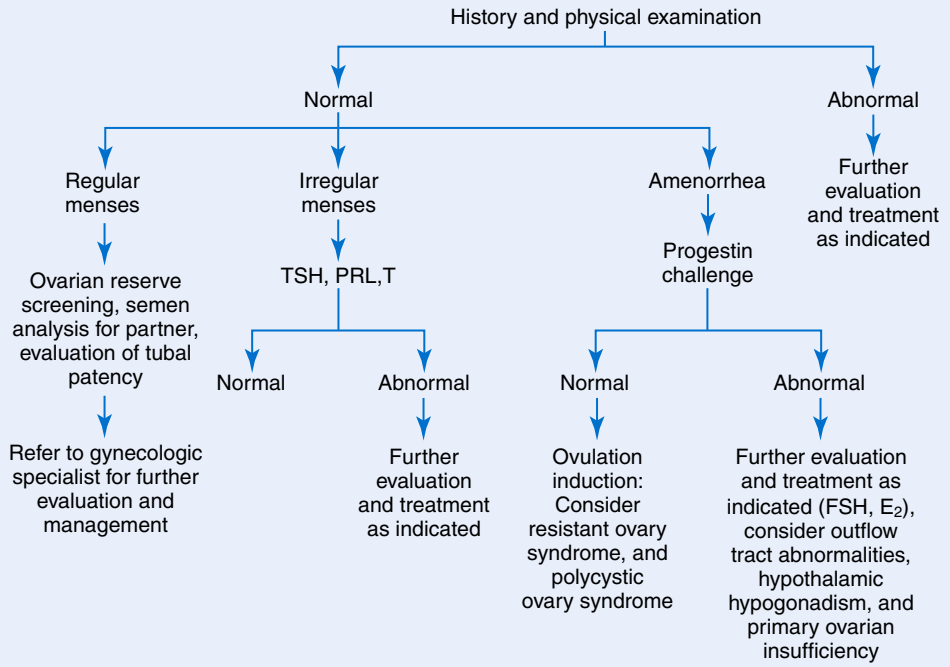
Immunologic (Antisperm Antibodies)

Modified from Morell, V. (1997). Basic infertility assessment. *Primary Care, 24,* 195–204.

Y-Chromosome Microdeletions

Deletions in either of the azoospermia factor regions (*AZF1* and *AZF2*) on the long arm of the Y chromosome are associated with an inability to make sperm. In addition, genes such as *SRY* (sex-determining region Y) are on the short arm of chromosome Y. Deletion of these regions is associated with azoospermia or, less frequently, oligospermia. The incidence of Y microdeletions in idiopathic nonobstructive azoospermic men is 8% to 18%. Testing for Y-chromosome microdeletions is performed to determine whether sperm might be retrieved on a testicular sperm extraction procedure.

AT A GLANCE: EVALUATING FEMALE INFERTILITY



Algorithm for the evaluation of female infertility. E_2, Estradiol; *FSH,* follicle-stimulating hormone; *LH,* luteinizing hormone; *PRL,* prolactin; *T,* testosterone; *TSH,* thyroid-stimulating hormone.

Female Infertility

Evaluating female infertility is more complex than evaluating infertility in the male. A list of the most common female infertility factors is given in Box 43.5. An algorithm for the evaluation of female infertility is shown in the At a Glance box.

Initial Evaluation of Female Infertility

The initial evaluation of female infertility should include a detailed history and physical examination. The physical examination should include evaluation of (1) the external genitalia and hair pattern (for signs of androgen excess, including clitoromegaly, hirsutism, and virilization), (2) the pelvis (for masses, nodularity, or tenderness), (3) the breasts (for signs of galactorrhea), (4) neurologic findings (sense of smell and visual impairments), (5) the thyroid (for enlargement or nodules), and (6) body mass index. All abnormalities in the history and physical examination should be pursued. A thorough medical and surgical history is also necessary, including an assessment of the patient's gravidity and parity, coital frequency, duration of infertility, and prior workup and treatment for infertility. History of sexually transmitted infections, assessment of previous cervical cytologic and human papillomavirus (HPV) testing and treatment, and a menstrual history should also be obtained. Concentrations of TSH, testosterone, and prolactin should be measured if menstrual cycles are absent or irregular or if signs of galactorrhea or thyroid abnormalities are present.

Evaluation of ovulation. In the menstruating woman, the next step would be to determine whether ovulation occurs. No current laboratory tests will confirm ovum release. However, measurement of the concentration of midluteal plasma progesterone does indicate that a corpus luteum has been formed. Other methods, such as measurement of the LH surge (to predict ovulation) and basal body temperature (to detect a midcycle rise in progesterone), have been used to assess ovulation.

Progesterone measurement. Measurement of the concentration of serum progesterone is the primary assay used for the evaluation of ovulation. It is important to note that an increase in progesterone concentration indicates that a corpus luteum has been formed, but it does not confirm that the oocyte was actually released. Beginning immediately after ovulation, serum progesterone concentrations rise (see Fig. 43.9); they peak within 5 to 9 days during the midluteal phase (days 21 to 23). If ovulation does not occur, the corpus luteum fails to form and the expected cyclic rise in progesterone concentration is subnormal. If pregnancy occurs, hCG maintains the corpus luteum, and progesterone production continues to rise. Midluteal progesterone concentrations greater than 300 ng/dL (9.5 nmol/L) indicate that ovulation has taken place.

Basal body temperature. Basal body temperature charts have long been accepted as simple, cost-effective indicators of ovulation. Ovulation is associated with a rapid rise in body temperature (by 0.2°F to 0.5°F, or 0.1°C to 0.3°C), which persists through the luteal phase. The rise in temperature is due

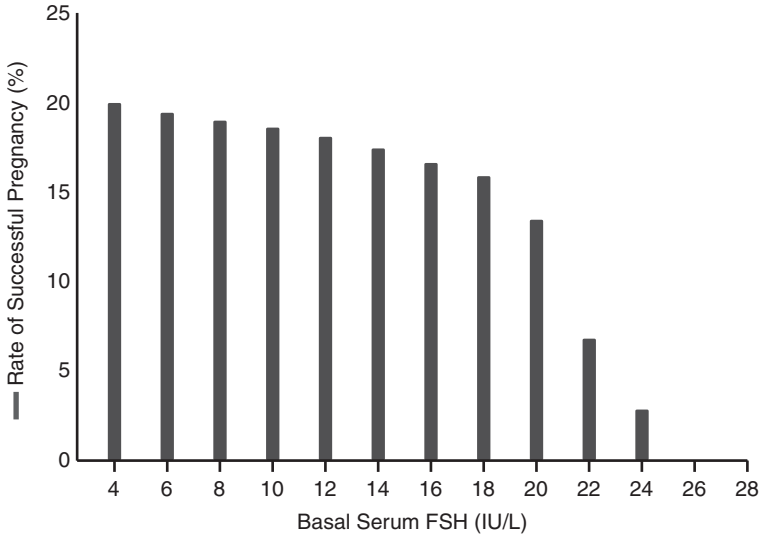


Fig. 43.11 The relationship between increasing follicle-stimulating hormone (FSH) concentrations and decreased percentage of successful pregnancies. (Modified from Jones, H., & Toner, J. P. [1993]. The infertile couple. *N Engl J Med, 329*, 1710–1715.)

to increased progesterone concentration. However, similar to progesterone, the rise in body temperature is evident only retrospectively and therefore does not predict imminent ovulation in a way helpful for timing intercourse.

Measurement of the luteinizing hormone surge. LH appears in the urine just after the serum LH surge and 24 to 36 hours before ovulation (see Fig. 43.9). Measurement of LH does not confirm ovulation or provide insight into the cause of anovulation but rather indicates when ovulation should occur and provides a guide with which to time intercourse.

Evaluation of Endocrine Parameters

Disorders of the hypothalamus, pituitary, and ovary are endocrine causes of infertility.

Hypergonadotropic hypogonadism. Premature ovarian failure is indicated by repeatedly elevated basal FSH concentrations (>30 IU/L) or a single elevation greater than 40 IU/L in a woman younger than 40 years. These patients are often hypoestrogenic (E_2 <20 pg/mL, 73 pmol/L) and do not respond to a progestin challenge because their endometrium is atrophic. A pelvic ultrasound will reveal a thin endometrium. Basal serum FSH has been used as an indicator of relative ovarian reserve. Fig. 43.11 shows the relationship between rising serum FSH and the reduced rate of successful pregnancy. A precipitous drop occurs at concentrations greater than 20 IU/L.

Assessing ovarian reserve. Women in their mid-to-late 30s and early 40s with infertility constitute the largest portion of the total infertility population. These women are also at increased risk for miscarriage. This reflects a diminished ovarian reserve as a result of follicular depletion and a decline in oocyte quality. As women age, serum FSH concentrations in the early follicular phase begin to increase. It has been suggested that this is due to a decline in the number of small follicles secreting inhibin B.

The concomitant measurement of follicular phase serum FSH and E_2 is a popular screening test for assessing ovarian reserve. In general, day 3 FSH concentrations greater than 20 to 25 IU/L are considered to be elevated and associated with a poor reproductive outcome. Concomitant measurement of serum E_2 concentration adds to the predictive power of an isolated FSH determination. Early follicular phase E_2 concentrations greater than 75 to 80 pg/mL (275 to 294 pmol/L) may be associated with a poor response to ovarian stimulation and pregnancy outcome, although these concentrations may also be slightly elevated in women with PCOS.

More recently AMH has emerged as a helpful indicator of ovarian reserve. AMH regulates follicle development and maturation as well as E_2 production. AMH is produced by small, growing follicles but ceases to be produced during FSH-dependent follicular growth or in atretic follicles. Serum AMH remains relatively constant throughout the menstrual cycle and decreases with age, leading to its promotion as a marker of ovarian reserve and reduced reproductive capacity.

Current literature suggests that the most beneficial application of AMH measurement is not in AMH's ability to predict the chance of pregnancy but rather in predicting response to controlled ovarian stimulation used in assisted reproductive techniques. Patients with elevated serum AMH concentrations are often "high responders" and at risk of ovarian hyperstimulation syndrome (OHSS) if a standard ovarian stimulation protocol is used. As a result, these women may undergo a modified stimulation protocol to minimize the risk of OHSS. At the other end of the spectrum, women with low serum AMH concentrations are considered "low responders" and often require greater ovarian stimulation to ensure successful oocyte collection.

Hypogonadotropic hypogonadism. In hypogonadotropic hypogonadism, serum E_2 concentrations are less than 40 pg/mL (147 pmol/L); therefore there is no withdrawal bleeding

with a progestin challenge because the endometrium is thin. Decreased LH (<10 IU/L) and decreased FSH (<10 IU/L) are also observed. Hyperprolactinemia can cause hypogonadotropic hypogonadic infertility. TSH should be measured to exclude hypothyroidism. Prolactin concentrations can be elevated in patients with PCOS and those taking medications such as antidepressants, cimetidine, and methyldopa as well as in stressful conditions. Prolactin concentrations can be elevated if drawn later in the day or after a meal, so they should be drawn fasting, early in the day. In cases where hyperprolactinemia is noted, radiographic imaging of the pituitary is indicated to rule out pituitary adenoma or empty sella syndrome.

Ovulatory factors. Ovulatory dysfunction is difficult to diagnose because it will manifest in the presence or absence of normal menses. Metabolic diseases of many types affect ovulatory function, including those that result in androgen excess. PCOS, which results in androgen excess, is the most common cause of anovulation. In women with hirsutism, CAH should be considered.

Historically, ovulation with inadequate luteinization and reduced progesterone secretion during the luteal phase was termed *luteal phase deficiency*. Currently, luteal phase deficiency is defined as a short luteal phase defined by less than 10 days between ovulation and menses or less than 13 days between the LH surge and menses. The clinical significance of luteal phase deficiency remains unclear.

Assisted Reproduction

Couples with a multitude of infertility problems, including unidentified causes and persistent infertility despite standard treatments, may benefit from assisted reproductive techniques (ART). If no definable cause is identified, standard initial therapy consists of ovulation induction with intrauterine insemination for three cycles before progression to more aggressive techniques such as controlled ovarian hyperstimulation (COH) with inseminations or in vitro fertilization (IVF) with or without intracytoplasmic sperm injection (ICSI).

The laboratory plays an important role in the process of COH. The principle involves the administration of gonadotropins to stimulate follicular growth, followed by hCG to stimulate follicular maturation and ovulation. Clinical, laboratory, and ultrasound monitoring of the treatment cycle is necessary to (1) identify the dose and length of therapy, (2) determine when or whether to administer hCG, and (3) obtain an adequate ovulatory response while avoiding hyperstimulation.

ANALYTIC METHODS FOR REPRODUCTIVE HORMONES

A variety of methods are available for measuring sex steroids in body fluids. Currently the most common method is nonisotopic immunoassay. However, the use of mass spectrometry to measure sex steroids is increasing.

Measurement of Sex Steroids by Mass Spectrometry

Although immunoassays are the predominant method for the detection and measurement of sex steroids (E_2, testosterone, progesterone, etc.), they are associated with considerable analytic problems. For example, automated testosterone immunoassays (nonisotopic) perform well in healthy adult males and estradiol immunoassays perform well in healthy premenopausal females, but they are unacceptable for use in children or adult patients with low hormone concentrations due to accuracy and imprecision problems at the low end of detectability.

Tandem mass spectrometry coupled with high-performance liquid chromatography (HPLC) offers several advantages over traditional immunoassays, including lower limits of detection, enhanced specificity, small sample size, decreased lot-to-lot variability, and the possibility of analyzing multiple steroids within the same sample. This ability to simultaneously measure a variety of analytes provides the opportunity for steroid profiling, which may enhance diagnostic capabilities. Because of these advantages, the Endocrine Society has suggested using extraction and chromatography followed by mass spectrometry (MS) or MS/MS as a potential gold standard for the measurement of testosterone.

However, this technology has some disadvantages, including the requirement for highly trained personnel, high costs of equipment and maintenance, and lack of standardization. Because testosterone and estradiol assays are not standardized, observations from clinical studies that demonstrate increased disease risk at a given steroid hormone concentration cannot be transferred to other institutions that utilize different assay methods. As a result, generalized cutoffs associated with disease risk cannot be defined, universally accepted reference intervals and clinical decision points cannot be established, and patient care guidelines either do not exist or cannot be applied. These differences in assay performance are attributed to differences in assay accuracy, specificity, imprecision, and calibration that are most pronounced in steroid hormone immunoassays, but mass spectrometric assays also require standardization.

Methods for Determining Total Testosterone in Blood

Circulating testosterone comprises three different forms or pools: a non–protein-bound or "free" form, a weakly bound form, and a tightly bound form. The weakly bound form is associated with albumin, and the tightly bound form is associated with SHBG, which is also known as testosterone/estradiol-binding globulin. The term *total testosterone* refers to serum measurements of free testosterone, albumin-bound testosterone, and SHBG-bound testosterone. Bioavailable testosterone includes circulating free testosterone and albumin-bound testosterone. Testosterone bound to SHBG is not biologically active, whereas the free form is available for target

cells. Albumin-bound testosterone is also available to target tissues because testosterone can dissociate from the albumin carrier and rapidly diffuse into target cells.

Methods

Direct (no extraction required) immunoassay methods have been developed for the determination of testosterone in serum or plasma. In these methods, the steroid must be displaced from its binding proteins (albumin and SHBG), and results of the assay depend on the effectiveness of the displacement. However, most routine immunoassays are not sensitive enough to measure very low testosterone concentrations, such as those found in women, children, or hypogonadal men.

Specimen collection and storage. Serum or heparinized plasma is used to measure total testosterone. Testosterone is subject to diurnal variation, reaching a peak concentration between 0400 and 0800 hours. Therefore morning specimens are preferred. DHEA supplementation should be avoided before testing.

Comments. Estimation of SHBG in serum is useful for interpreting blood concentrations of total testosterone and for calculating androgen index and bioavailable testosterone. Immunoassays for measurement of SHBG in the routine laboratory have been developed. SHBG concentrations are increased by estradiol and decreased by testosterone and therefore change with age.

Methods for Determining Free and Bioavailable Testosterone in Blood

In cases where SHBG concentrations are altered, as in women, aging men, and illness, it has been argued that measurements of free or bioavailable testosterone more accurately reflect androgen status. The three methods routinely used in clinical practice include the following:

1. Estimation of the free testosterone fraction by equilibrium dialysis or ultrafiltration
2. Estimation of combined free and weakly bound (bioavailable) testosterone fractions by selective precipitation of the tightly bound form
3. Calculation of free and weakly bound testosterone concentrations by mathematical modeling

Methods not recommended for use in clinical practice include the following:

1. Estimation of free hormone using a direct (analog tracer) radioimmunoassay
2. Calculation of the androgen index using indices that reflect the ratios of testosterone pools

Equilibrium Dialysis/Ultrafiltration

Only a small fraction (1% to 2%) of unconjugated testosterone exists in the free state (non–protein-bound) in serum or plasma. None of the conventional assay methods, including radioimmunoassay (RIA), is sufficiently sensitive to quantify free steroid directly in a protein-free ultrafiltrate of plasma. Instead, free steroid is estimated in plasma by adding a known amount of radiolabeled compound to the sample and allowing labeled and unlabeled compounds to reach equilibrium in their competition for the same binding sites on proteins. Bound and free radiolabeled fractions are then separated, and the ratio of free labeled to total labeled compound is determined. At equilibrium, this ratio is taken as a measure of the free testosterone fraction. An estimate of serum free testosterone can be calculated by multiplying the free testosterone fraction by the total testosterone concentration. Because of the labor-intensive nature of both equilibrium dialysis and ultrafiltration, these methods are used almost exclusively at reference laboratories.

Selective Precipitation

Selective precipitation of SHBG with ammonium sulfate is also used to measure bioavailable testosterone. With this technique, aliquots of serum or plasma are first incubated with radiolabeled testosterone. Testosterone bound to SHBG is then precipitated with 50% ammonium sulfate. The samples are centrifuged, and aliquots of the supernatant containing free and albumin-bound testosterone (also known as *non–SHBG-bound testosterone*) are radioactively counted. The percentage of radiolabel not bound to SHBG is subsequently multiplied by the total testosterone concentration to obtain the bioavailable testosterone. Similar to equilibrium dialysis and ultrafiltration methods, performance of selective precipitation is generally limited to reference laboratories.

Calculated Free Testosterone

Methods based on mathematical modeling use algorithms to derive non–SHBG-bound testosterone. These algorithms assume that when concentrations of total testosterone, SHBG, and albumin and the constants for binding of testosterone to SHBG and albumin are known, free testosterone and bioavailable testosterone can be calculated. These calculations are based on a proper estimation of the association constant for the binding of testosterone to SHBG and albumin. However, various association constants for SHBG have been reported, and conditions resulting in abnormal plasma protein concentrations—such as nephrotic syndrome, cirrhosis, and pregnancy—require adjustments in the assumption for albumin concentration.

Direct (Analog Tracer) Radioimmunoassay

Several RIA procedures are commercially available for the direct estimation of free testosterone. These assays use a labeled derivative (analog) of testosterone that, in theory, retains the ability to react with exogenous antitestosterone antibodies but is restricted from interacting with testosterone-binding proteins in the serum sample. In practice, development of an analog that does not interact with endogenous proteins has been difficult to achieve. As a result, direct RIAs have largely been replaced by calculation methods based on total testosterone, SHBG, and albumin concentrations.

Androgen Index

This index is a ratio of testosterone and SHBG multiplied by 100. Although this is only an indicator of free

testosterone, some have found it to be useful in the evaluation of hirsutism. Other reports have indicated that the free androgen index is not a reliable parameter of free testosterone because of its variability as a function of SHBG concentration.

Methods for Determining Testosterone Precursors and Metabolites in Blood

Several biosynthetic precursors and metabolites of testosterone are measured using specific immunoassays (directly or after sample extraction), chromatography, or LC-MS/MS.

Methods for Determining Dehydroepiandrosterone and Its Sulfate

Measurements of DHEA or its sulfated conjugate, DHEA-S, in serum and plasma are important for investigations of adrenal androgen production, such as assessment of adrenal hyperplasia, adrenal tumors, adrenarche, delayed puberty, or hirsutism. DHEA-S in circulation originates primarily from the adrenal glands. Although in men some may be derived from the testes; none is produced by the ovaries. DHEA is secreted almost entirely by the adrenal glands.

DHEA concentrations exhibit a circadian rhythm that reflects the secretion of adrenocorticotropic hormone (ACTH); these concentrations vary during the menstrual cycle. DHEA-S concentrations, however, do not exhibit a circadian rhythm because of their longer circulating half-life.

DHEA-S measurement is typically performed using nonisotopic immunoassays. Other methods include competitive protein-binding assays and GC-MS. Immunoassays for DHEA-S demonstrate significant cross-reactivity with DHEA, androstenedione, and androsterone, yet the relative concentrations of these steroids have a minimal effect on assay performance.

Specimen collection and storage

Serum or plasma (preserved with ethylenediaminetetraacetic acid [EDTA]) is suitable for DHEA or DHEA-S immunoassays. Early-morning collection, before 10:30 a.m., is preferred for DHEA. DHEA-S specimens are stable for at least 1 day at room temperature.

Determining 17-Ketosteroids in Urine

The main purpose of measuring the 17-ketosteroids (KS) is to assess adrenal androgen production. However, measurement of DHEA-S in serum serves as a more convenient marker for adrenal androgen production than does urinary 17-KS excretion because 24-hour urine collection is not required and because many drugs interfere with the 17-KS assay.

Methods for Determining Estradiol in Blood

Both chromatography–mass spectrometry and immunoassay-based methods are used to measure estrogens in blood. GC-MS methods utilizing isotope dilution provide the most accurate and reliable measurement of E_2. The main steps in these reference methods include solvent extraction, chromatographic fractionation, and chemical derivatization before analysis.

To measure E_2 directly without extraction and chromatography, the steroid must be displaced from its binding proteins. The displacing agents used in commercial methods often are not disclosed, but in some systems, effective displacement is achieved by adding 8-anilino-1-naphthalene sulfonic acid (ANS) or a large excess of a competing steroid such as DHT to the sample.

Caution should be exercised in assaying samples from subjects who are receiving oral contraceptives or estrogen replacement therapy because cross-reacting steroids may cause elevated results. Most notably, cross-reactivity with estrone is as high as 10% for some immunoassays but much lower (<1%) for others. This is likely due to a difference in the specificity of the antibody used for E_2.

Several immunoassays for E_2 have been developed and adapted for use on fully automated immunoassay systems. All are heterogeneous assays (separation step needed), but most are direct assays and do not require preliminary extraction. Most procedures offer the convenience of solid-phase separation methods. For routine clinical applications, the greatest experience is with enzyme immunoassays.

Specimen collection and storage

Serum and plasma (with EDTA or heparin as anticoagulant) have been used. Specimens should be centrifuged and separated within 24 hours.

Methods for Determining Estriol in Blood

Except for purposes of fetal aneuploidy screening, measurement of E_3 has little clinical value because in nonpregnant women E_3 is derived almost exclusively from E_2. Automated enzymatic immunoassay methods account for all unconjugated E_3 measurements. Several of these assays have been validated for use in maternal serum screening.

Methods for Determining Estrone in Blood

Estrone determinations have limited clinical utility. Normally, blood estrone concentrations parallel E_2 concentrations throughout the menstrual cycle, but at slightly lower concentrations.

Methods for Determining Progesterone in Blood

Enzyme immunoassays constitute the predominant method used for progesterone measurement. Initial immunoassays for serum progesterone measurement used organic solvents to remove the steroid from endogenous binding proteins such as corticosteroid-binding globulin and albumin. Direct (nonextraction) measurement of progesterone in serum or plasma is considered the method of choice for routine applications.

Several immunoassays are available on fully automated systems. All are heterogeneous assays that require separation of free and antibody-bound fractions. Although automated immunoassays are less labor-intensive than RIAs and yield results in less time without the need for radioactivity, these assays do not have adequate functional sensitivity for the measurement of low progesterone concentrations in men, postmenopausal women, and children.

Double-isotope derivative methods and competitive protein-binding assays have been applied to the measurement of serum progesterone, but these methods require extensive purification of the steroid and are labor-intensive. GC procedures using flame ionization, electron capture, or nitrogen detection have been used to improve the accuracy of progesterone analysis. However, these methods are time-consuming and often require solvent extraction, chromatography, and derivatization before the steroid is quantitated. GC-MS has been recommended as a reference method for progesterone determination.

Specimen Collection and Storage

Serum or plasma (with heparin or EDTA as anticoagulant) is used and should be separated within 24 hours. The patient need not be fasting, and no special handling procedures are necessary.

Measurement of Salivary Sex Steroids

Measurement of steroid concentrations in saliva has the potential to serve as a noninvasive and convenient procedure for the assessment of "serum free" steroid concentrations. However, primarily because of rapid fluctuations in salivary concentration necessitating the collection of multiple samples, salivary testing for sex steroids is not commonplace.

POINTS TO REMEMBER

- Measurement of total serum steroid hormone concentrations requires displacement of hormone from its binding proteins.
- Immunoassay remains the most common method of steroid hormone measurement, but mass spectrometry is becoming more widely used.
- Automated testosterone and estradiol immunoassays are acceptable for use in healthy adult men and women, respectively, but most lack sufficient accuracy and precision for use in children and adults with low steroid hormone concentrations.
- Mass spectrometry–based methods offer improved accuracy and a lower limit of detection but require highly trained personnel and increased equipment costs.
- Free testosterone is most accurately measured by equilibrium dialysis or ultrafiltration in a reference laboratory. Direct immunoassay measurement of free testosterone is not recommended.

REVIEW QUESTIONS

1. An elevated serum concentration of which of the following hormones is most useful in distinguishing between a pituitary versus placental source of hCG?
   a. FSH
   b. TSH
   c. Estradiol
   d. Progesterone
   e. ACTH
2. Which of the following best describes the application of mass spectrometry to steroid hormone measurement?
   a. Mass spectrometry is easily automated and does not require highly trained technologists.
   b. All mass spectrometry–based methods generate equivalent results.
   c. Mass spectrometry–based methods are usually more sensitive and specific than immunoassays.
   d. Mass spectrometers do not require routine maintenance.
   e. Mass spectrometry–based methods should not be used in women with low estradiol concentrations.
3. Which of the following laboratory tests would be most appropriate to assess anovulation?
   a. FSH on cycle day 4
   b. Estradiol on cycle day 10
   c. LH on cycle day 28
   d. Progesterone on cycle day 21
   e. Estrone on cycle day 14
4. Which of the following accurately describes free and bioavailable testosterone?
   a. Free testosterone represents the majority of circulating testosterone.
   b. Free testosterone provides an accurate reflection of androgen status in patients with altered SHBG concentrations.
   c. Measurement of free testosterone using a direct immunoassay generates the most accurate results.
   d. Free testosterone is not biologically active.
   e. Fluctuations in circulating SHBG do not affect free testosterone concentrations.
5. Which of the following is a likely diagnosis in a female patient experiencing primary amenorrhea characterized by delayed puberty with low serum FSH and LH concentrations?
   a. Gonadal dysgenesis
   b. 17α-hydroxylase deficiency
   c. Pituitary failure
   d. Uterine outflow abnormality
   e. FSH insensitivity
6. What is the most common cause of secondary amenorrhea?
   a. Prolactinoma
   b. Fragile-X primary ovarian insufficiency
   c. Autoimmune adrenal insufficiency
   d. Hypothyroidism
   e. Pregnancy

7. Which of the following correctly describes polycystic ovary syndrome?
   a. Women with PCOS experience hyperandrogenism and anovulation with no other known cause.
   b. Women with PCOS have undetectable serum estrogen concentrations.
   c. Women with PCOS have a decreased risk of developing diabetes.
   d. Women with PCOS have a decreased risk of developing coronary heart disease.
   e. Women with PCOS can be diagnosed with a single definitive laboratory test.

SUGGESTED READINGS

Bhasin, S., Cunningham, G. R., Hayes, F. J., et al. (2010). Testosterone therapy in men with androgen deficiency syndromes: an Endocrine Society clinical practice guideline. *Journal of Clinical Endocrinology and Metabolism*, *95*, 2536–2559.

Carr, B. R. (1998). Disorders of the ovary and female reproductive tract. In R. H. Williams, D. W. Foster, H. M. Kronenberg, et al. (Eds.), *Williams Textbook of Endocrinology* (9th ed) (pp. 751–817). Philadelphia: W.B. Saunders.

Demers, L. M. (2008). Testosterone and estradiol assays: current and future trends. *Steroids*, *73*, 1333–1338.

Dewailly, D., Andersen, C. Y., Balen, A., et al. (2014). The physiology and clinical utility of anti-Mullerian hormone in women. *Human Reproduction Update*, *20*, 370–385.

Gronowski, A. M., Fantz, C. R., Parvin, C. A., et al. (2008). Use of serum FSH to identify perimenopausal women with pituitary HCG. *Clinical Chemistry*, *54*, 652–656.

Jaffe, R. B. (1999). Disorders of sexual development. In S. S. C. Yen, R. B. Jaffe, & R. L. Barbieri (Eds.), *Reproductive Endocrinology: Physiology, Pathophysiology, and Clinical Management* (4th ed) (pp. 363–387). Philadelphia: W.B. Saunders.

La Marca, A., Giulini, S., Tirelli, A., et al. (2007). Anti-Mullerian hormone measurement on any day of the menstrual cycle strongly predicts ovarian response in assisted reproductive technology. *Human Reproduction*, *22*, 766–771.

Layton, J. B., Li, D., Meier, C. R., et al. (2014). Testosterone lab testing and initiation in the united kingdom and the united states, 2000 to 2011. *Journal of Clinical Endocrinology and Metabolism*, *99*, 835–842.

Legro, R. S., Arslanian, S. A., Ehrmann, D. A., et al. (2013). Diagnosis and treatment of polycystic ovary syndrome: an Endocrine Society clinical practice guideline. *Journal of Clinical Endocrinology and Metabolism*, *98*, 4565–4592.

Montag, M., Toth, B., & Strowitzki, T. (2013). New approaches to embryo selection. *Reproductive Biomedicine Online*, *27*, 539–546.

Morley, J. E., Patrick, P., & Perry, H. M. (2002). 3rd. Evaluation of assays available to measure free testosterone. *Metabolism*, *51* 554–549.

Nelson, S. M. (2013). Biomarkers of ovarian response: current and future applications. *Fertility and sterility*, *99*, 963–969.

Practice Committee of American Society for Reproductive M. (2012). Multiple gestation associated with infertility therapy: an American Society for Reproductive Medicine Practice Committee opinion. *Fertility and sterility*, *97*, 825–834.

Practice Committee of American Society for Reproductive Medicine. (2008). Current evaluation of amenorrhea. *Fertility and Sterility*, *90*, S219–S225.

Rosner, W., Auchus, R. J., Azziz, R., et al. (2007). Position statement: utility, limitations, and pitfalls in measuring testosterone: an Endocrine Society position statement. *Journal of Clinical Endocrinology and Metabolism*, *92*, 405–413.

Santen, R. J., Allred, D. C., Ardoin, S. P., et al. (2010). Postmenopausal hormone therapy: an Endocrine Society scientific statement. *Journal of Clinical Endocrinology and Metabolism*, *95*, s1–s66.

Vigen, R., O'Donnell, C. I., Baron, A. E., et al. (2013). Association of testosterone therapy with mortality, myocardial infarction, and stroke in men with low testosterone levels. *JAMA*, *310*, 1829–1836.

Yen, S. S. C. (1999). The human menstrual cycle: neuroendocrine regulation. In S. S. C. Yen, R. B. Jaffe, & R. L. Barbieri (Eds.), *Reproductive endocrinology: physiology, pathophysiology, and clinical management* (4th ed) (pp. 191–217). Philadelphia: W.B. Saunders.

44

Pregnancy and Prenatal Testing

*Melanie L. Yarbrough, Ann Gronowski**

OBJECTIVES

1. Define the following:
   - Alpha fetoprotein (AFP)
   - Amnion
   - Amniotic fluid
   - Chorion
   - Ectopic pregnancy
   - Embryo
   - Gestation
   - Meconium
   - Placenta
   - Human placental lactogen (hPL), or placental lactogen (PL)
   - Preeclampsia
   - Prenatal screening
   - Surfactant
2. Outline the events in the normal development of a fetus, beginning with conception and ending with birth.
3. Describe the structure and functions of the following, including hormones synthesized (if any), effect on maternal/fetal health if dysfunctional, and laboratory analyses and results used to determine health or disease:
   - Amnion/amniotic fluid
   - Blastocyst
   - Chorionic villus
   - Corpus luteum
   - Ovaries
   - Placenta
4. List the components of amniotic fluid and state their specific functions and what laboratory analyses are used to examine amniotic fluid.
5. Describe the major biochemical changes that take place during a normal pregnancy in maternal physiological systems, including renal, hemostatic, and endocrine changes.
6. State the clinical significance of the following laboratory analytes in the assessment of maternal and fetal health:
   - Alpha fetoprotein (AFP)
   - Human chorionic gonadotropin (hCG), or chorionic gonadotropin (CG)
   - Inhibin A (inhA)
   - Placental alpha macroglobulin-1 (PAMG-1)
   - Pregnancy-associated plasma protein A (PAPP-A)
   - Unconjugated estriol (uE_3)
7. Describe the fetal renal, gastrointestinal, hepatic, pulmonary systems, including functions, development, and maturation, and how and why these systems are assessed by the laboratory.
8. For the following disorders, list the causes and symptoms, and state the laboratory analytical procedures, specimen requirements, and laboratory results used to assess and diagnose:
   - Anencephaly
   - Down syndrome
   - Eclampsia
   - Ectopic pregnancy
   - HELLP syndrome
   - Hemolytic disease of the newborn (HDN)
   - Hyperemesis gravidarum
   - Preeclampsia
   - Respiratory distress syndrome (RDS)
   - Spina bifida
9. In regard to maternal serum screening for fetal defects, describe the composition, timing during pregnancy and utility of the quadruple and integrated tests.

KEYWORDS AND DEFINITIONS

Acute fatty liver of pregnancy A rare, life-threatening complication of pregnancy that occurs in the third trimester or the immediate period after delivery characterized by fatty infiltration of hepatocytes with little inflammation or necrosis.

Alpha fetoprotein (AFP) A protein produced in the fetal liver that is measured in maternal serum for predicting risk of anencephaly, spina bifida, and Down syndrome in the fetus.

Amniotic fluid Substance derived mostly from fetal urine that protects the developing fetus.

*The authors gratefully acknowledge the original contributions of David Grenache and Geralyn Lambert-Messerlian, upon which portions of this chapter are based.

Anencephaly A birth defect characterized by an underdeveloped brain, skull, and scalp.

Assisted reproductive technologies (ART) Procedures involving the manipulation of eggs or sperm to establish pregnancy in the treatment of infertility.

Blastocyst A thin-walled hollow structure in early embryonic development that contains a cluster of cells called the *inner cell mass* from which the embryo arises.

Cholestasis of pregnancy A condition during pregnancy characterized by impaired bile flow allowing bile salts to be deposited in the skin and the placenta.

Chorionic villi One of the minute vascular projections of the fetal chorion that combines with maternal uterine tissue to form the placenta.

Conception The union of the sperm and the ovum. Synonymous with fertilization.

Down syndrome A birth defect characterized by having three copies of chromosome 21 rather than the normal two copies. Also known as *trisomy 21*.

Eclampsia Convulsions and coma occurring in a pregnant woman or a woman who recently gave birth.

Ectopic pregnancy An embryo developing in the fallopian tube or abdomen instead of the uterus.

Embryo A developing infant that has not yet finished organ development (before 10 weeks' gestation).

Endometrium The glandular mucous membrane that lines the uterus.

Erythroblastosis fetalis A severe hemolytic disease of a fetus or newborn infant caused by the production of maternal antibodies against the fetal red blood cells, usually involving rhesus (Rh) blood group incompatibility between the mother and fetus.

Expected date of confinement (EDC) The date at which an infant is expected to be born, calculated from the date of the last menstrual period (LMP). Also called *due date*.

Fetal fibronectin (fFN) A protein produced during pregnancy that is thought to function as a "glue" attaching the fetal sac to the uterine lining.

Fetal lung maturity (FLM) A parameter that determines the likelihood a neonate will develop respiratory distress syndrome.

Fetus A developing infant that has finished organ development (following 10 weeks' gestation).

Gestation The process, state, or period of carrying an embryo or fetus from conception until birth; by convention, the time is measured clinically from the first day of the last menstrual period (LMP) and reported in weeks.

HELLP syndrome A life-threatening pregnancy complication usually considered to be a variant of preeclampsia and usually occurring between 23 and 39 weeks' gestation. The name is an acronym for the diagnostic features: *H*, Hemolysis; *EL*, elevated liver enzymes; *LP*, low platelets.

Hemolytic disease of the newborn (HDN) A disease of the fetus and newborn caused by maternal antibody–mediated fetal erythrocyte destruction.

Human chorionic gonadotropin (hCG) A placental glycoprotein hormone that stimulates the ovary to produce progesterone.

Human placental lactogen (hPL) A placental hormone, similar in structure and function to growth hormone, that disappears from the blood immediately after delivery.

Hydramnios (or polyhydramnios) An abnormality of pregnancy characterized by an accumulation of excess amniotic fluid.

Hydrops fetalis A condition in which a fetus or newborn baby accumulates fluids, causing swollen arms and legs and impaired breathing.

Hyperemesis gravidarum Extreme, excessive, and persistent vomiting in early pregnancy that may lead to dehydration and malnutrition.

In vitro fertilization (IVF) A procedure in which (1) eggs (ova) from a woman's ovary are removed, (2) the removed eggs are fertilized with sperm in a laboratory procedure, and (3) the fertilized egg (embryo) is returned to the woman's uterus.

Lamellar bodies Packages of phospholipids that are produced by type II alveolar cells and represent the storage form of surfactant. They are similar in size to platelets and are present in amniotic fluid.

Lamellar body count (LBC) Concentration of lamellar bodies in amniotic fluid. LBC increases as gestation advances.

Meconium A dark green fecal material that accumulates in the fetal intestines and is discharged at or near the time of birth.

Molar pregnancy A type of pregnancy caused by abnormal proliferation of placental tissue in the uterus. Also known as a *hydatidiform mole*.

Multiple of the median (MoM) In clinical screening, the statistic used to normalize analyte values.

Neural tube defect (NTD) A major birth defect resulting from the abnormal development of the neural tube present during embryonic life that gives rise to the central nervous system; the two most common NTDs are *spina bifida* and anencephaly.

Nuchal translucency (NT) test A measurement of the size of the translucent space behind the neck of the fetus; made using ultrasound between 10 and 14 weeks of pregnancy.

Oligohydramnios A condition in pregnancy characterized by a deficiency of amniotic fluid.

Omphalocele A birth defect in which the infant's intestine or other abdominal organs protrude from the navel.

Ovulation The release of the ripe egg (ovum) from the ovary.

Placenta A membranous vascular organ that develops in female mammals during pregnancy, lining the uterine wall and partially enveloping the fetus, to which it is attached by the umbilical cord. Following birth, the placenta is expelled.

Placental alpha microglobulin-1 (PAMG-1) A protein present in blood and the amniotic fluid and cervico-vaginal discharge of pregnant women.

Polyhydramnios (or hydramnios) The presence of excess amniotic fluid in the uterus.

Preeclampsia A disorder of widespread vascular endothelial malfunction and vasospasm that occurs after 20 weeks' gestation and can present as late as 4 to 6 weeks postpartum. It is clinically defined by hypertension and proteinuria, with or without pathologic edema.

Pregnancy The period from conception to birth. It usually lasts 40 weeks, beginning from the first day of the woman's last menstrual period (LMP), and is divided into three trimesters, each lasting 3 months.

Pregnancy-associated plasma protein-A (PAPP-A) A protein used in screening tests for Down syndrome.

Premature rupture of membranes (PROM) Breakage of the sac containing the developing fetus and the amniotic fluid prior to the start of labor; rupture before the 37th week of gestation is called *preterm PROM* (*PPROM*).

Preterm delivery The birth of an infant before 37 weeks' gestation.

Prolactin A pituitary hormone that stimulates and maintains the secretion of milk.

Puerperium The approximate 6-week period lasting from childbirth to the return of normal uterine size.

Pulmonary surfactant A fluid secreted by the cells of the alveoli that serves to reduce the surface tension of pulmonary fluids.

Respiratory distress syndrome (RDS) A disease of premature infants caused by a deficiency of lung surfactant.

Spina bifida A congenital disorder caused by the incomplete closure of the embryonic neural tube (see *neural tube defect*); the most common form is meningomyelocele (also called *myelomeningocele).*

Triploidy A condition characterized by the individual having three times the haploid number of chromosomes in the cell nucleus.

Trisomy 18 (Edwards syndrome) A genetic disorder caused by the presence of all or part of an extra 18th chromosome (trisomy: three chromosomes).

Trophoblasts The outermost layer of cells of the blastocyst that attaches the fertilized ovum to the uterine wall and serves as a nutritive pathway for the embryo.

Umbilical cord A flexible cordlike structure containing blood vessels and attaching a human or other mammalian fetus to the placenta during gestation.

Yolk sac A membranous sac attached to an embryo that provides early nourishment in the form of yolk in many animals, including humans, where it functions as the circulatory system before internal circulation begins.

Zygote The fertilized ovum or diploid cell resulting from the fusion of two haploid gametes.

The clinical laboratory has an important role in monitoring pregnancy. In contrast to most clinical situations, clinicians must simultaneously care for more than one patient. Normal reference intervals for many clinical measurements no longer apply during pregnancy, further complicating the management of the patient. This chapter reviews the biology of pregnancy and discusses laboratory tests used to detect, evaluate, and monitor both normal and abnormal pregnancies.

HUMAN PREGNANCY

To appreciate the role of laboratory tests in pregnancy healthcare, it is necessary to understand fundamental topics, such as (1) conception, embryo development, and fetal growth; (2) the role of the placenta; (3) the importance and composition of amniotic fluid; (4) maternal physiologic adaptations to pregnancy; and (5) functional maturation of the fetus.

Conception, Embryo, and Fetus

Normal human **pregnancy** (i.e., **gestation**) lasts approximately 40 weeks, as measured from the first day of the last normal menstrual period (LMP or LNMP). The anticipated date of birth of an infant is commonly referred to as the **expected date of confinement (EDC)**. Physicians customarily divide pregnancy into four time intervals. The first three time intervals are called *trimesters*, each of which is ~13 weeks. The last time interval, 37 to 42 weeks, is coined *term*. By convention, the first trimester, 0 to 13 weeks, begins on the first day of the last menses.

Ovulation occurs on approximately the 14th day of the menstrual cycle. If conception occurs, the ovum is fertilized, usually in the fallopian tube, and becomes a **zygote**, which is then carried down the tube into the uterus. The zygote divides, becoming a morula. After 50 to 60 cells are present, the morula develops a cavity, the primitive **yolk sac**, and thus becomes a **blastocyst**, which implants into the uterine wall about 5 days after fertilization. The cells on the exterior wall of the blastocyst, **trophoblasts**, synergistically invade the uterine **endometrium** and develop into **chorionic villi**, creating the placenta.

At this stage, the product of conception is referred to as an **embryo**. A cavity called the *amnion* forms and enlarges with the accumulation of amniotic fluid. Nourished by the placenta and protected by the amniotic fluid, an embryo undergoes rapid cell division, differentiation, and growth. From combinations of three primary cell types—ectoderm, mesoderm, and endoderm—organs begin to form through a process called *organogenesis*. At 10 weeks, an embryo has developed most major structures and is now referred to as a **fetus**. At 13 weeks, the fetus weighs approximately 13 g and is ~8 cm long.

Rapid fetal growth occurs during the 13 to 26 weeks of the second trimester. By the end of the second trimester, the fetus weighs approximately 700 g and is 30 cm long. In the third trimester fetal growth and maturation continue, and toward the

end of the third trimester there is a slight deceleration in the rate of fetal growth. By the end of the third trimester, the fetus weighs approximately 3200 g and is about 50 cm long. Term pregnancy is defined as greater than or equal to 37 weeks. Recently, the definition of term pregnancy was subdivided into early term (37 0/7 to 38 6/7 weeks), full term (39 0/7 to 40 6/7), and late term (41 0/7 to 41 6/7). Term birth is defined as delivery at or beyond 37 weeks. Normal labor, defined as rhythmic uterine contractions causing cervical dilation and eventually delivery of the fetus and placenta, normally occurs during this period.

Placenta

The **placenta** and the **umbilical cord** form the primary link between fetus and mother. The placenta grows throughout pregnancy and is normally delivered through the birth canal immediately after the birth of the infant.

Function

The placenta (1) keeps the maternal and fetal circulation systems separate, (2) nourishes the fetus, (3) eliminates fetal wastes, and (4) produces hormones vital to pregnancy. It is composed of large collections of fetal vessels called *villi*. These villi are finger-like projections that insert into blood-filled spaces called the *intervillous space*. The intervillous space contains maternal blood that bathes the fetal villi and facilitates bidirectional exchange between mother and fetus. For substances to move from maternal circulation to fetal circulation, they must cross through the trophoblasts and several membranes. The transfer of any substance depends largely on the (1) concentration gradient between the maternal and fetal circulatory systems, (2) presence or absence of circulating binding proteins, (3) lipid solubility of the substance, and (4) presence of facilitated transport, such as ion pumps or receptor-mediated endocytosis. The placenta is an effective barrier to the movement of large proteins and hydrophobic compounds bound to plasma proteins. Maternal immunoglobulin (Ig)G crosses the placenta via receptor-mediated endocytosis. Because of its long half-life, maternally produced IgG protects a newborn through passive immunity for the first 6 months of life. Antibody assays with low limits of detection may be positive in infants up to age 18 months because of the persistence of maternal antibodies.

Placental Hormones

The placenta produces several protein and steroid hormones. The major protein hormones are **human chorionic gonadotropin (hCG)** and **human placental lactogen (hPL)**. Steroid hormones including progesterone, estradiol, estriol, and estrone are synthesized in complex joint pathways involving maternal, placental, and fetal contributions. In general, hormone production by the placenta increases in proportion to the increase in placental mass (e.g., the increase in maternal serum hPL concentration with advancing gestational age is directly correlated with the increasing mass of placental tissue). hCG, which peaks at the end of the first trimester, is an exception.

Chorionic Gonadotropin

One of the most important placental hormones is hCG. It stimulates the ovary to produce progesterone by maintaining the corpus luteum that, in turn, prevents menstruation, thereby protecting the pregnancy. The chemistry, biochemistry, and methods for measuring hCG are discussed later in this chapter.

Placental Lactogen

Placental lactogen, also known as hPL and human chorionic somatomammotropin (hCS), is a single polypeptide chain of 191 amino acids with two intramolecular disulfide bridges. The structure of hPL is exceptionally homologous (96%) with growth hormone (GH) and less so with **prolactin** (67%). The placental secretion near term is 1 to 2 g/day, the largest of any known human hormone. hPL has many biological activities, including (1) lactogenic, (2) metabolic, (3) somatotropic, (4) erythropoietic, and (5) aldosterone-stimulating effects.

Placental Steroids

The placenta produces a wide variety of steroid hormones, including estrogen and progesterone, with large amounts of estrogens produced at term. Maternal cholesterol is the main precursor for placental progesterone production. Biosynthesis of estrogens by the placenta differs from that of the ovaries because the placenta has no 17α-hydroxylase. Thus each of the estrogens—estrone (E_1), estradiol (E_2), and estriol (E_3)—must be synthesized from C-19 intermediates that already have a hydroxyl group at position 17. In nonpregnant women, the ovaries secrete 100 to 600 μg/day of estradiol, of which about 10% is metabolized to estriol. During late pregnancy, the placenta produces 50 to 150 mg/day of estriol and 15 to 20 mg/day of estradiol and estrone. Secretion of estrogens and progesterone throughout pregnancy ensures (1) appropriate development of the endometrium, (2) uterine growth, (3) adequate uterine blood supply, and (4) preparation of the uterus for labor.

Amniotic Fluid

Throughout intrauterine life, the fetus lives within a fluid-filled compartment. The **amniotic fluid** provides a medium in which a fetus readily moves. It cushions a fetus against possible injury, helps maintain a constant temperature, and is inhaled and swallowed by the fetus during normal fetal lung development. This fluid is a dynamic medium whose volume and chemical composition are controlled within relatively narrow limits.

Volume and Dynamics

The volume of amniotic fluid increases progressively until 34 weeks' gestation, when it decreases slightly through the 40th week and then more sharply declines until the 42nd week. The volume is 200 to 300 mL at 16 weeks, 400 to 1400 mL at 26 weeks, 300 to 2000 mL at 34 weeks, and 300 to 1400 mL at 40 weeks. Pathological decreases and increases in amniotic fluid volume are encountered frequently in clinical practice. Intrauterine growth retardation and anomalies of the fetal

TABLE 44.1 Composition of Amniotic Fluid (Mean Values)

	GESTATIONAL AGE (WEEKS)		
Component	**15**	**25**	**40**
Sodium, mmol/L	136	138	126
Potassium, mmol/L	3.9	4.0	4.3
Chloride, mmol/L	111	109	103
Bicarbonate, mmol/L	16	18	16
Urea nitrogen, mg/dL (mmol urea/L)	11 (3.9)	11 (3.9)	18 (6.4)
Creatinine, mg/dL (μmol/L)	0.8 (71)	0.9 (80)	2.2 (194)
Glucose, mg/dL (mmol/L)	47 (2.6)	39 (2.2)	32 (1.8)
Uric acid, mg/dL (mmol/L)	4.0 (0.24)	5.7 (0.34)	10.4 (0.61)
Total protein, g/dL (g/L)	0.5 (5)	0.8 (8)	0.3 (3)
Bilirubin, mg/dL (μmol/L)	0.13 (2.2)	0.14 (2.4)	0.04 (0.7)
Osmolality, mOsm/kg H_2O	272	272	255

From Benzie, R. J., Doran, T. A., Harkins, J. L., Owen, V. M., & Porter, C. J. (1974). Composition of the amniotic fluid and maternal serum in pregnancy. *Am J Obstet Gynecol*, 119, 798–810.

urinary tract, such as bilateral renal agenesis or obstruction of the urethra, are associated with **oligohydramnios**, an abnormally low amniotic fluid volume. Increased fluid volume is known as **hydramnios** (also termed **polyhydramnios**). Conditions associated with hydramnios include (1) maternal diabetes mellitus, (2) severe Rh isoimmune disease, (3) fetal esophageal atresia, (4) multifetal pregnancy, (5) **anencephaly**, and (6) spina bifida.

Composition

Early in gestation, the composition of the amniotic fluid resembles a complex dialysate of the maternal serum. As a fetus grows, the amniotic fluid changes in several ways (Table 44.1). Most notably, the sodium concentration and osmolality decrease and concentrations of urea, creatinine, and uric acid increase. The activities of many enzymes in amniotic fluid have been studied with respect to both gestational age and fetal status, but have not been found to be clinically useful. The major lipids of interest are the phospholipids (PL), whose type and concentrations reflect **fetal lung maturity (FLM)** (discussed in more detail later). Numerous steroid and protein hormones are also present in amniotic fluid. The rare syndrome of congenital adrenal hyperplasia (CAH) has been diagnosed antenatally by measuring 17-hydroxyprogesterone and pregnanetriol in the amniotic fluid near term.

Early in pregnancy, little or no particulate matter is found in the amniotic fluid. By 16 weeks' gestation, large numbers of cells are present, having been shed from the surfaces of the amnion, skin, and tracheobronchial tree. These cells are of great utility in antenatal diagnosis and are the cellular source for DNA used for karyotype analysis after amniocentesis. As pregnancy continues to progress, scalp hair and lanugo (fine hair on the body of the fetus) are shed into the fluid and contribute to its turbidity. Production of surfactant particles in the lung, termed **lamellar bodies**, greatly increases the haziness of the fluid. At term, amniotic fluid contains gross particles of vernix caseosa, the oily substance composed of sebum and desquamated epithelial cells covering the fetal skin.

Normal fetuses do not defecate during pregnancy. If severely stressed, a fetus may pass stool that is called **meconium**. This heterogeneous material contains many bile pigments and therefore stains the amniotic fluid green. Meconium-stained amniotic fluid is a sign of fetal stress.

POINTS TO REMEMBER

Placenta

- A major function of the placenta is production of several protein and steroid hormones.
- hCG is a protein hormone that maintains the corpus luteum on the ovary to produce progesterone, which protects the pregnancy by preventing menstruation. Later in pregnancy, progesterone is synthesized by the placenta.
- hPL is a protein hormone that regulates maternal and fetal metabolism to facilitate growth and development of the fetus.
- Major steroid hormones, including progesterone and estriol, are synthesized by the fetoplacental unit.

Amniotic Fluid

- Amniotic fluid provides a medium to cushion and protect the fetus.
- Early in gestation, the composition of the amniotic fluid resembles a complex dialysate of the maternal serum.
- Toward the end of the first trimester, the fetal kidneys begin to produce urine, which becomes the main component of amniotic fluid.
- Fetal cells shed in amniotic fluid are a source of DNA for karyotype analysis for suspected aneuploidy.
- Decreases and increases in amniotic fluid volume are indicative of potential pathophysiological changes in pregnancy.

MATERNAL ADAPTATION

During pregnancy, a woman undergoes dramatic physiologic and hormonal changes. The large quantities of estrogens, progesterone, PL, and corticosteroids produced during pregnancy affect various metabolic, physiologic, and endocrine systems. In addition, the woman experiences (1) an increase in resistance to angiotensin, (2) a predominance of lipid metabolism over glucose use, and (3) increased synthesis by the liver of thyroid- and steroid-binding proteins, fibrinogen, and other proteins characteristic of pregnancy. As a result of such changes, many of the laboratory reference intervals for nonpregnant patients are not appropriate for pregnant patients. Reference intervals for more than 70 analytes in normal pregnancy have been developed. Mean values for selected tests expressed as a percentage of control means are presented in Table 44.2. It should be noted that these reference intervals will vary depending on testing method.

Hematologic Changes

Maternal blood volume increases during pregnancy by an average of 45%. Plasma volume increases more than red blood cell mass; therefore, despite augmented erythropoiesis,

TABLE 44.2 Mean Serum and Plasma Laboratory Values During Normal Pregnancies Expressed as a Percentage of the Nonpregnant Mean (*n* = 29)

	TIME OF GESTATION		
Analyte	**12 Weeks**	**32 Weeks**	**Term**
Sodium	97	98	97
Potassium	95	95	100
Bicarbonate	85	85	81
Chloride	98	100	99
Urea nitrogen	77	63	77
Creatinine	71	74	81
Fasting glucose	98	94	94
Bilirubin, unconjugated	56	67	78
Albumin	93	78	78
Protein	92	83	83
Uric acid	68	92	120
Calcium	98	94	97
Free ionized calcium	99	101	102
Parathyroid hormone, intact	—	—	140
1,25-Dihydroxyvitamin D	—	—	400
Phosphate	108	97	96
Magnesium	92	87	87
Alkaline phosphatase	90	203	347
Creatine kinase	87	86	135
α_1-Antitrypsin	129	174	191
Transferrin	105	160	170
Cholesterol	100	144	156
HDL-cholesterol	121	119	130
LDL-cholesterol	80	118	146
Fasting triglycerides	141	300	349
Iron	112	94	94
Iron-binding capacity	95	139	144
Transferrin saturation	136	68	64
Zinc protoporphyrin	107	109	144
Ferritin	81	33	59
Thyroxine	103	107	100
Triiodothyronine	100	121	121
Free thyroxine	98	72	74
Thyroxine-binding globulin	114	155	182
Thyroid-stimulating hormone	111	122	139
Cortisol	111	301	309
Aldosterone	—	—	1500
Prolactin	—	—	800
Hemoglobin	95	90	96
Hematocrit	94	91	97
Leukocyte count	144	167	240
Prothrombin time	99	97	97
Activated partial thromboplastin time	95	91	93
Platelet count	98	96	100
Fibrinogen	119	154	165

HDL, High-density lipoprotein; *LDL,* low-density lipoprotein.
Data from Lockitch, G, (Ed.), (1993). *Handbook of Diagnostic Biochemistry and Hematology in Normal Pregnancy.* Boca Raton: CRC Press.

hemoglobin concentration, erythrocyte count, and hematocrit decrease during normal pregnancy, producing the so-called physiologic anemia of pregnancy. Hemoglobin concentrations at term average 12.6 g/dL (126 g/L), compared with 13.3 g/dL (133 g/L) for the nonpregnant state. The leukocyte count varies considerably during pregnancy, from 4000 to 13,000/μL. During labor and **puerperium** (the interval immediately after delivery), leukocyte counts may be markedly increased.

The concentrations of several blood coagulation factors are increased during pregnancy. For example, plasma fibrinogen increases by approximately 65%, from 275 to 450 mg/dL (8.1 to 13.2 μmol/L); this increase contributes to the increase in sedimentation rate. Other clotting factors also increase, including factors VII, VIII, IX, and X. Prothrombin and factors V and XII do not change, whereas factors XI and XIII decrease slightly. Even though the platelet count remains unchanged in most women and the prothrombin time and activated partial thromboplastin time shorten slightly (see Table 44.2), pregnancy increases the risk of thromboembolism up to five times that of nonpregnant women.

Biochemical Changes

During pregnancy, electrolytes show little change, but an approximately 40% increase in serum triglycerides, cholesterol, PL, and free fatty acids is seen. Plasma albumin is decreased to an average of 3.4 g/dL (34 g/L) in late pregnancy; plasma globulin concentrations increase slightly. Several of the plasma transport proteins, including thyroxine-binding globulin (TBG), cortisol-binding globulin (CBG), and sex hormone-binding globulin (SHBG), increase markedly. Serum cholinesterase activity is reduced, whereas alkaline phosphatase activity in serum is tripled, mainly as the result of an increase in very heat-stable alkaline phosphatase of placental origin. In addition, creatine kinase can markedly increase upon delivery.

Renal Function

Pregnancy increases the glomerular filtration rate (GFR) to about 170 mL/min per 1.73 m^2 by 20 weeks, and therefore increases the clearance of urea, creatinine, and uric acid. Concentrations of these three analytes are slightly decreased in serum for much of the pregnancy. As term approaches, GFR begins to return to the nonpregnant rate. Urea and creatinine concentrations rise slightly during the last 4 weeks. During this time, tubular reabsorption of uric acid increases dramatically, which increases serum uric acid compared with the nonpregnant state. Glucosuria up to 1000 mg/day (5.55 mmol/day) may be present, owing to increased GFR, which presents more fluid to the tubules and therefore lowers the renal glucose threshold. Protein loss in the urine can increase to up to 300 mg/day.

Endocrine Changes

The action of progesterone prevents menses and thus allows pregnancy to continue. In early pregnancy, progesterone is produced by the corpus luteum of the maternal ovary in response to hCG. In later stages the placenta directly produces enough progesterone to maintain the pregnancy.

Throughout pregnancy, plasma parathyroid hormone (PTH) is increased by approximately 40%, with almost no change in the plasma free ionized calcium fraction, thus suggesting a new set point for the secretion of PTH. Calcitonin

does not increase predictably during pregnancy, whereas 1,25-dihydroxyvitamin D is increased during pregnancy and promotes increased intestinal calcium absorption to support the calcium requirements for fetal skeletal development.

Increased estrogen concentration stimulates increased hepatic production of CBG. The hepatic clearance of cortisol decreases. Thus the absolute plasma concentrations of both total and free cortisol are several times higher during pregnancy. The diurnal rhythm of cortisol, higher in the morning and lower in the evening, is maintained. Increased plasma aldosterone and deoxycorticosterone concentrations are observed.

Increasing estrogen concentrations throughout pregnancy increase the secretion of prolactin up to 10-fold. Conversely, high estrogen concentrations during pregnancy suppress the secretion of luteinizing hormone (LH) and follicle-stimulating hormone (FSH) below the detection limit. Baseline concentrations of other pituitary hormones such as thyroid stimulating hormone (TSH) remain nearly unchanged (see Table 44.2), but the GH response to provocative stimuli is blunted.

Although normal pregnancy is a euthyroid state, many changes occur in thyroid function. High concentrations of TBG raise the concentration of total thyroxine (T_4) and triiodothyronine (T_3), but a slight decrease in free T_4 concentration occurs during the second and third trimesters. A slight reciprocal increase in TSH has been reported. Thyroglobulin is significantly increased, especially in the third trimester. Very few (<0.2%) pregnant individuals develop hyperthyroidism, and hypothyroidism is very rare. Postpartum thyroid dysfunction is common and is frequently unrecognized and misdiagnosed as routine postpartum symptoms of fatigue and weight changes. The fetal thyroid-pituitary axis functions independently from the mother's axis by the end of the first trimester. However, if the mother has preexisting Graves disease (hyperthyroidism caused by autoantibodies that stimulate the thyroid), those antibodies can cross the placenta, causing hyperthyroidism in the fetus. If the mother has anti-TSH autoantibodies, the infant can develop transient hypothyroidism.

FUNCTIONAL DEVELOPMENT OF THE FETUS

Fetal organs mature during the third trimester but not at the same rate. This section reviews the lung, liver, kidneys, and blood maturation in the fetus.

Lungs and Pulmonary Surfactant

In normal air-breathing lungs, a substance called **pulmonary surfactant** coats the alveolar epithelium and responds to alveolar volume changes by reducing surface tension in the alveolar wall during expiration. Surfactant is needed because the surface tension is an inverse function of the radius of the airway. Thus small alveoli have a higher collapsing force than larger alveoli. Surfactant opposes the force and keeps the small alveoli from collapsing. Specialized alveolar cells synthesize pulmonary surfactant and package it into laminated storage granules called *lamellar bodies*. These storage granules are 1 to 5 μm in diameter and contain PL, cholesterol, and protein. Pulmonary surfactant production starts as early as 20 weeks' gestation, but adequate amounts do not accumulate until about 36 weeks. Exudation of pulmonary fluid (via the trachea) and fetal breathing movements transport lamellar bodies into the amniotic fluid.

Pulmonary surfactant is a complex mixture of lipids and proteins; less than 5% is composed of carbohydrates. The principal phospholipids present in surfactant are lecithin (phosphatidylcholine) and phosphatidylglycerol (PG), while phosphatidylinositol and sphingomyelin are present in much lower abundance. The synthesis of lecithin gradually increases from 28 weeks' gestation until birth, with the highest production occurring at week 36. Similarly, phosphatidylinositol first appears at 28 weeks' gestation, but reaches peak production between 32 and 35 weeks. PG is the last phospholipid to be produced, appearing in surfactant at 36 weeks and continuing to rise until birth. The protein fraction of lamellar bodies is approximately 4% and is composed of four surfactant-specific proteins: SP-A, SP-B, SP-C, and SP-D.

Liver

Hematopoiesis occurs in the liver during the first two trimesters and is transferred to the fetal bone marrow during the third trimester. The liver is also responsible for production of specific proteins (such as albumin and clotting factors), metabolism and detoxification of many compounds, and secretion of substances such as bilirubin. A clinically useful protein produced by the liver is **alpha fetoprotein (AFP)**. Bilirubin secretion and detoxification mechanisms are immature until late in pregnancy and even in the first few months after birth. Thus premature infants often have high serum bilirubin concentrations and metabolize drugs poorly.

Kidneys

Toward the end of the first trimester, the fetal kidneys begin to produce urine, which becomes the main component of amniotic fluid. Early nephrons cannot produce concentrated urine, and pH regulation is also limited. Complete maturation occurs after birth. Although kidneys are not required for fetal survival, amniotic fluid is required for normal lung development. Without fluid to breathe, the fetal lungs fail to properly develop. Thus newborns without kidneys die of pulmonary failure.

Fetal Blood Development

Fetal blood is produced first by the embryonic yolk sac, then by the liver, and finally by the fetal bone marrow. With the switch of erythropoiesis to the fetal liver and spleen, fetal hemoglobin (HbF) production begins. HbF consists of two α- and two γ-chains ($\alpha_2\gamma_2$). Small amounts of adult hemoglobin, HbA ($\alpha_2\beta_2$), are also produced, but HbF predominates during the remainder of fetal life.

As the fetal bone marrow begins red cell production, HbA production increases. At birth, fetal blood contains 75% HbF and 25% HbA. HbF production rapidly diminishes during the first year of postnatal life. In normal adults, less than 1%

of hemoglobin is HbF. The difference between fetal and adult hemoglobin is very significant, as HbF has a higher affinity for oxygen than HbA. Thus, in the placenta, oxygen is released from the maternal HbA, diffuses into the chorionic villi, and preferentially binds to the fetal HbF. In addition, 2,3-diphosphoglycerate (2,3-DPG) does not bind HbF and therefore cannot decrease its affinity for oxygen.

MATERNAL AND FETAL HEALTH ASSESSMENT

For optimum healthcare during pregnancy, a woman should consult her physician, preferably before conception. Preconception evaluation should include a medical, reproductive, and family history; physical examination; and laboratory testing.

Laboratory Testing

The following laboratory tests are recommended as part of a preconception evaluation: (1) hematocrit, (2) blood type and Rh compatibility, (3) erythrocyte antibody screen, (4) Papanicolaou smear (or human papillomavirus [HPV] test), (5) urinalysis, (6) rubella titer, (7) rapid plasma reagin test, (8) gonorrhea and chlamydia testing, (9) cystic fibrosis carrier status, (10) human immunodeficiency virus (HIV) antigen and antibody screen, and (11) hepatitis B surface antigen. Depending on demographic risks, genetic testing for disorders such as Tay-Sachs disease, thalassemia, and sickle cell disease should be offered. A careful diet history is warranted. Folic acid supplementation should be recommended to reduce the risk of **neural tube defects (NTDs)**. Women at high risk for diabetes mellitus should be screened for this disorder (see Chapter 33).

Many laboratory tests are useful for managing normal and abnormal pregnancies (see "At a Glance: Laboratory Evaluation of Maternal Health"). Screening for fetal NTDs and aneuploidy should be offered to all pregnant patients. Depending on diabetes risk, glucose tolerance testing should be performed immediately or at 24 to 28 weeks. Maternal observation and recording of fetal movements, ultrasound examination (biophysical profile), and fetal heart rate patterns (nonstress test and contraction stress tests) are the currently accepted methods for monitoring fetal well-being.

Diagnosis and Dating of Pregnancy

The most important aspects of pregnancy management are detection of pregnancy and establishing accurate estimates of gestational age. The most useful test for detecting pregnancy is the hCG test.

Qualitative tests for hCG in blood or urine are used to screen for pregnancy. In the first 8 weeks of pregnancy, the hCG concentration in maternal serum rises geometrically (Fig. 44.1). Detectable amounts (>5 IU/L) are present in the serum 8 to 11 days after conception. hCG usually becomes detectable in the urine 1 to 3 days later, although this interval is highly variable. For women aged 13 to 40 years, serum hCG concentrations of 5 IU/L or greater are consistent with pregnancy. Higher values are infrequently seen in older, nonpregnant women and are thought to be caused by hCG secreted by the pituitary gland. Concentrations in approximately half of pregnant women reach 25 IU/L on the first day of their

AT A GLANCE

Laboratory Evaluation of Maternal Health

First Prenatal Visit

- Complete blood count (CBC)
- Rh(D) typing and antibody screen
- Pap smear
- Urinalysis
- Rubella immunity
- Varicella immunity
- Hepatitis B screening
- HIV antibody test
- Sexually transmitted infection (STI) testing
  - Syphilis
  - Chlamydia
  - Gonorrhea

Women at increased risk may also receive testing for the following:

- Genetic testing for inherited disease
- Cystic fibrosis carrier testing
- Hemoglobinopathies
- Gonorrhea
- Tuberculosis
- Toxoplasma
- Hepatitis C

Additional testing later in pregnancy:

- Detection of fetal anomalies
  - Maternal serum screening (first and second trimester)
  - Cell free fetal DNA screening (at 10 weeks' gestation or later)
- Oral glucose tolerance testing (between 24 and 28 weeks' gestation)
- Group B *Streptococcus* screening cultures (between 35 and 37 weeks' gestation)

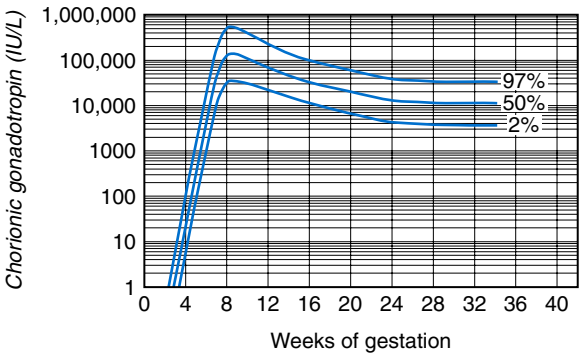


Fig. 44.1 Concentration of chorionic gonadotropin (hCG) in maternal serum as a function of gestational age. *Lines* represent the 2nd, 50th, and 97th percentiles. The maternal serum values from 14 to 25 weeks are medians calculated from 24,229 pregnancies from testing performed at ARUP Laboratories, Inc., from January to October 1997. (From Ashwood, E. R. [1992]. Evaluating health and maturation of the unborn: the role of the clinical laboratory. *Clin Chem, 38*, 1523–1529, with permission.)

missed period. The peak concentration occurs at about 8 to 10 weeks and is about 100,000 IU/L. Subsequently, hCG concentrations start to decline in serum and urine, and by the end of the second trimester, a 90% reduction from peak concentration has usually occurred. The presence of twins approximately doubles hCG concentrations.

COMPLICATIONS OF PREGNANCY

Although most pregnancies progress without problems, complications can arise in the mother, placenta, or fetus.

Abnormal Pregnancies

Conditions arising primarily in the mother include (1) ectopic pregnancy, (2) hyperemesis gravidarum, (3) preeclampsia, (4) HELLP syndrome, (5) liver disease, and (6) hemolytic disease of the newborn. The clinician must distinguish abnormal changes in laboratory tests from normal physiologic changes induced by pregnancy (see Table 44.2).

Ectopic Pregnancy and Threatened Abortion

When a fertilized egg implants in a location other than the body of the uterus, the condition is called an **ectopic pregnancy**. Most abnormal implantations occur in the fallopian tube; they can also occur in the abdomen, although this is rare. Common symptoms of ectopic pregnancy include abdominal pain, vaginal bleeding, and adnexal mass. Tubal rupture from the expanding nondistensible fallopian tube can cause a life-threatening hemorrhage and is a common cause of maternal death from ectopic pregnancy. Management of ectopic pregnancy can be surgical or medical (with methotrexate). Early detection and proper management of ectopic pregnancy are the most effective means of preventing maternal morbidity and mortality.

Ultrasound examination is used to evaluate women with symptoms. When ultrasound is nondiagnostic, serial quantitative measurements of serum hCG are used to identify women with ectopic pregnancy or abnormal intrauterine pregnancy. These conditions frequently produce abnormal hCG concentrations and slow rates of increase.

Preeclampsia and Eclampsia

Preeclampsia is a pregnancy condition characterized by hypertension and other end organ involvement, including proteinuria, cerebral vasospasm, hematologic abnormalities such as hemolysis and thrombocytopenia, and elevated transaminases (see "At a Glance: Diagnostic Criteria for Preeclampsia"). It can occur any time after 20 weeks (although there are cases of preeclampsia occurring at less than 20 weeks in abnormal pregnancies, such as **molar pregnancies**). It affects 3% to 5% of pregnancies and continues to be a major cause of maternal and perinatal mortality. If the mother develops generalized seizures, the condition is called **eclampsia**. Most maternal deaths are due to central nervous system complications, but ischemic liver damage may also occur. The only cure for preeclampsia is delivery of the placenta.

AT A GLANCE

Diagnostic Criteria for Preeclampsia

Mild Preeclampsia	Severe Preeclampsia
Hypertension: Blood pressure ≥140/90 mm Hg (or mean arterial pressure of ≥105 mm Hg)	Systolic blood pressure of ≥160 mm Hg and/or diastolic of ≥110 mm Hg on two occasions 6 h apart
AND	Oliguria (<500 mL in 24 h)
Proteinuria: ≥300 mg of protein in 24-h urine collection or ≥1+ protein on urine dipstick	Cerebral or visual disturbances
OR	Pulmonary edema or cyanosis
In patients with hypertension in the absence of proteinuria, new onset of any of the following is diagnostic of preeclampsia:	Epigastric or right upper-quadrant pain
Platelet count <100,000/μL	Impaired liver function
Serum creatinine >1.1 mg/dL (97 μmol/L) or doubling of serum creatinine in the absence of other renal disease	Thrombocytopenia
Liver transaminases elevated to twice the normal concentration	
Pulmonary edema	
Cerebral or visual symptoms	

Modified from American College of Obstetricians and Gynecologists; Task Force on Hypertension in Pregnancy. (2013). Hypertension in pregnancy: Report of the American College of Obstetricians and Gynecologists' Task Force on Hypertension in Pregnancy. *Obstet Gynecol, 122*, 1122.

HELLP Syndrome

The **HELLP syndrome** (hemolysis, elevated liver enzymes, and low platelet counts in association with preeclampsia) is a life-threatening obstetric complication that occurs in 0.1% of pregnancies. Its most prominent features are thrombocytopenia and disseminated intravascular coagulation. Most cases occur between 27 and 36 weeks' gestation, but the syndrome also may occur postpartum. Women typically present with epigastric or right upper quadrant pain, malaise, nausea, vomiting, and headache. Jaundice occurs in 5% of patients. Lactate dehydrogenase (LD) values may be very high reflecting high levels of red blood cell hemolysis, and alanine aminotransferase (ALT) and aspartate aminotransferase (AST) are usually 2 to 10 times their upper reference limits. Treatment for HELLP is delivery of the fetus.

Liver Disease

Several liver disorders are unique to pregnancy. These include (1) hyperemesis gravidarum, (2) **cholestasis of pregnancy**, and (3) fatty liver of pregnancy. These disorders must be

distinguished from the normal physiological changes of pregnancy (see Table 44.2). Significant changes normally seen in pregnancy include a dilutional decrease in serum albumin and an increase of alkaline phosphatase (from the placenta). Notably, total bilirubin, 5′-nucleotidase, gamma-glutamyl transpeptidase (GGT), ALT, and AST are unchanged in mothers with a normal pregnancy. Changes in these analytes reflect hepatobiliary disease. Also discussed in this section are non–pregnancy-related liver disease in pregnancy and the effect of pregnancy on preexisting liver disease.

Hyperemesis Gravidarum

Hyperemesis gravidarum is characterized by nausea and vomiting and, in severe cases, dehydration and malnutrition. It typically occurs in the first trimester. When hyperemesis is severe enough to cause dehydration, abnormal liver enzyme values—usually less than four times the upper reference limit—are seen in approximately 50% of patients. Mild hyperbilirubinemia may occur. However, significant liver disease does not occur, and liver biopsy results are normal. Low birth weight babies are common, especially for women who develop malnutrition.

Cholestasis of Pregnancy

Pregnancy does not preclude the acquisition or aggravation of non–pregnancy-related liver disease. Thus cholestasis during pregnancy may reflect the presence of (1) hepatotoxicity from drugs, (2) primary biliary cholangitis (formerly primary biliary cirrhosis), (3) Dubin-Johnson syndrome, or (4) cholelithiasis. Abdominal ultrasound, endoscopic retrograde cholangiography, or liver biopsy may be necessary to exclude these conditions.

The onset of cholestasis during pregnancy usually occurs in the third trimester and is manifested clinically by diffuse pruritus and, in 10% of patients, jaundice. The typical features of cholestasis, including pale stools and dark urine, are present and last until delivery. Women who experience cholestasis while taking oral contraceptives usually develop cholestasis of pregnancy. Serum bilirubin rarely exceeds 5 mg/dL (85.5 μmol/L). Alkaline phosphatase is typically two to four times the upper reference limit. Aminotransferase enzyme concentrations are mildly increased and may precede the increase of bile acids. Prothrombin time may be increased because of vitamin K malabsorption. The condition is associated with increased risk for preterm delivery and possibly fetal death, and has an increased risk for recurrence with subsequent pregnancies.

Fatty Liver of Pregnancy

Fatty liver of pregnancy occurs in approximately 1 in 10,000 pregnancies and is characterized by accumulation of microvesicular fat in the hepatocytes. Many of the cases of this maternal disorder are caused by an inherited mitochondrial fatty acid oxidation disorder in the fetus, long chain 3-hydroxyacyl-CoA dehydrogenase deficiency (LCHAD). Mothers carrying fetuses with this disorder are 50 times more likely to develop fatty liver of pregnancy. The disease typically occurs at week 37 and is manifested clinically by the rapid onset of malaise, nausea, vomiting, and abdominal pain. Mild increases in aminotransferase enzyme concentrations occur, with the AST increase typically greater than that of ALT but both typically less than six times the upper reference limit. Serum bilirubin is usually greater than 6 mg/dL (102.6 μmol/L). Life-threatening hypoglycemia may occur. Hyperuricemia, presumably from tissue destruction and renal failure, is characteristic. Liver histology shows acute fatty infiltration with little necrosis or inflammation. If untreated, fulminant hepatic failure with hepatic encephalopathy ensues (see Chapter 37). Treatment is immediate delivery, at which time rapid recovery usually occurs. Infant and maternal mortality is approximately 50% and 20%, respectively.

Effect of Pregnancy on Preexisting Liver Disease

Conception and full-term parturition do not usually occur in women who have cirrhosis. However, liver disease is not a reason for termination of a pregnancy. The hypervolemia associated with pregnancy may aggravate cirrhosis and predispose to bleeding from esophageal varices. Autoimmune chronic hepatitis is usually associated with amenorrhea, but pregnancy may occur after treatment to remission with corticosteroids.

Neonatal Thyroid Function

During the first trimester of pregnancy, the fetus is dependent on the mother for its supply of thyroid hormone. Low maternal thyroid hormone concentrations (overt or subclinical hypothyroidism) have been associated with adverse outcomes, such as preterm delivery, fetal death, and a reduced intelligence quotient (IQ) in children. Later in pregnancy, the fetal thyroid-pituitary axis functions independently from the mother's axis in most cases. However, if the mother has preexisting Graves disease, her IgG autoantibodies can cross the placenta and stimulate the fetal thyroid gland. Thus the fetus can develop hyperthyroidism. Measurement of TSH-receptor antibodies by thyrotropin-binding inhibitory immunoglobulin (TBII) or thyroid-stimulating immunoglobulin (TSI) assay is useful for assessing risk of fetal or neonatal Graves disease.

Hemolytic Disease of the Newborn

Hemolytic disease of the newborn (HDN) is a fetal hemolytic disorder caused by maternal antibodies directed against antigen on fetal erythrocytes. Commonly used synonyms for this disorder are *isoimmunization disease*, *Rh isoimmune disease*, *Rh disease*, or *D isoimmunization*. Any of a large number of erythrocyte surface antigens may be responsible for isoimmune hemolysis. When severe, the disorder is known as **erythroblastosis fetalis** and is life threatening to fetus and newborn. Disease severity is commonly assessed using noninvasive ultrasonographic determination of middle fetal cerebral artery velocity.

Etiology

Maternal sensitization may occur in response to blood transfusion or a pregnancy in which the mother is exposed

to fetal red blood cells carrying the antigen that the mother lacks. Antibodies against RhD are the most common cause of HDN, although antibodies against other erythrocyte antigens can also cause disease. The resulting maternal antibodies are actively transported across the placenta and into the fetus, where they cause destruction of the fetal erythrocytes. The severity of the resulting hemolysis is influenced by antibody specificity, titer, and transfer rate, as well as the functional maturity of the fetal spleen in which the sensitized erythrocytes are destroyed.

Destruction of the fetal erythrocytes leads to fetal anemia that imposes an extra burden on the fetal heart to provide an adequate oxygen supply to fetal tissues. Anemia stimulates the fetal marrow and extramedullary erythropoiesis in the liver and spleen to replace the destroyed erythrocytes. Extramedullary erythropoiesis destroys hepatocytes and leads to decreased production of serum albumin and decreased oncotic pressure in the intravascular space. When severe, the outcome is congestive heart failure and generalized fetal edema, a condition referred to as **hydrops fetalis**, which carries a very grave prognosis. Without therapeutic intervention via red blood cell transfusion into the umbilical cord, intrauterine demise soon follows.

Clinical Management of Rh-D Sensitized Mothers

To identify sensitized women, an alloantibody screen is performed at the first prenatal visit. If an antibody to an erythrocyte antigen is identified, the titer is determined. The critical anti-RhD titer, defined as the titer associated with risk for fetal hydrops, is usually 1:8 to 1:32. For all sensitized women, the paternal erythrocyte phenotype is determined. If the father is homozygous RhD-negative, then no follow-up studies are required. If he is D-positive, then zygosity is determined. Although this has historically been estimated from Rh antigen phenotypes (D, C/c, E/e) in conjunction with gene frequency tables based on race, DNA testing for RhD zygosity is more reliable. If the father is homozygous, then all of his offspring can be assumed to be RhD positive, negating the need for fetal RhD testing. Fetal RhD genotyping from cultured amniocytes is required if the father is heterozygous or is not available for testing. To guard against a false negative caused by a paternal RhD gene rearrangement (occurring in about 1.5% of Caucasians), the father can also be genotyped. A frequent occurrence in those of African ancestry is an RhD pseudogene; the patient is RhD-negative by serology, but RhD-positive on genotype. If the fetus is RhD-genotype-positive, the mother (who is RhD-negative serologically) should be tested for RhD genotype. In the United States, as well as other parts of the world, fetal Rh genotyping using cell-free fetal DNA that circulates in maternal blood is now available.

For sensitized mothers with an at-risk fetus, serial titers are performed on maternal serum every month until 24 weeks' gestation, and then every 2 weeks thereafter. If a critical titer anti-D is detected, ultrasound Doppler measurements are used to determine the peak velocity of blood flow in the fetal middle cerebral artery. Higher velocity is a strong indicator of fetal anemia. Historically, amniocentesis was performed every 10 to 14 days to assess the bilirubin concentration in amniotic fluid. Therapy for fetal anemia involves intrauterine percutaneous umbilical cord red blood cell transfusion. If fetal hydrops develops later in gestation (weeks 35 or beyond) or multiple transfusions have been required over the course of pregnancy, delivery timing is decided to balance the risks of continued intrauterine development versus the risk for prematurity.

Trophoblastic Disease

Serum hCG determinations are very useful for monitoring patients with germ cell–derived neoplasms or other hCG-producing tumors, such as lung carcinoma. The use of hCG in these diseases is discussed in Chapter 20.

Preterm Delivery

The leading cause of neonatal morbidity and mortality in the United States is **preterm delivery**, defined as delivery before 37 weeks' gestation. Rupture of the fetal membranes prior to the onset of uterine contractions is known as **premature rupture of membranes (PROM)**. When PROM occurs at less than 37 completed weeks' gestation, it is referred to as *preterm PROM* and is responsible for nearly one-third of preterm deliveries.

Infants born before 37 weeks' gestation are usually of low birth weight (<2500 g) and are vulnerable to numerous complications, including (1) infection, (2) necrotizing enterocolitis, and (3) intraventricular hemorrhage, and often develop (4) **respiratory distress syndrome (RDS)**. The cause of preterm labor is unknown, but it is likely that many factors are involved.

Premature Rupture of Membranes

Preterm PROM (PPROM) is a complication in 3% of pregnancies. Risk factors for PPROM include (1) a history of PPROM, (2) genital tract infection, (3) antepartum bleeding, and (4) smoking. Most women who experience PPROM will deliver their infants within 1 week. Management of PPROM varies according to the gestational age of the fetus and the presence or absence of maternal/fetal infection.

The diagnosis of PROM is often difficult, and the commonly used tests to detect it lack clinical sensitivity and specificity. These include (1) direct observation of fluid leaking from the cervix or pooling in the posterior fornix of the vagina, (2) ultrasound for the detection of oligohydramnios, (3) nitrazine (pH) and (4) fern tests, and (5) detection of placental alpha microglobulin-1 protein.

Respiratory Distress Syndrome

Respiratory distress syndrome (RDS), also called *hyaline membrane disease*, is the most common critical problem encountered in clinical management of preterm newborns. The worldwide incidence of RDS is 1% of live births and 10% to 15% of live preterm births (<37 weeks or <2500 g). The risk of RDS is affected strongly by the gestational age at the time of birth, with a likelihood of less than 5% at 37 weeks,

20% at 34 weeks, and 60% at 29 weeks. In 2010, RDS was the eighth leading cause of death in infants in the United States. Affected infants require supplemental oxygen and mechanical ventilation to remain properly oxygenated. The disorder is caused by a deficiency of pulmonary surfactant. In normal lungs, surfactant coats the alveolar epithelium and responds to alveolar volume changes by reducing the surface tension in the alveolar wall during expiration. When the quantity of surfactant is deficient, many of the alveoli collapse on expiration and thereby overinflate the remaining airways. The lungs become progressively noncompliant (stiff), and blood flowing through the capillary beds of collapsed alveoli fails to oxygenate. Infants at risk for developing RDS are treated with intratracheal administration of exogenous surfactant immediately at birth.

POINTS TO REMEMBER

Complications of Pregnancy

- When a fertilized egg implants in a location other than the body of the uterus, the condition is called an *ectopic pregnancy.*
- Preeclampsia is characterized by hypertension, proteinuria, and often edema. Eclampsia is characterized by the development of seizures in conjunction with preeclampsia.
- The diagnosis of PROM is often difficult, and the commonly used tests to detect it lack clinical sensitivity and specificity.
- Respiratory distress syndrome in a neonate is most commonly caused by a deficiency of pulmonary surfactant due to prematurity, which results in collapse of the alveoli.

Differential Diagnosis of Liver Disease

- **Acute fatty liver** is suggested in a patient with nausea and vomiting in the presence of jaundice, hypoglycemia and encephalopathy with a small or normal-sized liver.
- Preeclampsia and the HELLP syndrome are characterized by an enlarged liver, lower ALT concentrations than in liver disease, mildly increased or normal bilirubin, normal glucose, and hypertension.
- Liver biopsy may differentiate non–pregnancy-related causes of liver disease. Biopsy is not necessary to differentiate acute fatty liver from preeclampsia or HELLP syndrome because the treatment is the same for all of these conditions—delivery of the infant.

Hemolytic Disease of the Newborn

- HDN is a fetal hemolytic disorder caused by maternal antibodies directed against antigen on fetal erythrocytes.
- Maternal sensitization may occur in response to blood transfusion or exposure of the mother to fetal RBCs carrying antigen that the mother lacks.
- An alloantibody screen is performed to identify sensitized mothers. In positive women, a titer is determined to monitor risk for fetal hydrops.
- Fetal RhD genotyping using cultured amniocytes or circulating cell-free fetal DNA may be required in a sensitized mother if the father is heterozygous or is not available for testing.
- Noninvasive ultrasonographic determination of middle fetal cerebral artery velocity is useful to assess disease severity.

PRENATAL SCREENING FOR FETAL DEFECTS

Prenatal screening is the process of identifying pregnancies at sufficiently high risk of a serious birth defect such as an open NTD or **Down syndrome**. The risk for Down syndrome, calculated using screening tests, is more accurate than the use of maternal age alone, and women of all ages should be offered screening. In addition to prenatal screening tests for maternal serum analytes, detection of cell-free fetal DNA (cffDNA) in maternal blood holds promise for high quality noninvasive screening or diagnosis of some fetal aneuploidies. Prenatal screening of cffDNA is highly sensitive and specific for Down syndrome in both high-risk and routine pregnancies. (The reader is referred to chapter 53 of the *Tietz Textbook of Clinical Chemistry and Molecular Diagnostics*, 6th edition, for a more detailed discussion of this technique and its applications.) However, cffDNA testing does not assess risk of NTDs and is not recommended for multiple gestations. Due to this and the lack of data on low-risk pregnancies, the American College of Obstetricians and Gynecologists (ACOG) recommends measurement of maternal serum analytes for routine prenatal screening.

Terminology and Method of Risk Calculation in Prenatal Screening

Multiple of the Median

Understanding prenatal screening requires an understanding of the **multiple of the median (MoM)**, the statistic used to normalize analyte values. The initial step in calculating a MoM is to develop a set of median values for each week (or day) of gestation, using the laboratory's own assay values measured on the population to be screened. Fig. 44.2 graphically illustrates typical median values for the second-trimester markers used in prenatal screening. Individual test results are then expressed as a MoM by dividing each individual test result by the median for the relevant gestational age. The MoM is universally used as a common factor for converting analyte values into an interpretative unit and serves as the starting point for calculating screening results for NTDs, Down syndrome, and trisomy 18.

Calculating Individualized Patient-Specific Risks Using Multiple Biochemical Measurements

Measurements of each analyte are made on a serum sample, and the results in mass units are converted to MoM for the appropriate week of gestation. This MoM value is then adjusted for other variables, such as maternal weight and race. The individualized risk (patient-specific risk) for any given condition is determined by multiplying the a priori risk for that condition by a likelihood ratio that is calculated using the woman's MoM values, as shown in the following equation:

$$\text{Patient risk} = \text{A priori risk} \times \text{Likelihood ratio}$$

The a priori risk is obtained from large epidemiological studies that ascertain the prevalence for the condition under consideration. For example, a woman's age is used to define her a priori risk for having a fetus with Down syndrome. The

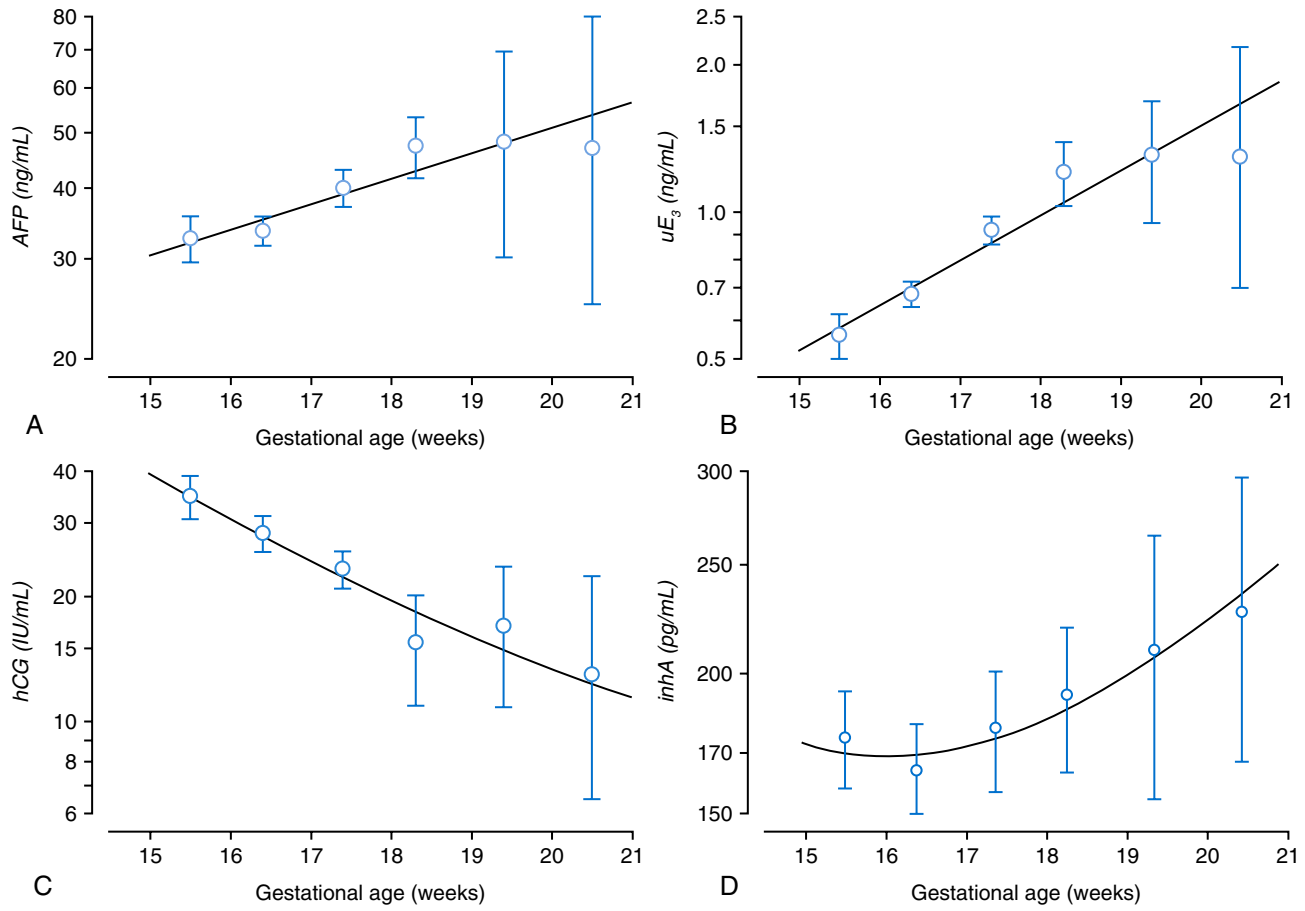


Fig. 44.2 Typical median values for maternal serum alpha-fetoprotein *(AFP)* (A), unconjugated estriol *(uE_3)* (B), human chorionic gonadotropin *(hCG)* (C), and inhibin A (inhA) (D) in the second trimester. The observed average gestational age is plotted on the horizontal axis, and the median of that week's marker measurements is plotted on the vertical axis. The thin vertical lines represent the 95% confidence interval of the medians.

likelihood ratio is determined by calculating the ratio of the heights of the affected and unaffected overlapping population distributions for any specified MoM value. When multiple tests are used, a single likelihood ratio is calculated using the overlapping distributions for each test but with the correlation between tests taken into account. This final calculated risk, rather than the analyte concentrations themselves, is the screening variable upon which clinical decisions are made.

PRENATAL SCREENING TESTS

Maternal serum screening for NTDs began in the 1970s using second trimester serum specimens. Since then, maternal serum screening has expanded to include partial detection of Down syndrome and trisomy 18, among others.

Neural Tube Defects

NTDs are serious abnormalities that occur early in embryonic development. By 19 days after fertilization, the area that is to form the central nervous system (brain and spinal cord) has differentiated into a plate of cells. The flat plate then rolls up, and its edges fuse into a hollow neural tube that drops into the embryo to develop just underneath what will become the skin of the back. Neural tube formation is normally complete 4 weeks after fertilization. Failure of neural tube fusion leads to permanent developmental defects of the brain, spinal cord, or both. These defects include *anencephaly*, *meningomyelocele* (which is commonly called **spina bifida**), and *encephalocele*. Although there are many heterogeneous causes, about 90% fall into the classification of multifactorial inheritance. Folic acid deficiency is clearly associated with increased frequency of neural tube closure defects.

The birth prevalence of open NTDs varies with factors such as geographic location (higher in the Eastern United States, lower in the West), race (lower in African Americans), ethnicity (higher in Scotch-Irish), family history (higher with prior births of affected individuals), and maternal weight (higher in obese women). An average figure for the United States is 1 open NTD per 1000 pregnancies. Almost all cases of anencephaly and about 95% of meningomyeloceles are open, with no overlying skin, and therefore are in direct communication with the amniotic fluid. Thus the fetal serum proteins normally present in amniotic fluid at low concentrations gain access in large quantities to the amniotic fluid. The increased amniotic fluid AFP concentration leads to increased amounts in the maternal circulation. Thus only open NTDs are detected by maternal serum AFP screening.

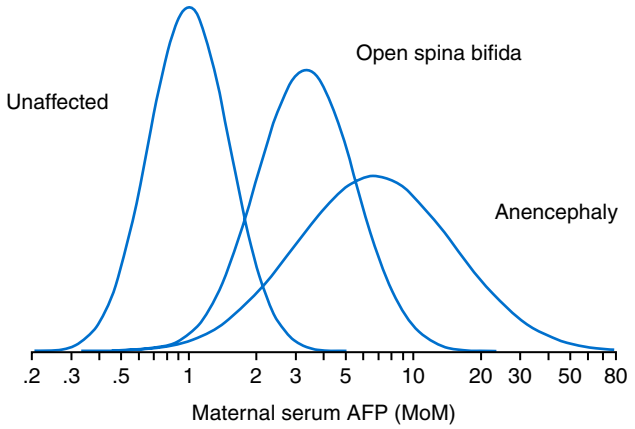


Fig. 44.3 Distribution of maternal serum alpha-fetoprotein *(AFP)* measurements in unaffected pregnancies and in pregnancies affected with open spina bifida or anencephaly. *MoM,* Multiple of the median.

Prenatal Screening

AFP concentrations are increased in amniotic fluid and serum (in the second trimester) of mothers carrying fetuses affected with an open NTD (e.g., anencephaly, open spina bifida). Maternal serum AFP testing thus provides a screening method to identify pregnancies at high risk or to estimate the numerical risk of having a fetus with an open NTD. Because concentrations in serum in affected and unaffected pregnancies overlap considerably, maternal serum AFP is useful as an initial screening test to identify women at high risk for having an affected fetus (Fig. 44.3). These women are then referred for diagnostic procedures (e.g., high resolution ultrasound and, if indicated, amniocentesis for measurement of AFP and acetylcholinesterase in amniotic fluid) to determine if the fetus has an open NTD.

Down Syndrome

Down syndrome is the most common serious disorder of the autosomal chromosomes in liveborn infants, occurring in 1 in 700 live births. An extra copy of chromosome 21 results in a phenotype consisting of moderate to severe intellectual and developmental disabilities, hypotonia, congenital heart defects, and flat facial profile. Most often, an affected child has three copies of chromosome 21 (i.e., trisomy 21), but 5% of cases are caused by translocations and 1% of cases are mosaics. A woman's risk increases slowly up to age 30 and then steadily increases between ages 30 and 45 to a plateau.

Prenatal Screening

Maternal serum AFP concentrations are about 25% lower in Down syndrome than in unaffected pregnancies. The association of AFP and Down syndrome is independent of maternal age. The independence of maternal serum AFP measurements and maternal age established that it was possible to offer a screening method to younger women, in whom most cases of Down syndrome occur.

Unconjugated estriol (uE_3), a product of the fetoplacental unit, is also about 25% lower in pregnancies with Down syndrome. In contrast, concentrations of hCG and inhibin A are about twice as high. A screening method that combines maternal age with measurements of these four analytes permits determination of a single Down syndrome risk estimate. This "quadruple test" provides approximately 80% detection at a 5% false positive rate. Table 44.3 illustrates general patterns of first- and second-trimester analytes associated with various disorders.

Trisomy 18

Trisomy 18 (Edwards syndrome) is caused by a nondisjunction event during meiosis that results in a fetus having an extra copy of chromosome 18. Although it occurs in only 1 in 8000 births, it is probably the most common chromosome defect at the time of conception. The dramatic change in prevalence is due to the very high fetal loss rate both before 8 weeks (>80%) and during the second and third trimesters (~70%). Following birth, half of the infants die within the first 5 days and 90% die within 100 days.

Prenatal Screening

Prenatal screening studies have found that fetal trisomy 18 has a distinctive triple-marker pattern that is different from the Down syndrome pattern (see Table 44.3). In trisomy 18 pregnancies, the AFP, uE_3, and hCG concentrations are low. A method for trisomy 18 risk calculation is available and is often included in prenatal screening programs.

Other Aneuploidies

Although chromosome disorders other than trisomies 21 and 18 are not part of routine screening, the marker patterns exhibited in maternal serum for some aneuploidies are similar. In particular, hydropic Turner syndrome and **triploidy** of paternal origin have serum marker patterns resembling that of Down syndrome pregnancy and are sometimes identified as high risk using this algorithm. Similarly, Turner syndrome without hydrops and triploidy of maternal origin are sometimes identified by the trisomy 18 risk algorithm. The pattern for trisomy 13 is variable (see Table 44.3).

Newer Screening Algorithms

Following extensive use of screening tests in the second trimester, new tests have been introduced that use serum collected in the first trimester.

First Trimester Combined Test

For patients seeking early diagnosis, screening for Down syndrome in the first trimester (10 to 13 weeks' gestation) is available. First trimester screening involves measurement of maternal serum **pregnancy-associated plasma protein-A (PAPP-A)** concentration, which is relatively low in Down syndrome pregnancy, and hCG (free beta subunit or total) concentration, which is relatively high in Down syndrome. A third marker used to improve first trimester screening performance is ultrasound measurement of **nuchal translucency (NT)**, the subcutaneous space between the skin and cervical spine, which is increased in fetuses with Down syndrome. A combination of NT and serum tests (the combined test) was comparable or slightly better than the second-trimester quadruple test for Down syndrome screening, detecting 85% of cases at a 5% false-positive rate.

TABLE 44.3 Conditions Associated With Various Maternal Serum Screening Result Patterns

	SECOND TRIMESTER				FIRST TRIMESTER		
Condition	**AFP**	**hCG**	**uE_3**	**InhA**	**PAPP-A**	**hCG**	**NT**
Amniocentesis	Normal to high	—	—	—	—	—	—
Anencephaly	Very high	—	Very low	—	—	—	—
Congenital nephrosis, duodenal atresia, encephalocele, esophageal atresia, gastroschisis, hydrocephalus, Meckel syndrome, **omphalocele**, sacrococcygeal teratoma	High	—	—	—	—	—	—
Cystic hygroma	High	—	—	—	—	—	Increased
Down syndrome	Low	High	Low	High	Low	High	Increased
Fetal blood contamination	High to very high	Unchanged	Unchanged	Unchanged	—	—	—
Molar pregnancy	Very low	Very high	Very low	Normal	—	—	—
Molar pregnancy (partial)	Low to normal	Very high	Low to normal	Normal	—	—	—
Myelomeningocele (open spina bifida)	High	—	—	—	—	—	—
Normal pregnancy	Low, normal, or high	Low, normal, or high	Low, normal, or high	Low, normal, or high	—	—	—
Overestimated gestational age	Low	High	Low	Normal	—	—	—
Preeclampsia	Normal to high	High	—	—	—	—	—
Pseudocyesis (imaginary pregnancy)	Undetectable	Undetectable	Undetectable	Undetectable	—	—	—
Smith-Lemli-Opitz syndrome	Low	Low	Very low	—	—	—	—
Spontaneous or impending pregnancy loss	Variable	Low or high	Low	Low or high	—	—	—
Steroid sulfatase deficiency (fetal)	Unchanged	Unchanged	Very low	Unchanged	—	—	—
Triploidy (paternal)	Variable	High	Low	High	Variable	Very high	—
Triploidy (maternal)	Variable	Low	Low	Low	Variable	Very low	—
Trisomy 13	Variable	Variable	Variable	Variable	Very low	Low	Increased
Trisomy 18	Low	Low	Very low	Normal	Very low	Very low	Increased
Turner syndrome without hydrops	Low	Low	Very low	Low	Variable	Variable	Increased
Turner syndrome with hydrops	Low	High	Very low	High	Variable	Variable	Increased
Twins and other multiple gestations	High	High	High	High	High	High	—
Underestimated gestational age	High	Low	High	Normal	—	—	—

AFP, α-Fetoprotein; *hCG*, human chorionic gonadotropin; *inhA*, inhibin A; *NT*, nuchal translucency; *PAPP-A*, pregnancy-associated plasma protein A; *uE_3*, unconjugated estriol.

The pattern of markers in trisomy 18 affected pregnancies consists of an increased NT measurement and reduced PAPP-A and hCG in the first trimester. Estimates for the performance of first trimester combined testing in the detection of trisomy 18 are varied, but detection rates have been suggested to be about 80% for a 0.3% false positive rate.

Adding First and Second Trimester Markers Into a Single Integrated Screening Test

The integrated test takes advantage of first- and second-trimester markers and avoids most of the limitations of stand-alone first-trimester screening (Fig. 44.4). With this approach, measurements of NT and PAPP-A are made in the first trimester but are not interpreted or acted upon until testing in the second trimester is complete. In the second trimester, a quadruple test is performed on maternal serum. Results from all six tests (NT, PAPP-A, AFP, uE_3, hCG, and inhibin A) are then combined into a single risk estimate. This approach detects 85% of Down syndrome cases with only a 1% false-positive rate.

The integrated screening approach can be offered effectively and at a lower cost by using a modified test based only on the serum markers (serum integrated). Maternal serum

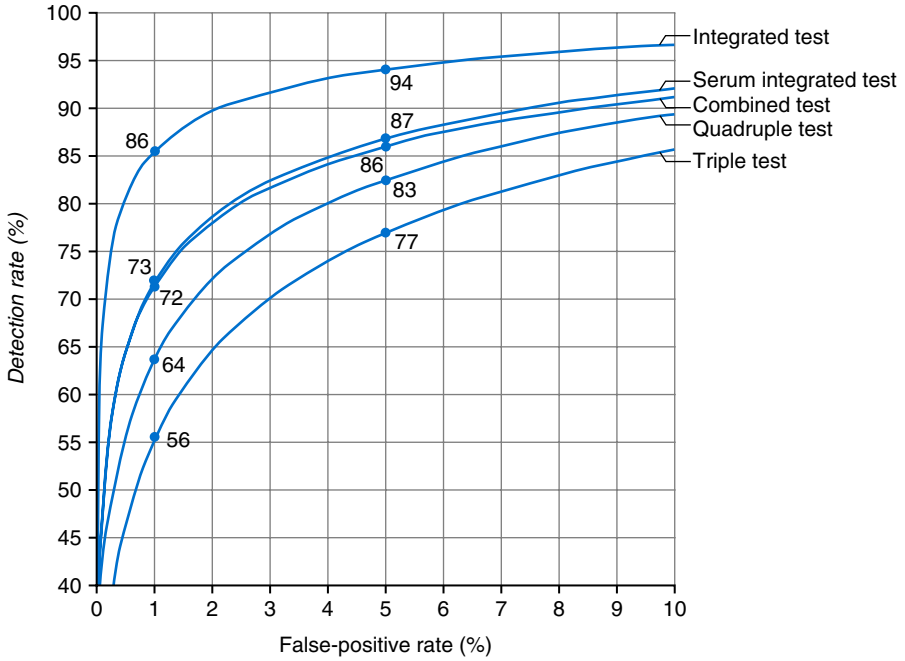


Fig. 44.4 Receiver operating characteristic curves for tests used to screen for Down syndrome. (Reprinted with permission from Wald, N. J., Rodeck, C., Hackshaw, A. K., Rudnicka, A. [2005]. SURUSS in perspective. *Semin Perinatol*, *29*, 225–235.)

PAPP-A concentrations are measured at 10 to 13 weeks and combined with a second trimester quad test for a five-marker screening panel. The serum integrated test gives an 85% detection of Down syndrome cases with about a 3% false-positive rate. A summary of detection and false positive rates for second-trimester and integrated screening tests is illustrated in Fig. 44.4.

Adjustments for Factors That Influence Analyte Measurements

Prenatal screening for both open NTDs and Down syndrome is optimized when each woman's analyte concentration is compared with those of other women (the reference group) who are "similar" to her in many respects. In addition to gestational age, this "similarity" extends to other factors that have been shown to affect analyte concentrations, including (1) maternal weight, (2) race, (3) insulin-dependent diabetes (IDD), and (4) multiple pregnancy. Taking into account these factors increases the accuracy of the interpretation.

Pregnancies Achieved by Assisted Reproductive Technologies

The use of **assisted reproductive technologies (ART)**, such as **in vitro fertilization (IVF)**, for conception is increasing. Women who achieve pregnancy by IVF are about twice as likely to have a positive result after second-trimester Down syndrome screening as are women who achieve pregnancy spontaneously. Thus MoM values of certain analytes are adjusted to restore an appropriate screen positive rate. In particular, concentrations of uE_3 tend to be reduced, while hCG and inhibin A are increased in pregnancies achieved by IVF. Pregnancies achieved by intrauterine insemination, with or without ovulation induction, show a similar trend in marker concentrations.

POINTS TO REMEMBER

Prenatal Screening

- Screening of maternal serum for AFP, hCG, uE_3, and inhibin A together with nuchal translucency measurement is useful to identify women with increased risk for having a fetus with anomalies such as a neural tube defect, Down syndrome, and trisomy 18.
- Down syndrome is associated with increased nuchal thickness and decreased concentrations of AFP and uE_3 and increased concentrations hCG and inhibin-A in maternal serum.
- Increased risk for neural tube defects is associated with increased serum AFP.
- Trisomy 18 is associated with decreased concentrations of AFP, estriol, uE_3, and hCG.

ANALYTICAL METHODOLOGY

In this section of the chapter, methods for measurement of (1) hCG, (2) AFP, (3) uE_3, (4) inhibin A, (5) PAPP-A, and (6) tests for FLM are reviewed.

Human Chorionic Gonadotropin

Measurement of hCG is used to (1) detect pregnancy and its abnormalities (e.g., ectopic and molar pregnancies), (2) screen for Down syndrome and trisomy 18, and (3) monitor the course of a patient with a hCG-producing cancer. Because hCG has these diverse uses, many assays are used, including qualitative urine and serum devices and quantitative serum assays.

Biochemistry

hCG is a glycoprotein containing a protein core with branched carbohydrate side chains that usually terminate with sialic acid. The hormone is a heterodimer composed of two nonidentical, noncovalently bound glycoprotein subunits—alpha (α) and beta (β). When the hCG dimer is dissociated, the hormone activity is lost. However, a major part of the original activity is restored by equimolar recombination of the two subunits.

hCG is synthesized in the placenta. A single gene located on chromosome 6 encodes the α subunit of all four glycoprotein hormones (TSH, LH, FSH, and hCG). Chromosome 19 contains a family of seven genes that encode the hCGβ subunit, though only three appear to be active. Separate messenger RNAs are transcribed from the respective genes, and the α and β subunits are translated from each. The subunits spontaneously combine in the rough endoplasmic reticulum and are then continuously secreted into the maternal circulation.

In the first weeks of pregnancy, hCG stimulates the corpus luteum in the ovary to make progesterone, a steroid that prevents menses and thus facilitates pregnancy. hCG binds weakly to TSH receptors in the maternal thyroid, and extremely increased hCG concentrations have the potential to be thyrotropic.

Methods

Qualitative human chorionic gonadotropin tests. Numerous tests for the qualitative detection of hCG in urine or serum are available as over-the-counter (home) or point-of-care (POC) devices, and their use for the rapid identification of pregnancy is well established. Qualitative detection of hCG in urine is a test granted waived status under the Clinical Laboratory Improvement Amendments of 1988. In contrast, the use of serum as a sample matrix is considered a moderately complex test, even when the test device itself is approved for use with urine or serum, due to the requirement that centrifugation of the specimen is necessary to obtain serum. POC devices are single-use tests that utilize immunochromatography for the rapid qualitative detection of hCG when its concentration exceeds a detection threshold, frequently 10 to 25 IU/L. First-morning specimens are preferred for qualitative urine pregnancy tests because they are concentrated and contain abundant hCG.

Quantitative human chorionic gonadotropin tests. Commonly used quantitative hCG tests are high-performance immunometric assays designed to measure hCG over a wide range of concentrations. Upper limits of detection vary from 400 to 15,000 IU/L. Thus specimen dilution is frequently required to obtain an absolute measurement. The lowest detectable concentration of these assays is from 1 to 2 IU/L.

The measurement of hCG is complicated by its molecular heterogeneity, and considerable variation in measured hCG concentrations is observed between the different assays. Because of variation between hCG assays, median hCG values calculated for maternal serum screening are not transferable and should be considered method specific. hCG measurements in early pregnancy generally are expressed as international units per liter (IU/L). A typical hCG concentration at 16 weeks is approximately 30,000 IU/L.

Alpha Fetoprotein

Measurement of AFP in maternal serum and amniotic fluid is used extensively for prenatal detection of some serious fetal anomalies. Maternal serum AFP is increased in 85% to 95% of cases of fetal open NTD and is decreased in about 30% of cases of fetal Down syndrome. Additionally, AFP measurements are used in nonpregnant patients for monitoring certain cancers.

Chemistry

AFP is a glycoprotein encoded by a gene located within q11–22 on chromosome 4 that is part of a family of genes that also encodes for albumin and vitamin D–binding protein. The protein is composed of carbohydrate and a single polypeptide chain containing 591 amino acids. The carbohydrate composition varies depending on the organ of synthesis, the length of gestation, and source of the specimen (fetal serum vs. amniotic fluid).

Biochemistry

AFP is produced initially by the fetal yolk sac in small quantities and then by the fetal liver in larger quantities as the yolk sac degenerates. Trace amounts are also produced in the fetal gut and kidneys. Early in embryonic life, this protein has a high concentration in fetal serum, reaching about one-tenth the concentration of albumin. Maximal concentration in the fetal serum (~3 million μg/L) is reached at about 9 weeks' gestation. The concentration then declines steadily to about 20,000 μg/L at term. The increase and decrease in concentration of AFP in the amniotic fluid roughly parallel those in the fetal serum but the concentration is two to three orders of magnitude lower (~15,000 μg/L at 16 weeks' gestation). Amniotic fluid AFP has been measured as early as 8 weeks. It rapidly decreases to a low point at 11 weeks and then increases to reach a second maximum at 13 weeks. The concentration then falls in a log-linear fashion until 25 weeks, when the decline steepens.

AFP is first detectable (~5 μg/L) in maternal serum at about the 10th week of gestation. The concentration increases about 15% per week to a peak of approximately 180 μg/L around 25 weeks. The concentration in maternal serum then subsequently declines slowly until term. After birth, the maternal serum AFP rapidly decreases to less than 2 μg/L. In an infant, serum AFP declines exponentially to reach adult concentrations by the 10th month of life.

Methods

Although AFP was traditionally measured by radioimmunoassay (RIA), newer methods use immunoenzymometric assay (IEMA) or chemiluminescent immunoassay (CIA) because of their (1) lower detection limits, (2) increased precision, (3) speed, (4) avoidance of radioactivity, and (5) ease of automation.

Amniotic fluid AFP is measured using the same immunoassays as for maternal serum AFP after a suitable dilution (usually 1:50 to 1:200). AFP in amniotic fluid is less stable than in serum, and leaving samples at room temperature for prolonged periods results in degradation of amniotic fluid AFP. The presence of fetal blood in amniotic fluid samples has been known to increase AFP results, and laboratories should note the presence of blood on the report. In the event of an increased amniotic fluid AFP result (usually >2.0 or 2.5 MoM), the laboratory should test for the presence of fetal blood.

Measurement of acetylcholinesterase (AChE) is a confirmatory test for samples with increased amniotic fluid AFP. Cerebrospinal fluid contains high concentrations of the neural enzyme AChE, and in cases of fetal open NTDs, fluid leaks from the open lesion and allows AChE to enter the amniotic fluid.

Unconjugated Estriol

Any disruption in the biosynthetic pathway will lead to very low maternal serum uE_3. Conditions that cause disruption include (1) fetal anencephaly, (2) placental sulfatase deficiency, (3) fetal death, (4) chromosome abnormalities, (5) molar pregnancy, and (6) Smith-Lemli-Opitz syndrome (SLOS). SLOS is a serious, rare birth defect that is the result of an inborn error in cholesterol metabolism, 7-dehydrosterol-7-reductase deficiency. Down syndrome leads to a modest decrease in uE_3. This steroid, rather than total estriol (unconjugated plus conjugated estriol), is the most specific of the estrogens for identifying a fetus with Trisomy 21.

Chemistry

Estriol is an estrogen with hydroxyl groups at positions 3, 16, and 17. Although present in nonpregnant patients in very low concentrations, during late pregnancy this estrogen predominates. Only a minor amount (~10%) of the hormone circulates in plasma unconjugated, and this form is strongly bound to sex hormone binding globulin (SHBG) because of its low solubility. The majority exist as conjugates of glucuronate and sulfate. Conjugation, which occurs in the maternal liver, makes the hormone more soluble and permits renal clearance.

Biochemistry

Estriol is produced in very large amounts during the last trimester of pregnancy. The biosynthetic pathway requires three organs to be fully functioning: fetal adrenal, fetal liver, and placenta. The fetal adrenal avidly binds low-density lipoprotein to take in cholesterol, which is converted to two major steroid intermediates: pregnenolone sulfate and dehydroepiandrosterone sulfate (DHEA-S). These intermediates are secreted into the fetal circulation. The fetal liver, possessing 16α-hydroxylase, converts DHEA-S to 16α-hydroxy-DHEA-S, which is secreted back into the fetal circulation. Finally, the placenta synthesizes estriol from 16α-hydroxy-DHEA-S. Approximately 90% of maternal serum estriol is derived from this fetal-placental pathway. A minor amount is made using precursors from the maternal ovary. Concentrations typical for the second trimester of pregnancy are 0.70 to 2.50 μg/L (2.43 to 8.68 nmol/L), depending on the assay used.

Methods

The determination of uE_3 is made difficult by its low concentration. Values obtained with various uE_3 assays vary widely as judged using the CAP proficiency testing Survey FP. Conversion to MoM reduces the between-method differences, but uE_3 is still the most variable of the screening analytes.

Of the four analytes currently used for screening, uE_3 is the least stable, and consequently requirements for collection, storage, and shipment are dictated by this analyte. The uE_3 concentration increases in blood at room temperature and at 4°C because the conjugated forms are able to spontaneously deconjugate to form the parent hormone. Therefore collected blood should be allowed to clot, and serum should be removed promptly. uE_3 is stable in serum for up to 7 days at 2°C to 4°C. The concentration of uE_3 increases when sera has been stored for longer than 4 days at room temperature.

Inhibin A

Inhibins are members of the transforming growth factor-β (TGFβ) superfamily of proteins. Inhibin is a negative feedback regulator of FSH secretion in both males and females. The placenta produces large quantities of inhibin A that completely suppress FSH. In addition to the usefulness of inhibin A as a predictor of Down syndrome risk, inhibin A and B measurements have found additional applications as tumor markers for ovarian cancer and in the evaluation of male infertility.

Chemistry

Inhibins are proteins consisting of dimers of dissimilar subunits (α and β) linked by disulfide bridges. The mature form of inhibin is produced by cleavage of larger precursor forms. In follicular fluid and serum, mature inhibins, precursors of inhibins, and intermediate molecules of varying molecular weight are present. Another group of related molecules, the activins, are dimers consisting of just the β-subunits. Inhibin A is the only form within the inhibin/activin family of proteins that provides sufficient discrimination to be useful for use in Down syndrome screening.

Biochemistry

In the reproductive system, inhibin and activin subunits are expressed in the placenta, as well as in the granulosa cells of the ovary and by the Sertoli cells of the testis. Inhibin A and inhibin B have distinctive serum profiles during the human menstrual cycle. Inhibin A rises in the follicular phase and peaks at ovulation, before decreasing to basal amounts in the luteal phase. In postmenopausal women, the concentrations of both forms of inhibin are nondetectable.

Inhibin A is produced by the fetoplacental unit beginning in early pregnancy. Inhibin A concentrations exhibit a

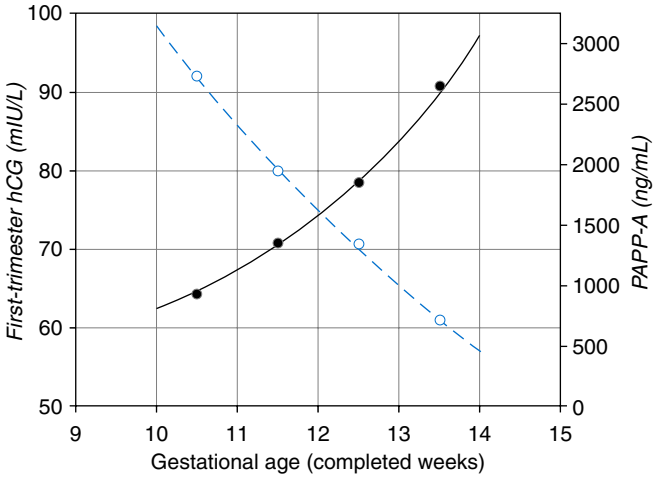


Fig. 44.5 Typical median values for maternal serum pregnancy–associated plasma protein-A *(PAPP-A)* and total human chorionic gonadotropin *(hCG)* in the first trimester. The observed average gestational age is plotted on the horizontal axis and the median of that week's PAPP-A *(black solid line)* or total hCG *(blue dotted line)* measurements on the vertical logarithmic axis.

complex pattern during the course of pregnancy, rising to a peak at 8 to 10 weeks' gestation, declining to a minimum at 17 weeks of pregnancy, and then resuming to slowly increase at term. Unlike the other screening tests, average inhibin concentrations change relatively little from 15 to 20 weeks' gestation. Values in the second trimester range from 50 to 400 ng/L, with typical levels at 17 weeks' gestation reaching a nadir of about 175 ng/L (175 pg/mL).

Methods

Inhibin assays used for Down syndrome screening must measure only dimeric inhibin A and not the free α subunits and the precursors of higher molecular weight, which also circulate in blood. Highly specific assays using monoclonal antibodies are available that measure only dimeric inhibin A. Specific inhibin A assays provide better screening performance than the nonspecific total inhibin assays.

Pregnancy-Associated Plasma Protein A

PAPP-A concentrations in the maternal serum during the first trimester of pregnancy are associated with fetal and placental health and development. Low PAPP-A concentrations early in pregnancy have been associated with (1) Down syndrome, (2) a high rate of fetal loss, (3) poor fetal growth (intrauterine growth restriction [IUGR]), (4) premature delivery, (5) hypertension, and (6) preeclampsia.

Biochemistry

PAPP-A is a zinc-containing metalloproteinase glycoprotein. It is expressed at low concentrations in all tissues, but high amounts of PAPP-A protein and mRNA are localized in placental tissues throughout gestation. Circulating PAPP-A is part of a larger molecular complex that includes two subunits of PAPP-A covalently bound to two subunits of pro major basic protein (pro MBP), forming a heterotetramer. PAPP-A immunoreactivity is found in syncytiotrophoblast cells.

Maternal serum PAPP-A concentrations increase as gestation proceeds to term (Fig. 44.5). Concentrations of PAPP-A are critical to normal fetal growth because of its role as an insulin-like growth factor binding protein (IGFBP) protease. PAPP-A regulates the action of insulin-like growth factor II (IGF-II) by cleaving its binding protein, primarily IGFBP-4, thereby increasing bioavailable forms.

Methods

PAPP-A immunoassays are available on routine clinical chemistry platforms. There are significant between-method differences, and efforts to achieve standardization are currently underway. A small increase in PAPP-A concentrations has been reported for collection in plastic versus glass tubes. First-trimester serum PAPP-A concentrations are stable in serum at 4°C for 1 week or longer, depending on the assay used for testing.

Fetal Fibronectin

Determining the risk of preterm labor would help clinicians manage those at high risk more aggressively, thereby lowering the incidence of preterm delivery. **Fetal fibronectin (fFN)** concentration in cervical and vaginal secretions was proposed as a test to aid in predicting preterm delivery. While patients with a negative test result have only a ~1% chance of delivering in 1 week, the positive predictive value of fFN is low. Thus fFN is of limited utility in predicting preterm birth within 7 days of sample collection. fFN may be measured by immunoassay of cervical or vaginal mucus collected with a Dacron polyester swab.

Placental Alpha Microglobulin-1

Placental alpha microglobulin-1 (PAMG-1) is a placental glycoprotein. During pregnancy, PAMG-1 is secreted into the amniotic fluid, where it is present at a high concentration (2000 to 25,000 ng/mL) relative to that of maternal blood (5 to 25 ng/mL) or cervicovaginal fluid with intact membranes (0.05 to 2 ng/mL). A test for PAMG-1 that exploits these large differences in concentrations has been developed for clinical use as an aid for the detection of PROM.

A rapid immunochromatographic method is commercially available for the rapid detection of PAMG-1 in cervicovaginal fluid. The specimen is collected using a polyester swab, and the fluid is eluted off the swab by rinsing it in a vial containing a buffer solution. A test strip is then placed into the buffer for 10 minutes. The test result is determined by visual inspection of a test and a control line on the lateral flow device. The analytical limit of detection of the test is 5 ng/mL.

This test is more sensitive and specific for the detection PROM than clinical assessment or the nitrazine or ferning tests. Contamination with large amounts of blood will cause a false-positive result. False-negative results may occur if the specimen is collected 12 or more hours after a rupture that is subsequently obstructed by the fetus or is resealed.

Fetal Lung Maturity

FLM tests help assess the risk of a fetus to develop RDS. Numerous tests of amniotic fluid for FLM have been developed that exhibit high sensitivity for immaturity and an excellent predictive value of a mature result. Despite performing well, the number of FLM tests performed each year in the United States has declined, and several studies have demonstrated that FLM testing does not improve fetal morbidity or mortality.

In 2008, the ACOG issued a revised practice bulletin concerning FLM. This bulletin discourages FLM testing before 32 weeks of gestation or when delivery is mandated due to fetal or maternal indications and specifies that assessment of FLM is indicated prior to a scheduled delivery at less than 39 weeks of gestation if it cannot be inferred from clinical history. However, the practice of scheduled deliveries at less than 39 weeks of gestation is now strongly discouraged by ACOG, unless delivery is justified by medical or obstetric complications. Thus testing for FLM is becoming infrequent.

Fetal Lung Maturity Assessment

Fetal lung liquids contribute to amniotic fluid, resulting in an exchange of surfactants from the developing lungs to the amniotic fluid. Therefore amniotic fluid is the required specimen for FLM testing, and it is obtained by transabdominal amniocentesis, usually during real-time sonographic visualization. Currently available methods for FLM testing include (1) **lamellar body count (LBC)**, (2) phosphatidylglycerol measurement, and the (3) lecithin/sphingomyelin ratio. Readers are referred to previous editions of this textbook for a detailed discussion of FLM testing methodology.

REVIEW QUESTIONS

1. The most severe form of hemolytic disease of the newborn (HDN) is referred to as:
   a. respiratory distress syndrome (RDS).
   b. ectopic pregnancy.
   c. jaundice.
   d. erythroblastosis fetalis.
2. In a developing fetus, the major blood-forming organ until 24 weeks' gestation is the:
   a. bone marrow.
   b. liver.
   c. lung.
   d. kidney.
3. Preeclampsia is diagnosed by the presence of which findings?
   a. Hypertension and proteinuria ≥300 mg of protein in 24-hour urine
   b. Hypertension and proteinuria ≥100 mg of protein in 24-hour urine
   c. Hypertension and new onset of platelet count >100,000/μL
   d. Hypertension and decreased liver transaminases
4. The composition of amniotic fluid is most typically like that of which one of the following?
   a. Extracellular fluid
   b. Intracellular fluid
   c. Cerebrospinal fluid
   d. Blood
5. Hemolytic disease of the newborn (HDN) is caused by:
   a. lack of folic acid in the maternal diet.
   b. increased formation of fetal hemoglobin (Hb F).
   c. maternal antibodies against fetal erythrocytes.
   d. blockage of placental transfer of immunoglobulin A (IgA) antibodies.
6. What is the function of the hormone hCG (human chorionic gonadotropin) in pregnancy?
   a. It signals the corpus luteum to produce progesterone to maintain pregnancy.
   b. It inhibits the production of estrogen and progesterone by the ovary.
   c. It stimulates the production of estrogen to promote uterine growth.
   d. Its concentration peaks near birth, which helps initiate the birth process.
7. Which of the following best describes changes in concentrations of placental hormones during pregnancy?
   a. Estrogen peaks in the first trimester, and then remains stable throughout gestation.
   b. Estrogen concentration increases throughout gestation.
   c. hCG concentration is stable throughout gestation.
   d. Progesterone concentration declines throughout gestation.
8. The integrated screen for prediction of Down syndrome risk includes testing for:
   a. PAPP-A in the first trimester and hCG, AFP, and uE_3 in the second trimester.
   b. AFP, hCG, uE_3, and inhibin A in the second trimester.
   c. PAPP-A and nuchal translucency in the first trimester and AFP, hCG, uE_3, and inhibin A in the second trimester.
   d. hCG in the first trimester and nuchal translucency, AFP, and PAPP-A in the second trimester.
9. The majority of waste produced by a fetus is in the amniotic fluid. Which one of the following structures is responsible for removal of this waste from the amniotic fluid?
   a. Fetal kidneys
   b. Fetal liver
   c. Fetal lungs
   d. Placenta
10. When a fertilized egg implants in a location other than the body of the uterus, the condition is called:
   a. Down syndrome.
   b. spina bifida.
   c. ectopic pregnancy.
   d. hyperemesis gravidarum.

SUGGESTED READINGS

Yarbrough, M. L., Stout, M., & Gronowski, A. M. (2017). Pregnancy and Its Disorders. In N. Rifai, A. R. Horvath, & C. T. Wittwer (Eds.), *Tietz textbook of clinical chemistry and molecular diagnostics* (6th ed.) (pp. 1655–1696). St. Louis: Elsevier.

American College of Obstetricians and Gynecologists. (2007). ACOG practice bulletin no. 77: Screening for fetal chromosomal abnormalities. *Obstetrics Gynecology*, *109*, 217–227.

Lockitch, G. M. (Ed.). (1993). *Handbook of diagnostic biochemistry and hematology in normal pregnancy*. Boca Raton, FL: CRC Press.

Haddow, J. E., Palomaki, G. E., Knight, G. J., Williams, J., Pulkkinen, A., Canick, J. A., et al. (1992). Prenatal screening for down's syndrome with use of maternal serum markers. *New England Journal of Medicine*, *327*, 588–593.

Workowski, K. A., Berman, S., & Centers for Disease C, Prevention. (2010). Sexually transmitted diseases treatment guidelines, 2010. *MMWR Recommendations and Reports*, *59*, 1–110.

McChesney, R., Wilcox, A. J., O'Connor, J. F., Weinberg, C. R., Baird, D. D., Schlatterer, J. P., et al. (2005). Intact hCG, free hCG beta subunit and hCG beta core fragment: longitudinal patterns in urine during early pregnancy. *Human reproduction*, *20*, 928–935.

Grenache, D. G. (2004). Hemolytic disease of the newborn. In A. M. Gronowski (Ed.), *Handbook of clinical laboratory testing during pregnancy* Vol. (pp. 219–243). Totowa, NJ: Humana Press.

American College of Obstetricians and Gynecologists. (2008). ACOG practice bulletin no. 97: fetal lung maturity. *Obstetrics Gynecology*, *112*, 717–726.

Palmer, O. M., Grenache, D. G., & Gronowski, A. M. (2007). The NACB laboratory medicine practice guidelines for point-of-care reproductive testing. *Point of Care*, *6*, 265–272.

Wald, N. J., Huttly, W. J., & Hackshaw, A. K. (2003). Antenatal screening for Down's syndrome with the quadruple test. *Lancet*, *361*, 835–836.

Yarbrough, M. L., Grenache, D. G., & Gronowski, A. M. (2014). Fetal lung maturity testing: the end of an era. *Biomarkers in medicine*, *8*, 509–515.

Gala, R. (2008). Ectopic pregnancy. In J. S. J. Schorge, L. Halvorson, B. Hoffman, K. Bradshaw, & F. Cunningham (Eds.), *Williams gynecology* Vol. (1st ed.). New York: McGraw-Hill.

Malone, F. D., Canick, J. A., Ball, R. H., Nyberg, D. A., Comstock, C. H., Bukowski, R., et al. (2005). First-trimester or second-trimester screening, or both, for Down's syndrome. *New England Journal of Medicine*, *353*, 2001–2011.

Norton, M. E., Jacobsson, B., Swamy, G. K., Laurent, L. C., Ranzini, A. C., Brar, H., et al. (2015). Cell-free DNA analysis for noninvasive examination of trisomy. *New England Journal of Medicine*, *372*, 1589–1597.

Chiu, R. W. K., & Lo, Y. M. D. (2017). Circulating Nucleic Acid for Prenatal Diagnostics. In N. Rifai, A. R. Horvath, & C. T. Wittwer (Eds.), *Tietz textbook of clinical chemistry and molecular diagnostics* (6th ed.) (pp. 1655–1696). St. Louis: Elsevier.

45

Newborn Screening and Inborn Errors of Metabolism

Marzia Pasquali, Nicola Longo

OBJECTIVES

1. Define the following:
   - Aminoacidopathy
   - Fatty acid oxidation disorder
   - Inborn error of metabolism (IEM)
   - Multiplex analysis
   - Newborn screening
   - Organic acidemia
   - Second-tier testing
   - Tandem mass spectrometry (MS/MS)
   - Urea cycle disorder
   - Lysosomal storage disorder
2. Describe autosomal recessive inheritance patterns, including their significance in inborn errors of metabolism (IEMs) and percent risk of having an affected child in each pregnancy; diagram an autosomal recessive inheritance pattern pedigree.
3. List the six components of a newborn screening program; state the criteria required of any newborn screening program.
4. State the classes of metabolic disorders identifiable by newborn screening and list an example of each class.
5. For the following disorders, list causes and symptoms, treatments, available screening, and second-tier laboratory analyses and results:
   - Galactosemia
   - Glutaric acidemia type I
   - Homocystinuria
   - Maple syrup urine disease (MSUD)
   - Medium-chain acyl-**coenzyme A** (CoA) dehydrogenase (MCAD) deficiency
   - Phenylketonuria (PKU)
   - Tyrosinemia type 1
   - Urea cycle disorders
6. Outline how the results of tandem mass spectrometry are interpreted with regard to inborn errors of metabolism (IEMs); state an advantage of using this type of analysis.
7. Discuss the issue of false positive results in newborn screening, including tests most commonly misinterpreted and the use and principle of second-tier tests.
8. List four types of confirmatory laboratory tests used to verify out of range newborn screening test results.

KEYWORDS AND DEFINITIONS

Aminoacidopathy A disorder of amino acid metabolism caused by defective activity of an enzyme leading to elevated amino acids in blood (and urine).

Argininosuccinate lyase (ASL) deficiency A genetic disorder that affects the body's ability to clear the nitrogen already incorporated into the urea cycle as argininosuccinate. This disorder often presents with rapid-onset hyperammonemia in the newborn period. Also known as *argininosuccinic aciduria.*

Argininosuccinate synthetase (ASS) deficiency A genetic disorder characterized by a deficiency of argininosuccinate synthase and an increase in the ammonia concentration in the blood. Also known as *citrullinemia type I.*

Autosomal recessive disorder A disorder characterized by the requirement that both copies of a gene (called alleles) are abnormal (see following definition for *autosomal recessive inheritance*).

Autosomal recessive inheritance A mendelian inheritance pattern in which (1) nonsex chromosomes (autosomes) carry a DNA sequence (gene) for the inherited trait and (2) two abnormalities (one from each parent) in the DNA sequence must be present for the trait to appear in an individual; heterozygous parents have a 25% chance of having an affected offspring.

Citrullinemia type II An autosomal recessive disease characterized by increased concentrations of citrulline in serum and urine. Ammonia intoxication is another manifestation. This disorder is characterized by jaundice and liver failure early in life; neuropsychiatric symptoms, including abnormal behaviors, loss of memory, seizures, and coma in older individuals.

Dihydropteridine reductase (DHPR) An enzyme that catalyzes the reversible formation of tetrahydrobiopterin (BH4) from dihydrobiopterin; deficiency of BH4 results in malignant hyperphenylalaninemia.

Disorders of amino acid metabolism A group of disorders of amino acid metabolism caused by defective activity of an enzyme leading to increased concentration of amino acids in blood (and urine).

Disorders of carbohydrate metabolism A group of disorders caused by loss of an enzyme in the metabolic pathway of a carbohydrate, leading to increased accumulation of that carbohydrate within organs and tissues.

Disorders of fatty acid oxidation A group of disorders caused by deficiency of an enzyme in the oxidation pathway of fatty acids, leading to inability to use fat as an energy source.

Dried blood spot (DBS) A form of biosampling where blood samples are blotted and dried on filter paper.

Galactosemia An inherited disease in which the transformation of galactose to glucose is blocked, allowing galactose to increase to toxic concentrations in the body.

Glucose-6-phosphate dehydrogenase (G-6-PD) deficiency An X-linked recessive hereditary disease characterized by abnormally low quantities of the enzyme G-6-PD involved in the pentose phosphate pathway, especially important in red blood cell metabolism.

Glutaric acidemia type 1 (GA1) An inherited disorder in the metabolism of lysine and tryptophan causing accumulation of excessive quantities of glutaric acid (GA) and related compounds that can damage the brain.

Glycogen storage diseases Genetic diseases, also known as *glycogenoses* that involve the enzymes regulating glycogen accumulation and breakdown.

Homocystinuria A condition characterized by accumulation of excess homocysteine in blood. This biochemical abnormality has a variety of genetic causes.

Inborn error of metabolism (IEM) Inherited disorder due to deficiency of an enzyme or transporter impairing the transformation of body chemicals.

Long-chain 3-hydroxyacyl-CoA dehydrogenase (LCHAD) deficiency A metabolic disease characterized by defective fatty acid oxidation due to a defect in an enzyme in the β-oxidation pathway.

Lysosomal storage disorder A disorder characterized by impaired degradation of complex macromolecules in the lysosomes leading to accumulation of these substrates in organs and tissues.

Maple syrup urine disease (MSUD) An autosomal recessive metabolic disorder affecting branched-chain amino acids (leucine, isoleucine, and valine). The urine of affected individuals smells like maple syrup.

Medium-chain acyl-CoA dehydrogenase (MCAD) deficiency A disorder of fatty acid oxidation that impairs the body's ability to break down medium-chain fatty acids into acetyl-CoA.

Methylmalonic acidemia An autosomal recessive metabolic disorder resulting from defects in the metabolic pathway in which methylmalonyl-coenzyme A (CoA) is converted into succinyl-CoA by the enzyme methylmalonyl-CoA mutase.

Microcephaly A condition characterized by an abnormally small head due to failure of brain growth.

Multiplex analysis Simultaneous assessment of multiple analytes in a single sample.

Organic acidemia A disorder of intermediary metabolism in which lack of an enzyme leads to buildup of an organic acid (from deamination of the amino acid) as opposed to the buildup of the parent amino acid.

Peroxisomal disorders A group of conditions caused by impaired function of peroxisomes resulting from either impaired peroxisomes biogenesis or impaired activity of one of the peroxisomal enzymes.

Phenylketonuria (PKU) A disorder characterized by the accumulation of phenylalanine in blood caused by deficiency of phenylalanine hydroxylase activity leading to production of phenylketones that are excreted in urine.

Propionic acidemia An organic acid disorder due to deficiency of an enzyme involved in amino acid and fatty acid catabolism. It is characterized by excess of propionic acid in blood and urine.

Tandem mass spectrometry (MS/MS) An analytical technique where two mass spectrometers are connected in series. It involves separation and identification of substances and chemicals based on their mass to charge *(m/z)* ratio.

Tyrosinemia A genetic disorder involving the metabolism of the tyrosine. Affected individuals have high concentrations of tyrosine in blood (hypertyrosinemia).

Urea cycle disorder Any disorder in which the body is unable to excrete waste nitrogen–ammonia, resulting in mental and behavioral dysfunction, coma, and death.

Inborn errors of metabolism (IEM) are genetically determined disorders affecting an individual's ability to convert nutrients or to use them for energy production. They are caused by the impaired function of (1) enzymes, (2) transporters, or (3) cofactors, and result in accumulation of abnormal metabolites (substrates) proximal to the metabolic block or by lack of necessary products (path A to D, Fig. 45.1). Abnormal by-products can also be generated when alternative pathways are used to dispose of the excess metabolites (path A to F, see Fig. 45.1).

Typically IEMs present in the newborn period or in infancy. Some diseases, however, such as fatty acid oxidation defects or milder variants of classic metabolic disorders, may not be detected until adulthood. Despite the long asymptomatic period, their consequences are still devastating and may result in death. Therefore it is critical to identify and treat these diseases before irreversible damage occurs. The frequency of individual diseases is rare, varying from 1:10,000 to 1:200,000 or even rarer. Their cumulative frequency, however, is substantial and approaches a frequency of 1:2,000.

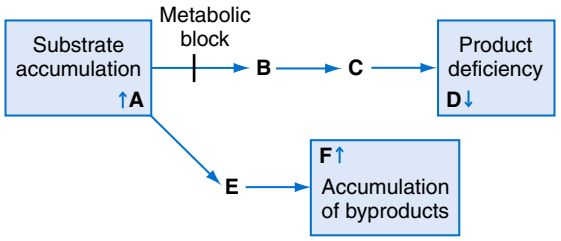


Fig. 45.1 General schematic of a metabolic pathway. The substrate *A* is converted by a series of reactions into product *D*. If one of the enzymes *(arrows)* is defective (metabolic block), the substrate of the reaction will accumulate (*A* in this case) and can enter alternative pathways of metabolism, leading to the formation of byproducts (*E* and *F* in this case). At the same time, the concentration of the product of the reaction *(D)* will decrease.

POINTS TO REMEMBER

- IEMs are disorders affecting the conversion of nutrients into energy.
- IEMs usually present in the newborn period or in early infancy, although there are cases that can become symptomatic in adulthood.
- The majority of IEMs are inherited as autosomal recessive traits.
- Symptoms vary and include vomiting, lethargy, delays in development, seizures, and coma leading to death.
- For some IEMs symptoms are triggered by environmental stress, such as fasting, infections, exercise.
- The diagnosis requires analysis of specific metabolites (amino acids, carnitine, acylcarnitines, organic acids), measurement of enzyme activity, and DNA testing.
- Treatment is available for many IEMs, and it is effective if it is initiated early.
- Newborn screening can identify many of these disorders pre-symptomatically.

INHERITANCE PATTERN OF METABOLIC DISORDERS

Metabolic disorders are caused by mutations in genes that code for specific enzymes or transporters involved in metabolic pathways. The majority of metabolic disorders have **autosomal recessive inheritance**, in which affected individuals have a mutation in both alleles encoding for a specific enzyme or transporter (Fig. 45.2) and girls are affected as often as boys. In the vast majority of cases, the parents of children with one of these metabolic conditions are carriers of the condition—that is, they carry one normal allele and one mutant allele, and they do not show clinical signs of the condition. In each singleton pregnancy when both parents are carriers, there is (1) a 25% chance that the child is affected, (2) a 50% chance that the child is a carrier like the parents, and (3) a 25% chance that the child has two normal alleles (see Fig. 45.2).

Clinical Presentation of Metabolic Disorders

The medical consequences of IEMs vary from failure to thrive to acute illness leading in some cases to (1) brain damage, (2) coma, and (3) death. In many cases, the acute presentation is

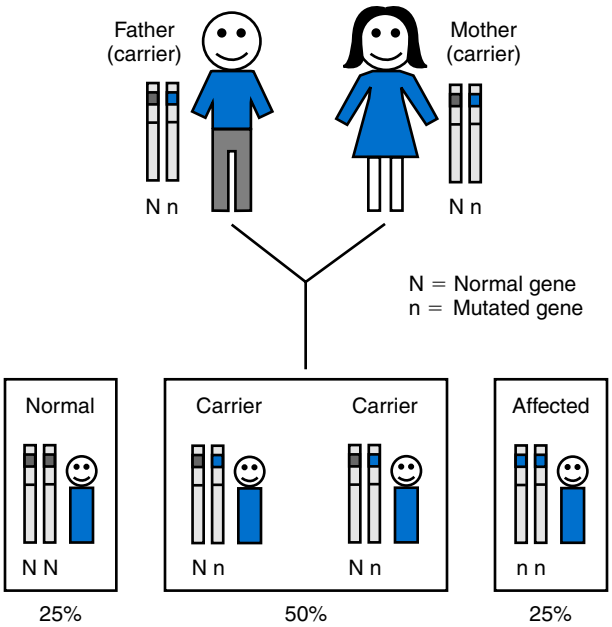


Fig. 45.2 Autosomal recessive inheritance pattern.

preceded by a symptom-free period that is variable in length depending on the specific disease. In many cases, there is a treatment available for these disorders consisting of special diets (formulas) that do not contain the specific nutrients that the patients are unable to metabolize. The treatment is effective if begun early before symptoms occur, but damage that has already occurred is usually irreversible. Thus the ideal time for identifying patients with metabolic disorders is at birth or earlier.

Biochemical Diagnosis

The biochemical diagnosis of IEMs and treatment monitoring involve analysis of (1) metabolites, (2) enzymatic activity, and/or (3) DNA sequence. Because of technological advances, such as **tandem mass spectrometry (MS/MS)**, which allows for the simultaneous detection of multiple analytes, many IEMs are now included in newborn screening programs.

NEWBORN SCREENING

Newborn screening is a public health activity aimed at early identification of conditions for which timely intervention is expected to result in elimination or reduction of (1) morbidity, (2) mortality, and (3) disabilities.

Background

Newborn screening was originally instituted in the 1960s as a result of the pioneering efforts of Robert Guthrie and his colleagues, who developed a screening assay for measuring the phenylalanine content of a **dried blood spot (DBS)** collected from the blood of newborn babies on filter paper.

Criteria of Newborn Screening Programs

The disease to be screened must be serious, must be fairly common, must have a natural history that is understood, and must have helpful treatment and genetic counseling (in

the case of genetic disease) available. In addition, the screening test must be acceptable to the public, reliable, valid, and affordable. The newborn screening program requires the availability of expedient diagnosis and treatment of the disease and effective communication of results. Newborn screening programs must be effectual public health approaches to the diagnosis of treatable disorders early in life. Because newborn screening is a program, its efficacy depends on the integration and collaboration among its different components, involving (1) public health, (2) screening and diagnostic laboratories, (3) physicians, and (4) families of affected children.

After the efforts of the American College of Medical Genetics and Genomics to uniform newborn screening practices in the United States, the newborn screening panel is periodically updated by the Advisory Committee on Heritable Disorders in Newborns and Children, and it is known as the recommended uniform screening panel (RUSP). States then decide if, when, and how they will implement the recommendations of the committee. Current recommendations for the uniform screening panel of core conditions (https://www.hrsa.gov/advisorycommittees/mchbadvisory/heritabledisorders/recommendedpanel/) include at least (1) five fatty acid oxidation disorders, (2) nine organic acidemias, (3) six aminoacidopathies, (4) two endocrinopathies (congenital hypothyroidism and congenital adrenal hyperplasia), (5) three hemoglobinopathies/thalassemias, (6) two lysosomal storage disorders (Pompe disease and Hurler syndrome), and (6) seven other disorders (Table 45.1), two of which (hearing loss and critical congenital heart disease) are tested directly at the birth facility. In addition, many secondary conditions can be identified using MS/MS and are routinely reported by most newborn screening programs (Table 45.2).

Advances in therapeutic interventions for IEMs are continuously expanding the role of newborn screening. Of note, newborn screening does not identify all metabolic disorders, and some patients with these disorders can be missed. Therefore a symptomatic patient at any age should be investigated despite normal newborn screening results.

Newborn screening is accomplished using MS/MS, which as a multiplex methodology allowing the measurement of multiple metabolites, which identifies the largest number of diseases, and other screening methods, such as (1) enzyme assays, (2) immunoassays, (3) electrophoresis, and (4) DNA tests.

Steps to be Followed in a Newborn Screening Program

The different steps that need to be coordinated and tracked in a newborn screening program are:

1. Screening (sample collection and delivery, laboratory testing)
2. Follow-up (complete demographic information, satisfactory specimens, abnormal screening results, access to testing and clinical care)
3. Diagnosis (confirmatory tests, clinical consultation)
4. Clinical management (primary care physician, specialist physician, genetic counselor, dietitian)
5. Education (healthcare professionals, parents)
6. Quality assurance: analytical (proficiency testing, quality controls, standards), efficiency of follow-up system, efficacy of treatment, long-term outcome

Each of these components should have specifically-written protocols that deal directly with the performance of the tasks involved. To assist laboratories in writing such protocols and procedures, the Clinical and Laboratory Standards Institute has published guidelines for newborn Screening, including guidelines on (1) blood collection on filter paper (NBS01-A6), (2) laboratory testing by tandem mass spectrometry (NBS04), (3) follow-up activities (NBS02-A2), and (4) document preparation (QMS02-A6). These documents describe the basic principles, scope, and range of activities within a newborn screening program and how to prepare the prerequisite documents. They are available at https://clsi.org/standards/products/newborn-screening/documents/.

Second-Tier Testing

One of the major pitfalls in newborn screening is the number of false-positive results associated with certain screening tests. To reduce the number of infants requiring additional and unnecessary confirmatory testing, second-tier tests have been developed. These second-tier tests involve further analysis of the same blood spot with an abnormal result targeting different, more specific analytes and, often, using a methodology different from the one used in the primary screening test. They include (1) DNA analysis for cystic fibrosis (see Chapter 49) (2) MS/MS steroid profiling for congenital adrenal hyperplasia (see Chapter 41), and (3) metabolic disorders. Second-tier testing has become a crucial component of newborn screening programs to increase the sensitivity and the specificity of the screening test.

INBORN ERRORS OF METABOLISM

In this section, disorders of metabolism included in the RUSP are discussed in general, with selected individual disorders of each class reviewed in more detail as examples. An online database containing a catalog of human genetic disorders is found at www.ncbi.nlm.nih.gov/omim/. This database is entitled "Online Mendelian Inheritance in Man" (OMIM) and is a comprehensive and authoritative compendium of human genes and genetic phenotypes. In this database, individual disorders are assigned a specific OMIM number.

Disorders of Amino Acid Metabolism

Disorders of amino acid metabolism collectively affect approximately 1 in 8000 newborns. Almost all are transmitted as autosomal recessive traits and result from a lack of a specific enzyme in the metabolic pathway of an amino acid, leading either to the accumulation of (1) the parent amino acid, (2) its by-products, or (3) the catabolic products (organic acids), depending on the location of the enzyme block. Disorders of amino acid metabolism are divided into two groups: (1) **aminoacidopathies**, in which the parent amino acid accumulates in excess in blood and spills over into urine; and (2) **organic acidemias**, in which

TABLE 45.1 Uniform Screening Panel Recommended by the Secretary of the Department of Health and Human Services for States to Screen as Part of Their State Universal Newborn Screening Programs: Core Conditions[a]

		METABOLIC DISORDER					
ACMG Code	**Core Condition**	**Organic Acid Condition**	**Fatty Acid Oxidation Disorders**	**Amino Acid Disorders**	**Endocrine Disorder**	**Hemoglobin Disorder**	**Other Disorder**
PROP	Propionic acidemia	X					
MUT	Methylmalonic acidemia (methylmalonyl-CoA mutase)	X					
Cbl A,B	Methylmalonic acidemia (cobalamin disorders)	X					
IVA	Isovaleric acidemia	X					
3-MCC	3-Methylcrotonyl-CoA carboxylase deficiency	X					
HMG	3-Hydroxy-3-methylglutaric aciduria	X					
MCD	Holocarboxylase synthase deficiency	X					
βKT	β-Ketothiolase deficiency	X					
GA1	Glutaric acidemia type I	X					
CUD	Carnitine uptake defect/carnitine transport defect		X				
MCAD	Medium-chain acyl-CoA dehydrogenase deficiency		X				
VLCAD	Very long-chain acyl-CoA dehydrogenase deficiency		X				
LCHAD	Long-chain L-3 hydroxyacyl-CoA dehydrogenase deficiency		X				
TFP	Trifunctional protein deficiency		X				
ASA	Argininosuccinic aciduria			X			
CIT	Citrullinemia, type I			X			
MSUD	Maple syrup urine disease			X			
HCY	Homocystinuria			X			
PKU	Classic phenylketonuria			X			
TYR I	Tyrosinemia, type I			X			
CH	Primary congenital hypothyroidism				X		
CAH	Congenital adrenal hyperplasia				X		
Hb SS	S,S disease (Sickle cell anemia)					X	
Hb S/βTh	S, Beta-thalassemia					X	
Hb S/C	S,C disease					X	
BIOT	Biotinidase deficiency						X
CCHD	Critical congenital heart disease						X
CF	Cystic fibrosis						X
GALT	Classic galactosemia						X
GSD II	Glycogen storage disease Type II (Pompe)						X
HEAR	Hearing loss						X
SCID	Severe combined Immunodeficiencies						X
MPS I	Mucopolysaccharidosis type 1						X
X-ALD	X-linked adrenoleukodystrophy						X

[a]Disorders that should be included in every Newborn Screening Program. Selection of conditions based upon American College of Medical Genetics. (2006). Newborn screening: towards a uniform screening panel and system. *Genetic Med*, *8*(5, Suppl), S12–S252, as authored by the ACMG and commissioned by the Health Resources and Services Administration. Nomenclature for Conditions based upon Sweetman, L., et al. (2006). Naming and counting disorders (conditions) included in newborn screening panels. *Pediatrics*, *117*(5, Suppl), S308–S314. <https://www.hrsa.gov/advisorycommittees/mchbadvisory/heritabledisorders/recommendedpanel/>.

CoA, Coenzyme.

TABLE 45.2 Uniform Screening Panel: Secondary Conditions as Recommended by the Advisory Committee on Heritable Disorders in Newborns and Children[a]

		METABOLIC DISORDER				
ACMG Code	Secondary Condition	Organic Acid Condition	Fatty Acid Oxidation Disorders	Amino Acid Disorders	Hemoglobin Disorder	Other Disorder
Cbl C,D	Methylmalonic acidemia with homocystinuria	X				
MAL	Malonic acidemia	X				
IBG	Isobutyrylglycinuria	X				
2MBG	2-Methylbutyrylglycinuria	X				
3MGA	3-Methylglutaconic aciduria	X				
2M3HBA	2-Methyl-3-hydroxybutyric aciduria	X				
SCAD	Short-chain acyl-CoA dehydrogenase deficiency		X			
M/SCHAD	Medium/short-chain L-3-hydroxyacyl-CoA dehydrogenase deficiency		X			
GA2	Glutaric acidemia type II		X			
MCAT	Medium-chain ketoacyl-CoA thiolase deficiency		X			
DE RED	2,4-Dienoyl-CoA reductase deficiency		X			
CPT IA	Carnitine palmitoyltransferase type I deficiency		X			
CPT II	Carnitine palmitoyltransferase type II deficiency		X			
CACT	Carnitine acylcarnitine translocase deficiency		X			
ARG	Argininemia			X		
CIT II	Citrullinemia, type II			X		
MET	Hypermethioninemia			X		
H-PHE	Benign hyperphenylalaninemia			X		
BIOPT (BS)	Biopterin defect in cofactor biosynthesis			X		
BIOPT (REG)	Biopterin defect in cofactor regeneration			X		
TYR II	Tyrosinemia, type II			X		
TYR III	Tyrosinemia, type III			X		
Var Hb	Various other hemoglobinopathies				X	
GALE	Galactoepimerase deficiency					X
GALK	Galactokinase deficiency					X
	T-cell-related lymphocyte deficiencies					X

[a]These secondary conditions are detected in the differential diagnosis of the core disorders listed in Table 45.1. Selection of conditions based upon American College of Medical Genetics. (2006). Newborn screening: towards a uniform screening panel and system. *Genetic Med*, *8*(5, Suppl), S12–S252 as authored by the ACMG and commissioned by the Health Resources and Services Administration. Nomenclature for Conditions based upon Sweetman, L., et al. (2006). Naming and counting disorders (conditions) included in newborn screening panels. *Pediatrics*, *117*(5, Suppl): S308–S314. <https://www.hrsa.gov/advisorycommittees/mchbadvisory/heritabledisorders/recommendedpanel/>.

intermediary products (organic acids) in the catabolic pathway of certain amino acids accumulate.

An example of an aminoacidopathy is **phenylketonuria (PKU)**, a disorder of phenylalanine metabolism caused in the majority of cases by deficiency of phenylalanine hydroxylase, the enzyme responsible for the conversion of phenylalanine to tyrosine. Other examples include (1) **maple syrup urine disease (MSUD)**, (2) **homocystinuria**, and (3) **tyrosinemia type1**.

Glutaric acidemia type 1 (GA1), a disorder of lysine and tryptophan metabolism, is an example of an **organic acidemia**. Other organic acidemias include (1) isovaleric acidemia, (2) **methylmalonic acidemia**, and (3) **propionic acidemia** (see Table 45.1). The clinical manifestations of the organic acidemias vary from no observable clinical consequences to neonatal mortality. Conditions, such as (1) developmental retardation, (2) seizures, (3) alterations in sensorium, (4) failure to thrive or (5) behavioral disturbances, occur in more than half the disorders. Metabolic ketoacidosis, often accompanied by hyperammonemia, is a frequent finding in organic acidemias. The compound(s) accumulated

depend on the (1) site of the enzymatic block, (2) reversibility of the reactions proximal to the lesion, and (3) availability of alternative pathways of metabolic "runoff."

Phenylketonuria

PKU (OMIM #261600) is a disorder of phenylalanine metabolism that results from impaired phenylalanine hydroxylase activity leading to the accumulation of phenylalanine, lack of tyrosine, and production of phenylketones that are excreted in urine.

Phenylalanine is an essential amino acid, constituting 4% to 6% of all dietary protein. Phenylalanine not used for protein synthesis is converted to tyrosine by the enzyme phenylalanine hydroxylase and further degraded via a ketogenic pathway (Fig. 45.3). The frequency of hyperphenylalaninemia/PKU is 1:16,500 live births. The majority (98%) of cases of PKU are caused by mutations in the phenylalanine hydroxylase gene, with the remaining 2% being caused by defects in biosynthesis or recycling of tetrahydrobiopterin (BH_4), the cofactor for phenylalanine hydroxylase, or by a defect in chaperone protein.

Primary or secondary (due to a deficiency of the cofactor/chaperone) impairment of phenylalanine hydroxylase causes (1) increased phenylalanine, (2) increased phenylketones, and (3) deficiency of tyrosine (see Fig. 45.3). The greatly increased concentration of phenylalanine impairs brain development and function, affecting other organs minimally. Patients with classic PKU are clinically asymptomatic at birth, with developmental delays and neurological manifestations typically becoming evident at several months of life when brain damage has already occurred. Untreated PKU patients develop (1) **microcephaly**, (2) eczematous skin rash, (3) "mousy" odor (due to accumulation of phenylacetate), and (4) severe mental retardation. The treatment of PKU includes a diet (1) low in protein and phenylalanine and (2) supplemented with tyrosine, minerals, vitamins, and other nutrients to sustain healthy growth. Some patients respond to treatment with pharmacological amounts of a synthetic form of tetrahydrobiopterin. Treatment should be continued for life.

Newborn screening leads to early detection and therapy of patients with PKU with prevention of mental retardation. Ideally treatment should start before 2 weeks of age. Pregnant women with PKU should adhere to a strictly controlled diet that is low in phenylalanine and proteins, because phenylalanine is teratogenic and results in an increased risk of spontaneous abortions or of having a child with (1) growth retardation, (2) microcephaly, (3) significant developmental delays, and (4) birth defects.

The diagnosis of PKU is confirmed biochemically by demonstration of increased plasma phenylalanine and an increased phenylalanine: tyrosine ratio. Urine specimens from affected individuals contain increased concentrations of phenylketones (hence the name *phenylketonuria*). Enzymatic confirmation of phenylalanine hydroxylase deficiency is not usually performed (the enzyme is expressed only in the liver), but mutational analysis of the gene is increasingly used because there is a correlation between severity of the mutation and phenylalanine

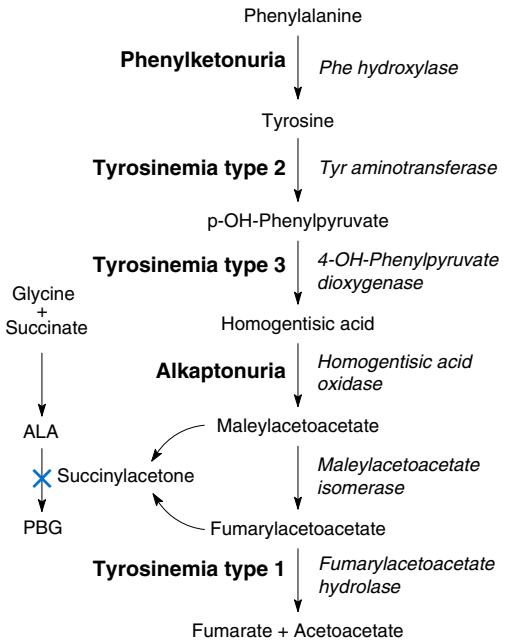


Fig. 45.3 Inborn errors of metabolism (IEMs) resulting from deficiency of key enzymes required in the catabolism of phenylalanine. Specific disorders caused by enzyme deficiency are shown on the left side of the figure. Phenylalanine is an essential amino acid obtained from the hydrolysis of almost all protein part of a normal diet. It is transformed to tyrosine by phenylalanine hydroxylase, the enzyme defective in phenylketonuria. The amino group of tyrosine is then removed by tyrosine hydroxylase, the enzyme defective in tyrosinemia type 2, to become *p*-hydroxyphenylpyruvate. *p*-Hydroxyphenylpyruvate dioxygenase (the enzyme defective in tyrosinemia type 3) adds oxygen to generate homogentisic acid. Homogentisic acid oxidase generates maleylacetoacetate, and a deficiency of this enzyme causes alkaptonuria, a condition in which urine turns black after standing at room temperature or with the addition of alkali. Maleylacetoacetate isomerase generates fumarylacetoacetate; a deficiency of this enzyme has not been reported in humans. Fumarylacetoacetate is very unstable and becomes succinylacetone, a very toxic compound capable of causing cancer and cell death. Succinylacetone also inhibits delta-aminolevulinic acid synthase, the enzyme that converts delta-aminolevulinic acid (ALA) into porphobilinogen (PBG). Accumulation of delta-aminolevulinic acid causes severe pain and neurological symptoms. Succinylacetone accumulates in tyrosinemia type 1 caused by deficiency of fumarylacetoacetate hydrolase. This enzyme normally generates fumarate and acetate that can enter the citric acid cycle for the production of energy, water, and carbonic anhydride.

tolerance, and it can exclude deficiency of the chaperone. All children with hyperphenylalaninemia should be screened for defects in BH_4 synthesis or recycling. This is performed by measuring the urinary pterin profile and by measuring the enzyme activity of **dihydropteridine reductase (DHPR)** in blood spotted on filter paper. Deficiency of BH_4 affects the synthesis of several neurotransmitters, including dopamine and serotonin because it is a cofactor of (1) phenylalanine hydroxylase, (2) tyrosine hydroxylase, (3) tryptophan hydroxylase, and (4) nitric oxide synthase. Patients with a defect in BH_4 synthesis or recycling have neurological symptoms and developmental regression in the first few months of life, despite adequate control of phenylalanine intake and plasma concentrations. They can develop seizures and have a characteristic hypotonia

of the trunk with hypertonia of the extremities. These patients require therapy with BH_4 and appropriate neurotransmitters. They may or may not require a low phenylalanine diet once BH_4 therapy is initiated.

Tyrosinemia

Tyrosinemia is characterized by increased blood concentrations of tyrosine. There are three types of tyrosinemia (I–III); each is caused by the deficiency of a different enzyme (Fig. 45.3). Tyrosinemia type I (TYR-I) is the most severe form and is caused by a deficiency of fumarylacetoacetate hydrolase. Tyrosinemia type II is caused by a deficiency of tyrosine aminotransferase, and tyrosinemia type III is caused by a deficiency of 4-hydroxyphenylpyruvate dioxygenase.

The incidence of TYR-I is approximately 1 in 100,000 with a clustering of cases in the Lac-St. Jean region of Quebec (Canada). Patients with TYR-I present, usually before 6 months of age, with severe liver involvement, or with (1) failure to thrive, (2) liver dysfunction, and (3) rickets due to renal Fanconi syndrome. They have extreme irritability caused by peripheral neuropathy mimicking acute intermittent porphyria. Untreated patients develop liver cirrhosis and are at very high risk for liver cancer.

Patients with TYR-I have increased concentrations of tyrosine in the plasma, but this increase usually is not as marked as in patients with other forms of tyrosinemia. Increased tyrosine can also be seen in (1) other types of tyrosinemia (type II and III), (2) transient tyrosinemia of the newborn, (3) prematurity, (4) liver disease from any cause, (5) mitochondrial DNA depletion syndrome, and (6) diets very rich in proteins. The biochemical diagnosis of TYR-I is based on the detection in urine of succinylacetone, derived from fumarylacetoacetic acid, the intermediate immediately upstream of the enzyme defect. TYR-1 is identified by newborn screening only when succinylacetone is used as the primary marker, because tyrosine might not be elevated in the newborn period in these patients. Therapy consists of (1) low dietary tyrosine, (2) low dietary phenylalanine, and (3) drug therapy with 2-(2-nitro-4-trifluoro-methylbenzoyl)-1,3-cyclohexanedione (NTBC), an inhibitor of 4-hydroxyphenylpyruvate dioxygenase (see Fig. 45.3). NTBC prevents the synthesis of succinylacetone, which is rapidly reduced within the normal range after initiation of treatment. Measurement of alpha fetoprotein is also used to monitor these patients, because liver cancer is a complication of this condition. Liver transplantation is indicated in patients who progress to liver failure and in those acquiring liver cancer.

Homocystinuria

Homocystinuria is characterized by increased concentrations of the sulfur-containing amino acid, homocysteine, in blood and urine (see Fig. 45.4). It can be caused by defects of methionine metabolism (classic homocystinuria) or defects in homocysteine remethylation. Classic homocystinuria is caused by reduced activity of cystathionine β-synthase. The incidence is approximately 1:450,000 live births in the United States with a very high incidence in the country of Qatar (1:1800). Clinical manifestations are nonspecific at first and may include failure to thrive and developmental delay. Patients usually develop lens dislocation (often requiring surgery) and a body habitus like that seen in Marfan syndrome. (Homocysteine interferes with disulfide formation in fibrillin, the protein defective in Marfan syndrome.) Patients whose blood homocysteine concentration remains increased are at risk of blood clots, which are a life-threatening complication of the condition.

The biochemical diagnosis of homocystinuria is obtained by plasma amino acid analysis showing increased plasma concentrations of methionine (especially in children) and the presence of the disulfide homocystine. Classic homocystinuria is detected in newborn screening by increased concentrations of methionine in whole blood spots, although often the concentration of methionine is not high in the newborn screening period and this condition could be missed. Therapy for classic homocystinuria requires (1) high doses of pyridoxine (the cofactor of cystathionine β-synthase), (2) a special diet low in methionine, and (3) administration of betaine that donates a methyl group to homocysteine to generate methionine.

Maple Syrup Urine Disease

Maple syrup urine disease (MSUD) is an **autosomal recessive disorder** with an incidence of approximately 1:200,000 live births. It is caused by a deficiency of the branched-chain alpha-keto acid dehydrogenase complex (BCKDC), leading to a buildup of the branched-chain amino acids leucine, isoleucine, and valine and their toxic ketoacids in the blood and urine. The disease is named for the presence of sweet-smelling urine with an odor similar to that of maple syrup. Several forms of this disease may occur, depending on the gene affected and the severity of the mutations. The enzyme complex consists of four subunits designated $E_1\alpha$, $E_1\beta$, E_2, and E_3. MSUD results from mutations in any of the genes that code for the enzyme subunits.

Individuals with classic MSUD present with poor feeding and vomiting during the first week of life followed by lethargy and coma within a few days. This usually follows a normal birth and an uneventful first few days of life. Routine laboratory work is mostly unremarkable except for the presence of ketones in urine. Diagnosis is established by measuring plasma amino acids and finding increased branched chain amino acids and the presence of alloisoleucine, which is characteristic of this disease.

Cornerstones of treatment include diets that have a restricted content of branched-chain amino acids and include supplementation with (1) high-dose thiamine and, in many patients, (2) valine and (3) isoleucine.

Urea Cycle Disorders

A **urea cycle disorder**, or urea cycle defect, is caused by a deficiency of one of the enzymes or a transporter in the urea cycle, which is responsible for removing ammonia from the blood stream. The urea cycle involves a series of biochemical steps in which nitrogen derived from protein metabolism is

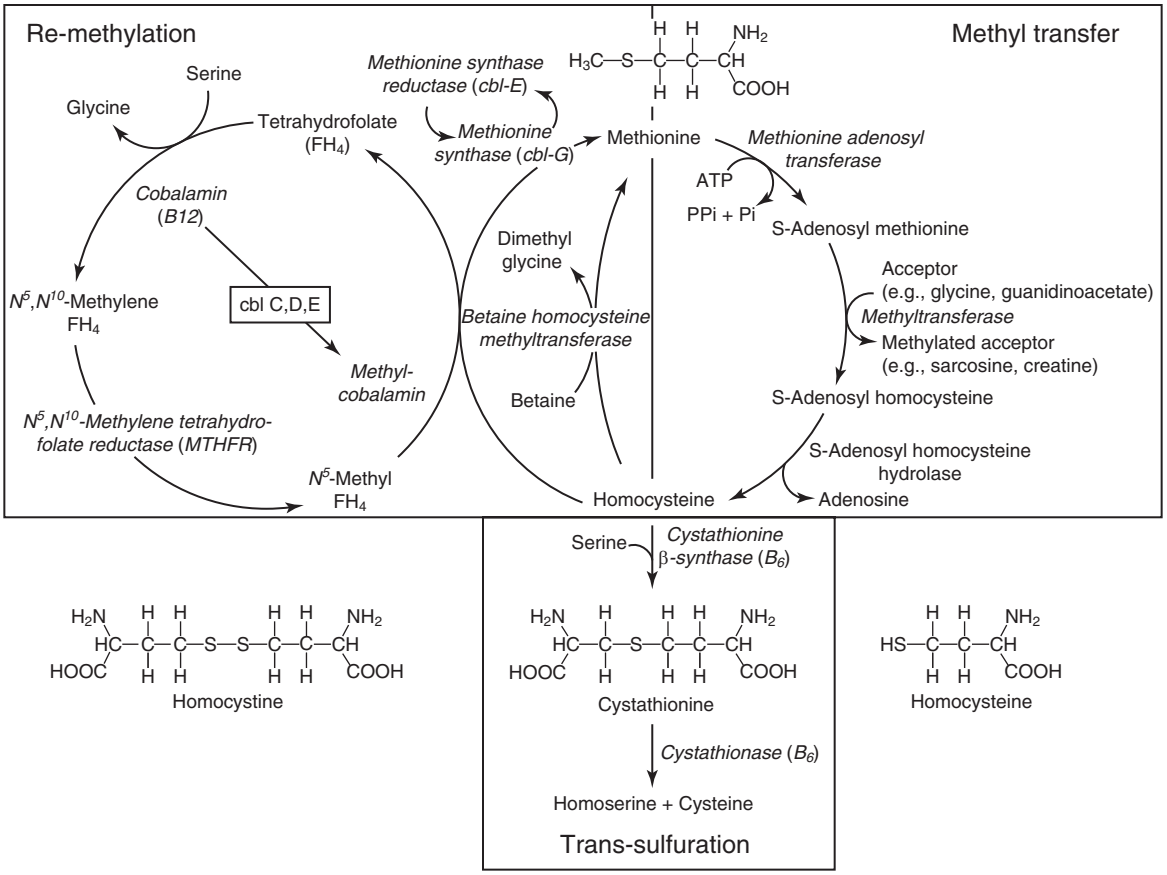


Fig. 45.4 Sulfur amino acid metabolism. Defects in methionine metabolism (classic homocystinuria caused by cystathionine beta synthase deficiency) or defects in homocysteine remethylation (defects in vitamin B_{12} or folic acid metabolism) cause homocystinuria. Methionine transfers a methyl group during its conversion to homocysteine. Defects in methyl transfer or in the subsequent metabolism of homocysteine by the pyridoxal phosphate (vitamin B_6)–dependent cystathionine β-synthase increase plasma methionine concentrations. Excess homocysteine is transformed back to methionine via remethylation. This reaction is catalyzed by methionine synthase and requires methylcobalamin and folic acid. Deficiencies in these enzymes or lack of vitamin B_{12} (cobalamin) or folic acid or defects in their metabolism (defects in vitamin B_{12} metabolism are named cblC, cblD, cblF; methylenetetrahydrofolate reductase [MTHFR] deficiency is the most common defect of folic acid metabolism) are associated with decreased or normal methionine concentrations and with elevated homocysteine. In an alternative pathway, homocysteine is remethylated by betaine:homocysteine methyltransferase.

removed from the blood and converted to urea, which is then excreted in urine. In urea cycle disorders, the nitrogen accumulates in the form of ammonia, a toxic substance, and is not removed from the body.

Urea cycle disorders can result in (1) mental and behavioral dysfunction, (2) coma, and (3) death. The urea cycle disposes of the nitrogen groups of amino acids before their carbon skeleton is metabolized to gluconeogenic (most amino acids) or ketogenic precursors (leucine and lysine), or both (isoleucine, phenylalanine, tyrosine, and tryptophan). This cycle requires the combined action of different enzymes and mitochondrial transporters (Fig. 45.5). Deficiency of any of these enzymes or transporters impairs the function of the urea cycle, causing hyperammonemia.

Newborn screening can effectively identify **argininosuccinate synthetase (ASS) deficiency** (or citrullinemia type I), **citrullinemia type II** (mitochondrial aspartate transporter/citrin deficiency), **argininosuccinate lyase (ASL) deficiency**, and arginase deficiency. OTC (ornithine transcarbamylase) deficiency, an X-linked disorder that is also the most common of the urea cycle defects, CPS-1 (carbamylphosphate synthase 1) and NAGS (N-acetylglutamate synthase) deficiencies are not effectively detected by NBS. Patients with urea cycle defects may have symptoms at any age. In the neonatal period, there is usually a brief interval between birth and clinical manifestations, with the most severe cases having symptoms before the results of newborn screening are available. Hyperammonemia and the accumulation of glutamine in the brain lead to (1) poor feeding, (2) vomiting, (3) lethargy, or (4) irritability progressing to coma and death.

Diagnosis is accomplished by measuring plasma amino acids and identifying the compound present in excess. Patients are treated with a diet low in proteins and administration of nitrogen scavengers, such as sodium benzoate and phenylacetate or its precursor phenylbutyrate, which bind and remove glycine and glutamine, respectively.

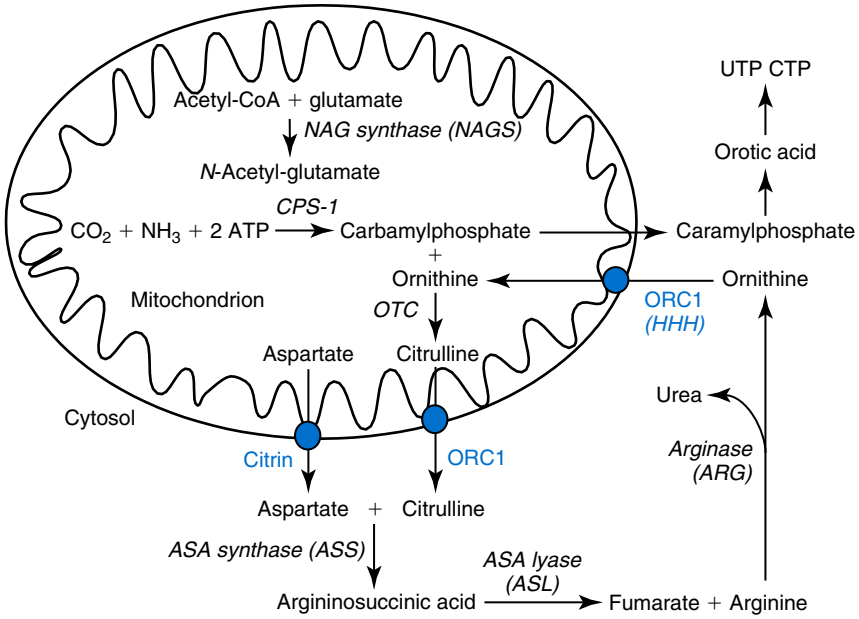


Fig. 45.5 The urea cycle. In this cycle, urea is formed starting from ammonia *(NH_3)*. It requires several enzymes (in italics) and mitochondrial transporters *(blue)*, any of which can be defective and may impair the function of the urea cycle. Within the mitochondrial matrix, ammonia, derived from amino acid catabolism, is combined with carbonic anhydride and phosphorus to generate carbamoylphosphate by carbamyl phosphate synthase 1 *(CPS-1)*. This enzyme is active only in the presence of *N*-acetylglutamate, an allosteric activator synthesized by *N*-acetylglutamate synthase *(NAGS)*. Carbamoylphosphate is then attached to ornithine by ornithine transcarbamylase *(OTC)* to generate citrulline. Citrulline then exits the mitochondrial matrix using the ornithine/citrulline mitochondrial transporter *(ORC1)*. This transporter is an exchanger and couples the exit of citrulline from mitochondria to the entry of ornithine. Citrulline in the cytoplasm combines with aspartate (transported from the mitochondrial matrix by the citrin transporter) to generate argininosuccinic acid by the action of a synthase *(ASS)*. Argininosuccinic acid is then split to arginine and fumarate by a lyase *(ASL)*. Fumarate enters the citric acid cycle, and arginine is converted to urea (excreted in urine) and ornithine by arginase *(ARG)*. Ornithine can then enter the mitochondrial matrix again to restart another cycle. The two nitrogens in urea excreted in urine derive from ammonia and the amino group of aspartate. Carbamoylphosphate not utilized in the urea cycle (usually because of deficiency in ornithine transcarbamylase or any of the downstream enzymes) exits to the cytoplasm to generate nucleotides (UTP and CTP). The concentration of orotic acid (an intermediate in nucleotide synthesis) can be measured in the urine of patients with OTC, ASS, ASL or ARG deficiency. *citrin,* Aspartate/glutamate exchanger; *CP,* carbamoylphosphate; *CPS-1,* carbamyl phosphate synthase 1; *CTP,* cytidine triphosphate; *UTP,* uridine triphosphate.

Glutaric Acidemia Type I

Glutaric acidemia type 1 (GA1; OMIM #231670) is an autosomal recessive disorder of the metabolism of (1) lysine, (2) hydroxylysine, and (3) tryptophan; it is caused by deficiency of glutaryl-**coenzyme A** (CoA) dehydrogenase. In this condition, glutaric acid (GA) and 3-hydroxyglutaric acid (3-OH-GA), formed in the catabolic pathway of the above mentioned amino acids, accumulate, especially in urine. Affected patients have brain atrophy and macrocephaly (head circumference often increases dramatically following birth), and they can develop acute dystonia secondary to degeneration of the corpus striatum (a component of the motor system in the brain). In most cases this is triggered by an infection with fever between 6 and 18 months of age.

GA1 is identified by an increased concentration of glutarylcarnitine (C5DC) on newborn screening. The diagnosis is biochemically confirmed by urine organic acid analysis that indicates the presence of excess 3-OH-GA, with plasma and urine acylcarnitine profiles showing glutarylcarnitine as the major peak. Therapy consists of (1) carnitine supplementation to remove GA, (2) a diet restricted in amino acid precursors of GA, and (3) prompt administration of intravenous calories in the patient who is unable to eat for any reason such as (1) infections, (2) fever, and (3) gastroenteritis. Early diagnosis and therapy reduce the risk of acute dystonia in patients with GAI.

Treatment of Organic Acidemias and Aminoacidopathies

As with the treatment regimens for aminoacidopathies, therapy for organic acidemias consists of (1) special diets restricting the compounds (usually amino acids) that result in the formation of the abnormal organic acid, (2) supplementation with vitamins specific for each disorder, (3) carnitine supplements, and (4) sometimes avoidance of fasting. For some of these conditions, aggressive therapy of illnesses with intravenous fluids containing glucose is essential to avoid catabolism that aggravates clinical symptoms.

POINTS TO REMEMBER

Disorders of Amino Acid Metabolism

- They result from impaired activity of an enzyme involved in the metabolic pathway of an amino acid.
- The impaired enzyme activity leads to the accumulation of the parent amino acid, its by-products, or its catabolic products (organic acids).
- In aminoacidopathies the parent amino acid accumulates in blood; in organic acidemias the intermediary products in the catabolic pathway of amino acids accumulate and are excreted in the urine.
- Phenylketonuria, tyrosinemia type I, maple syrup urine disease, and homocystinuria are examples of aminoacidopathies.
- Glutaric acidemia type 1 is an example of organic acidemia.
- The treatment consists of a special diet that restricts the offending amino acid(s) and includes supplements of cofactors and vitamins, including carnitine for organic acidemias. Fasting should also be avoided.

Disorders of Fatty Acid Oxidation

Fatty acids are metabolized within mitochondria to produce energy. Carnitine and the carnitine cycle are required to transfer long-chain fatty acids into mitochondria for subsequent beta-oxidation (Fig. 45.6). Within the mitochondrial matrix, long-chain fatty acids are progressively shortened by two carbon units at each cycle to generate acetyl-CoA, which is used by the Krebs cycle to produce energy (Fig. 45.7). **Disorders of fatty acid oxidation**, such as **medium-chain acyl-CoA dehydrogenase (MCAD) deficiency**, occur when an enzyme is missing in the metabolic pathway and fatty acids fail to undergo oxidation to supply energy. These disorders are usually silent and become evident only when the body needs energy from fat during times of increased energy demand, such as (1) fasting, (2) infections/fever, or (3) strenuous exercise. In these cases, apparently healthy children with these disorders (1) become acutely sick, (2) lose consciousness, (3) become comatose, and (4) may die. When symptomatic, patients with fatty acid oxidation disorders develop hypoglycemia and might show increased serum transaminases indicating liver damage. Some fatty acid oxidation disorders (such as **long-chain 3-hydroxyacyl-CoA dehydrogenase [LCHAD] deficiency**) also (1) affect skeletal muscle, (2) affect cardiac muscle, and (3) affect the mother during pregnancy. Other disorders include carnitine transporter defects and very long-chain acyl-CoA dehydrogenase deficiency (see Table 45.1).

Medium-Chain Acyl-CoA Dehydrogenase Deficiency

MCAD (OMIM #201450) deficiency is the most common disorder of fatty acid oxidation with a frequency of about 1:18,000 in the United States. The clinical presentations of this disease are (1) asymptomatic, (2) hypoglycemia, (3) lethargy, (4) coma, and (5) sudden death, which is usually triggered by prolonged fasting, acute illness, or both. The majority of patients have symptoms in the first year of life, but clinical symptoms can occur at any time during life, and often the first episode is fatal. The treatment consists of (1) avoidance of fasting, (2) consumption of low-fat foods after 1 year of age, (3) carnitine supplementation, and (4) institution of an emergency plan in case of illness or other metabolic stress. Early diagnosis through newborn screening and early initiation of treatment leads to a favorable prognosis. Patients with MCAD deficiency are identified by MS/MS newborn screening because of the characteristic acylcarnitine profile with increased concentrations of (1) C6- (hexanoyl), (2) C8- (octanoyl), and (3) C10:1- (decenoyl) carnitine (with C8-carnitine being the highest) and increased C8/C2 and C8/C10 ratios.

The diagnosis of MCAD is confirmed biochemically by urine organic acid and acylglycine analyses showing (1) excess hexanoylglycine and suberylglycine, (2) plasma acylcarnitine profile (confirming increased C6-, C8-, and C10:1 carnitine), and (3) DNA analysis. A common pathogenic variant (c.985A>G, p.Lys329Glu) is prevalent in symptomatic patients (80% of symptomatic patients are homozygous for this mutation; 98% carry at least one copy).

Treatment of Fatty Acid Oxidation Disorders

Treatment of fatty acid oxidation disorders consists of (1) avoidance of fasting, (2) adherence to a low-fat diet, and (3) carnitine supplementation. For some disorders of fatty acid oxidation, therapy includes administration of medium chain triglycerides (oil) that enter mitochondria independently from carnitine and bypass the metabolic block. In addition, conditions that increase catabolism such as (1) fever, (2) vomiting, and (3) infections need to be aggressively treated with antibiotics (if necessary), antipyretics, and intravenous glucose.

POINTS TO REMEMBER

Disorders of Fatty Acid Oxidation

- Fatty acids are metabolized in mitochondria via beta-oxidation to produce energy.
- Most disorders of fatty acid oxidation are triggered by environmental stress (fasting, infections, exercise) when energy demands increase.
- The clinical presentation includes hypoketotic hypoglycemia, cardiomyopathy, sudden death.
- The diagnosis relies on plasma acylcarnitines and urine organic acids analyses, in addition to DNA testing.
- The treatment consists of avoidance of fasting, low-fat diet, in addition to special types of fat and carnitine supplementation for some of these disorders.

Lysosomal Storage Disorders

Lysosomal storage disorders are a group of diseases caused by deficiency of one or more lysosomal enzymes responsible for the degradation of complex macromolecules, such as mucopolysaccharides. The impaired degradation of these macromolecules results in their accumulation in organs and tissues. The disease phenotype depends on the substrate accumulated and its sites of turnover—for example, muscle and heart in Pompe disease and connective tissue in

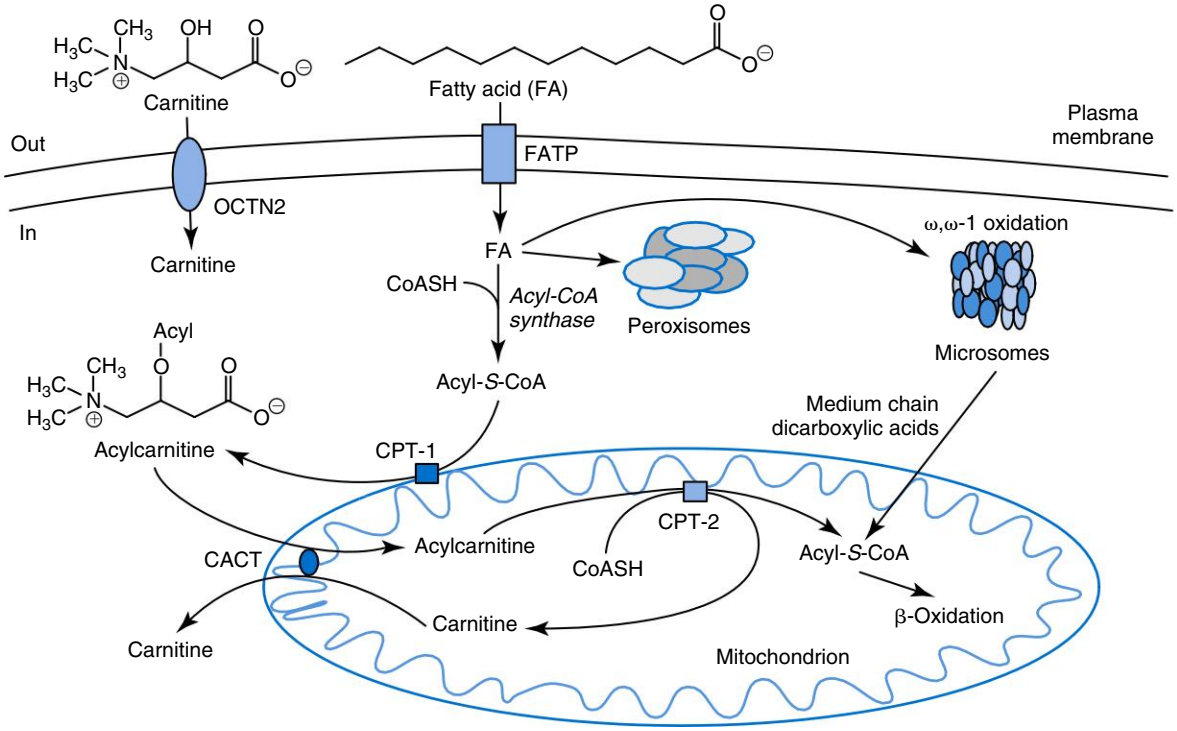


Fig. 45.6 The carnitine cycle in fatty acid *(FA)* oxidation. The carnitine cycle is responsible for delivering long-chain fatty acid to the mitochondrial matrix for subsequent beta oxidation. FAs bound to albumin are transferred across the plasma membrane by the action of fatty acid transport proteins *(FATP)*. Inside the cell, fatty acids undergo vectorial acylation, a process catalyzed by acyl-**coenzyme A** (CoA) synthases, that traps them in the cytoplasm as acyl-CoA thioesters (acyl-*S*-CoA). The acyl-CoA thioesters are then conveyed through different metabolic pathways in mitochondria, peroxisomes, and microsomes based on the cell energy status. Metabolism of fatty acid by peroxisomes and microsomes is usually minimal, except in disorders of fatty acid oxidation. The mitochondrial membrane is impermeable to acyl-CoAs and fatty acids must be conjugated to carnitine to enter mitochondria. Carnitine is accumulated inside the cell by the high-affinity organic cation transporter novel 2 (OCTN2) carnitine transporter. Carnitine forms a high-energy ester bond with long-chain fatty acids by the action of carnitine palmitoyl transferase 1 *(CPT-1)*, located in the inner aspect of the outer mitochondrial membrane, with the formation of acylcarnitines. Acylcarnitines are then translocated across the inner mitochondrial membrane by the carnitine acylcarnitine translocase *(CACT)*. Once inside mitochondria, carnitine palmitoyl transferase 2 *(CPT-2)*, located in the inner mitochondrial membrane, removes carnitine from acylcarnitines and re-generates acyl-CoAs. Carnitine then returns to the cytoplasm for another cycle (using CACT), while the acyl-CoAs can enter (in aerobic conditions and in the presence of low levels of ATP) β-oxidation with final production of acetyl-CoA for oxidative phosphorylation or production of ketone bodies in the liver.

mucopolysaccharidoses. In addition to symptomatic therapy to alleviate the most severe clinical symptoms, specific therapies for lysosomal storage disorders are available or in development: (1) enzyme replacement therapy; (2) substrate reduction therapy; (3) chaperone therapy; (4) bone marrow and hematopoietic stem cell transplantation; (5) gene therapy. Currently newborn screening for two lysosomal storage disorders, Pompe disease and mucopolysaccharidosis type I (Hurler, Hurler-Scheie, and Scheie syndrome), have been approved for inclusion in the RUSP.

Pompe Disease

Pompe disease is an autosomal recessive lysosomal storage disorder caused by deficiency of the enzyme, acid alpha-glucosidase (GAA), with an estimated incidence of approximately 1:40,000 births. In this condition, glycogen accumulates within the lysosomes in the skeletal muscle and heart. There are two main forms of Pompe disease: infantile and late-onset. In infantile Pompe disease, hypertrophic cardiomyopathy, muscle weakness, and hypotonia are usually evident by 2 months of age and most patients die within the first year of life. Late-onset Pompe disease is characterized by variable age of onset and presentation, with symptoms including progressive muscle weakness and respiratory insufficiency. The neonatal screening for Pompe disease is based on the measurement of enzyme activity in blood spots by (1) LC-MS/MS, which measures the enzymatic conversion of a substrate to a product, or (2) fluorimetric methods, which detect the release of a fluorescent compound from the substrate mediated by the enzyme. The diagnosis of Pompe disease is confirmed by measurement of enzyme activity in leukocytes or in fibroblasts and by DNA testing. DNA testing is important to identify patients with pseudodeficiency of enzymatic activity: a condition in which the enzyme functions normally against the natural substrate, but not against the synthetic substrate used to measure activity in the laboratory.

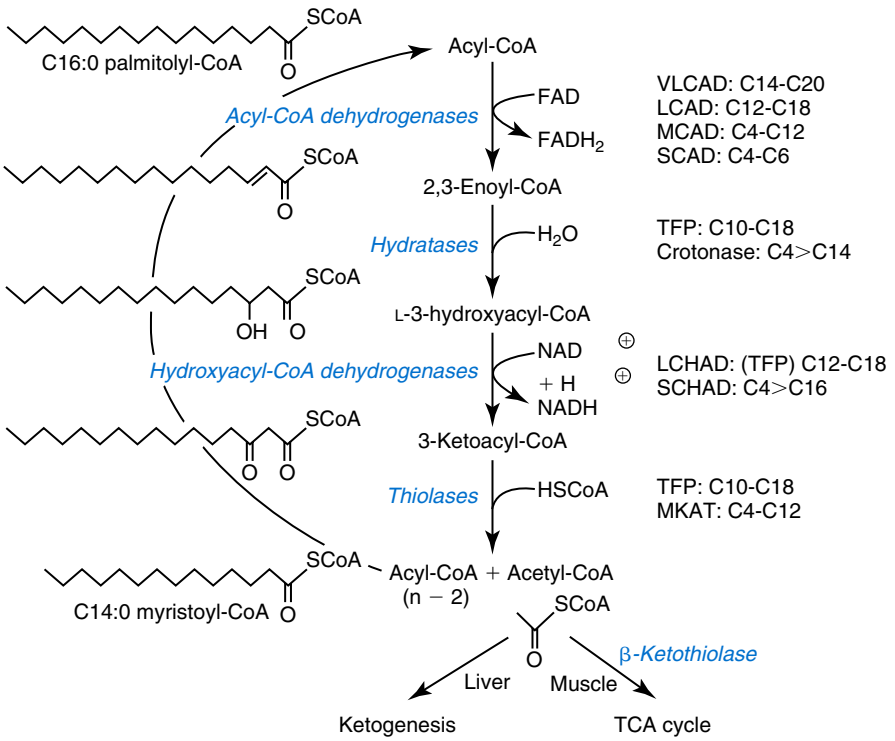


Fig. 45.7 Beta oxidation of fatty acids. In the mitochondrial matrix, long-chain fatty acids undergo a series of steps to progressively shorten the fatty acids of two carbon units (acetyl-**coenzyme A** [CoA]) through a series of enzymatic reactions. Chemical structures are shown on the left, enzymes *(blue)* and reactions are in the middle, and the carbon chain lengths for various disorders are on the right. Dehydrogenases introduce a double bond between C2 and C3 and are specific for fatty acids with different chain lengths: *LCAD,* Long-chain acyl CoA dehydrogenase; *MCAD,* medium-chain acyl-CoA dehydrogenase; *MKAT,* medium-chain acyl-CoA thiolase; *SCAD,* short-chain acyl-CoA dehydrogenase; *SCHAD,* short-chain 3-hydroxyacyl-CoA dehydrogenase; *VLCAD,* very long-chain acyl-CoA dehydrogenase. A trifunctional protein (TFP) adds water and cleaves two carbon atoms from the long-chain fatty acid. This is done through sequential action of a hydratase (enoyl-CoA hydratase), a L-3-hydroxyacyl-CoA dehydrogenase (*LCHAD,* long-chain 3-hydroxyacyl-CoA dehydrogenase), and a thiolase (acyl-CoA acetyltransferase). The two carbon units generated can be completely oxidized in the muscle to CO_2 or can generate ketone bodies in the liver that may be exported to other organs to provide energy.

Enzyme replacement therapy with human recombinant acid alpha-glucosidase improves the manifestations of the disease, including survival in affected children.

Mucopolysaccharidosis Type I

Mucopolysaccharidosis type I (MPS I) is an autosomal recessive lysosomal storage disorder caused by deficiency of the lysosomal hydrolase alpha-iduronidase. The incidence is estimated at 0.7 to 4:100,000. Alpha-iduronidase is involved in the degradation of the glycosaminoglycans heparan sulfate and dermatan sulfate, and its impaired activity leads to their accumulation in organs and tissues. MPS I is a progressive, debilitating, and often life-threatening disease. Typical features of MPS I are coarse facial features, hepatosplenomegaly, cardiac involvement, joint stiffness, skeletal deformities, and corneal clouding. There is a wide spectrum of phenotypes, with the most severe, MPS-IH or Hurler syndrome, presenting in the first 6 months of life with a rapid disease progression involving the central nervous system (CNS) and death by 10 years of age. In patients with the attenuated presentation, Hurler-Scheie (MPS-IHS) and Scheie (MPS-IS), the disease progression is slower, does not involve the CNS, and life expectancy is close to normal in the milder affected individuals. Newborn screening methods for MPS I are based on measurement of alpha-iduronidase activity. As for Pompe disease, there are two methods available for high throughput screening: (1) LC-MS/MS based and (2) fluorimetric. The diagnosis is confirmed by measuring enzyme activity in freshly isolated leukocytes and by sequencing of the *IDUA* gene. As for Pompe disease, DNA can identify individuals with enzyme pseudodeficiency. Enzyme replacement therapy with human recombinant alpha-iduronidase improves organ and skeletal involvement in MPS I patients, but is not effective in the neuropathic form (Hurler syndrome). In this case, hematopoietic stem cell transplantation early in life remains the only option to prevent CNS involvement, often in conjunction with enzyme replacement therapy to prevent peripheral complications.

Peroxisomal Disorders

Peroxisomes are single membrane organelles containing enzymes responsible for the alpha- and beta-oxidation of very long-chain fatty acids, the synthesis of the membrane glycerophospholipids plasmalogens, and the oxidation of bile acids

and cholesterol. Defects in the biogenesis of peroxisomes or defects in one or more peroxisomal function lead to severe clinical manifestations.

X-Linked Adrenoleukodystrophy

X-linked adrenoleukodystrophy (X-ALD) is an X-linked **peroxisomal disorder** resulting in defective oxidation of very long-chain fatty acids. Its incidence is approximately 1:20,000. X-ALD is caused by mutations in the *ABCD1* gene located on Xq28 that encodes for the adrenoleukodystrophy protein (ALDP) in the peroxisomal membrane, which transfers very long-chain fatty acids inside the peroxisomes for degradation. X-ALD can present in males before 5 years of age with progressive cognitive and neurological deterioration leading to a vegetative state and death. A milder variant, adrenomyeloneuropathy (AMN) presents usually in the 2nd to 3rd decade of life with progressive spastic paraplegia. Adrenal dysfunction occurs in about 2/3 of males with AMN. A total of 50 % to 65% of female carriers develop symptoms similar to AMN; however, the onset is usually in the 4th or 5th decade of life. X-ALD can be diagnosed by newborn screening measuring C26:0 lysophosphatidylcholine (C26:0 LPC) by LC-MS/MS. C26:0 LPC is elevated in DBS of patients with X-ALD, with peroxisomal biogenesis defects, and in approximately 80% of female carriers for X-ALD. The diagnosis is confirmed by measurement of very long-chain fatty acids (VLCFA) in plasma and by DNA sequencing of the *ABCD1* gene. VLCFAs in plasma are elevated and diagnostic in all males. However, they are elevated and informative only in 80% to 85% of heterozygous females for whom DNA testing should be considered. The only effective treatment for X-ALD is hematopoietic stem cell transplantation, to be performed before the appearance of neurological symptoms.

Disorders of Carbohydrate Metabolism

Enzyme deficiency in the metabolic pathways for carbohydrates results in accumulation of sugars within organs or tissues impairing their function, inability to obtain energy from them, or toxicity from the excess of monosaccharides or their derivatives (phosphorylated sugars). **Disorders of carbohydrate metabolism** include the (1) glycogen storage diseases, (2) **glucose-6-phosphate dehydrogenase (G-6-PD) deficiency**, and (3) classic **galactosemia**.

Glycogen storage disorders affect primarily the liver or the skeletal muscle. The accumulation of glycogen impairs organ function and, if the liver is affected, prevents release of glucose with resulting hypoglycemia. They are usually treated with avoidance of fasting and a special diet rich in protein, devoid of simple sugars and supplements with uncooked cornstarch.

G-6-PD deficiency affects red blood cells, causing them to break down (hemolysis). It can cause mild to severe jaundice in newborns and in some cases hemolytic anemia. Symptoms can be triggered by (1) infections, (2) certain drugs (some antibiotics and medications used to treat malaria), or (3) exposure to fava beans (a reaction called favism). The gene for this condition is on the X chromosomes, and the condition affects mostly males. Avoidance of stressors can mitigate or avoid symptoms.

Classic galactosemia results from absence of galactose-1-phosphate uridyl transferase. Galactose is derived from the disaccharide lactose found in milk. In galactosemia, galactose is phosphorylated to galactose-1-phosphate, but it is not further metabolized. Increased concentrations of galactose-1-phosphate in cells are toxic. Infants have (1) failure to thrive, (2) jaundice, (3) liver failure, and (4) predisposition to life-threatening infections with *E. coli* and other Gram-negative bacteria. The treatment of galactosemia involves removal of lactose (contained in human or animal-derived milk, but not in soy milk) and avoidance of all foods containing galactose. In galactosemia, intervention early in life provides the best prognosis, although some long-term complications (speech apraxia, learning disorders, and ovarian insufficiency in females) can still occur in treated individuals.

DIAGNOSTIC TESTS FOR INHERITED DISORDERS OF METABOLISM

A newborn screening result highly suggestive of a metabolic disorder typically leads to an immediate evaluation using confirmatory tests and a referral of the newborn to a metabolic center.

In asymptomatic patients, the confirmation of diagnosis relies on specific tests, such as (1) ion-exchange chromatography and (2) liquid chromatography-tandem mass spectrometry (LC-MS/MS) for amino acids analysis, (3) gas chromatography-mass spectrometry (GC-MS) for organic acids analysis, and (4) MS/MS without or with liquid chromatographic separation for acylcarnitines and acylglycines analyses, respectively. A combination of tests is often key in the confirmation of abnormal newborn screening results, especially for those with borderline values. DNA testing and enzyme assays are available for further confirmation of most of these conditions.

These techniques are described in Chapters 12, 13, 18, and 48. MS/MS is a technique of choice in newborn screening programs.

Tandem Mass Spectrometry

The introduction of MS/MS as a routine laboratory technique has dramatically increased the number of disorders amenable to newborn screening, since multiple metabolites are detected simultaneously in the same blood spot (**multiplex analysis**). This allows the identification of several disorders at once, whereas traditional screening techniques were based on one test for one disorder.

Methodological Details

MS/MS measures the ratio of the mass *(m)* of a chemical to its charge *z* (see Chapter 13). For newborn screening, a tandem mass spectrometer is configured to measure only acylcarnitines and amino acids using the information about their mass and fragmentation pattern. Labeled internal standards are added to the extraction mixtures to quantify the different species.

Interpretation

With MS/MS screening, several analytes are detected at the same time and the interpretation of the results is based heavily on pattern recognition, whereas the measurement of the concentration of the different metabolites supports the interpretation. The ability to detect multiple metabolites by MS/MS allows the use of ratios of metabolites to define whether an increased value is due to a metabolic derangement or to the clinical and nutritional status of the newborn.

Assessment of congenital disorders at the appropriate time after birth is crucial. For example, although amino acid concentrations do not change significantly with newborn age, acylcarnitine concentrations vary significantly. For most acylcarnitines, the concentrations are highest in the first week of life and decrease rapidly afterward. Typically the first screen sample should be collected within 24 to 48 hours of life, while the second screen sample, in states mandating two screens, is collected between 7 and 28 days of age. Age-appropriate cutoff values for all screening tests should be used for the interpretation of newborn screening results.

REVIEW QUESTIONS

1. An inherited disorder that affects the conversion of nutrients into energy is referred to as a(n):
   a. aminoacidopathy.
   b. autosomal recessive disorder.
   c. inborn error of metabolism (IEM).
   d. hemolytic disease of a newborn.
2. Tandem mass spectrometry (MS/MS) involves:
   a. an ultraviolet laser to ionize small amounts of matrix and analyte that are directed into the mass analyzer.
   b. screening a large number of infants for disorders of amino acid metabolism using different growth antagonists.
   c. an ion trap designed to trap ions in three dimensions instead of two dimensions.
   d. assessment of the mass-to-charge ratio of a chemical.
3. An inborn disorder of metabolism that is triggered by prolonged fasting and/or acute illness and that is confirmed by abnormal acylcarnitine concentrations is:
   a. galactosemia.
   b. medium-chain acyl-CoA dehydrogenase (MCAD) deficiency.
   c. phenylketonuria (PKU).
   d. homocystinuria.
4. Deficiency of an enzyme that results in excess of monosaccharides in the blood of a newborn leads to a specific disorder of metabolism. An example of this type of disorder would be:
   a. galactosemia.
   b. medium-chain acyl-CoA dehydrogenase (MCAD) deficiency.
   c. phenylketonuria (PKU).
   d. glutaric acidemia type I (GAI).
5. The disorder that is identified by increased blood glutarylcarnitine on newborn screening is characterized by dysfunctional metabolism of:
   a. hydroxylysine and phenylalanine.
   b. tyrosine and tryptophan.
   c. tryptophan, hydroxylysine, and lysine.
   d. cystine and hydroxylysine.
6. In a disorder that has an autosomal recessive inheritance pattern, what is the risk of two carrier parents having an affected child with that disorder?
   a. 0%
   b. 25%
   c. 50%
   d. 100%
7. In relation to newborn screening for inborn errors of metabolism (IEMs), second-tier testing is done to:
   a. assess the parents of the newborn for possible carrier status.
   b. assess the siblings of the newborn for similar symptoms or disorders.
   c. determine the number of false negatives that might have occurred with initial screening.
   d. further assess a positive screening test result by targeting more specific analytes.
8. The disorder of amino acid metabolism in which the parent amino acid accumulates in blood and spills over into urine is classified as:
   a. an organic acidemia.
   b. an aminoacidopathy.
   c. a carbohydrate disorder.
   d. a disorder of fatty acid oxidation.
9. The simultaneous analysis of multiple analytes using a single sample, such as a blood spot from a newborn, is referred to as:
   a. multiplex analysis.
   b. newborn screening.
   c. tandem mass spectrometry (MS/MS).
   d. multiple of means.
10. The enzyme that is absent in the aminoacidopathy phenylketonuria (PKU) is:
    a. phenylalanine decarboxylase.
    b. ornithine translocase.
    c. phenylalanine hydroxylase.
    d. succinyl-coenzyme A (CoA) mutase.

SUGGESTED READINGS

Guthrie, R., & Susi, A. (1963). A simple phenylalanine method for detecting phenylketonuria in large populations of newborn infants. *Pediatrics*, *32*, 338–343.

Schwarz, E., Liu, A., Randall, H., Haslip, C., Keune, F., Murray, M., et al. (2009). Use of steroid profiling by UPLC-MS/MS as a second tier test in newborn screening for congenital adrenal hyperplasia: The Utah experience. *Pediatr Res*, *66*, 230–235.

Shigematsu, Y., Hata, I., & Tajima, G. (2010). Useful second-tier tests in expanded newborn screening of isovaleric acidemia and methylmalonic aciduria. *J Inherit Metab Dis*, *33*, S283–S288.

Wong, D., Tortorelli, S., Bishop, L., Sellars, E. A., Schimmenti, L. A., Gallant, N., et al. (2016). Outcomes of four patients with homocysteine remethylation disorders detected by newborn screening. *Genet Med*, *18*, 162–167.

Therrell, B. L., Jr., Lloyd-Puryear, M. A., Camp, K. M., & Mann, M. Y. (2014). Inborn errors of metabolism identified via newborn screening: ten-year incidence data and costs of nutritional interventions for research agenda planning. *Mol Genet Metab*, *113*, 14–26.

Anikster, Y., Haack, T. B., Vilboux, T., Pode-Shakked, B., Thony, B., Shen, N., et al. (2017). Biallelic mutations in DNAJC12 cause hyperphenylalaninemia, dystonia, and intellectual disability. *Am J Hum Genet*, *100*, 257–266.

Longo, N., Arnold, G. L., Pridjian, G., Enns, G. M., Ficicioglu, C., Parker, S., et al. (2015). Long-term safety and efficacy of sapropterin: The PKUDOS registry experience. *Mol Genet Metab*, *114*, 557–563.

Hedlund, G. L., Longo, N., & Pasquali, M. (2006). Glutaric acidemia type 1. *Am J Med Genet C Semin Med Genet*, *2006*(142C), 86–94.

Viau, K., Ernst, S. L., Vanzo, R. J., Botto, L. D., Pasquali, M., & Longo, N. (2012). Glutaric acidemia type 1: outcomes before and after expanded newborn screening. *Mol Genet Metab*, *106*, 430–438.

Longo, N., Frigeni, M., & Pasquali, M. (2016). Carnitine transport and fatty acid oxidation. *Biochim Biophys Acta*, *1863*, 2422–2435.

Wilcken, B. (2010). Fatty acid oxidation disorders: Outcome and long-term prognosis. *J Inherit Metab Dis*, *33*, 501–506.

Matern, D., & Rinaldo, P. (1993-2017). Medium-chain acyl-coenzyme a dehydrogenase deficiency. In M. P. Adam, H. H. Ardinger, R. A. Pagon, & S. E. Wallace (Eds.), *GeneReviews [Internet]*. Seattle: Univ Washington.

Grabowski, G. A. (2010). Gaucher disease and other storage disorders. *Hematology Am Soc Hematol Educ Program*, *2012*, 13–18.

Chien, Y. H., Hwu, W. L., & Lee, N. C. (2013). Pompe disease: Early diagnosis and early treatment make a difference. *Pediatr Neonatol*, *54*, 219–227.

Burton, B. K. (2012). Newborn screening for Pompe disease: An update, 2011. *Am J Med Genet C, Semin Med Genet*, *160C*, 8–12.

Matern, D., Gavrilov, D., Oglesbee, D., Raymond, K., Rinaldo, P., & Tortorelli, S. (2015). Newborn screening for lysosomal storage disorders. *Semin Perinatol*, *39*, 206–216.

de Ru, M. H., Boelens, J. J., Das, A. M., Jones, S. A., van der Lee, J. H., Mahlaoui, N., et al. (2011). Enzyme replacement therapy and/or hematopoietic stem cell transplantation at diagnosis in patients with mucopolysaccharidosis type I: Results of a European consensus procedure. *Orphanet J Rare Dis*, *6*, 55.

Vogel, B. H., Bradley, S. E., Adams, D. J., D'Aco, K., Erbe, R. W., Fong, C., et al. (2015). Newborn screening for X-linked adrenoleukodystrophy in New York state: diagnostic protocol, surveillance protocol and treatment guidelines. *Molecular genetics and metabolism*, *114*, 599–603.

Berry, G. T. (1993-2017). Classic galactosemia and clinical variant galactosemia. In M. P. Adam, H. H. Ardinger, R. A. Pagon, & S. E. Wallace (Eds.), *GeneReviews [Internet]*. Seattle: Univ Washington.

Chace, D. H. (2009). Mass spectrometry in newborn and metabolic screening: Historical perspective and future directions. *J Mass Spectrom*, *44*, 163–170.

46

Pharmacogenetics

Gwendolyn A. McMillin, Mia Wadelius, Victoria M. Pratt

OBJECTIVES

1. Define the following:
   - Adverse drug reaction (ADR)
   - Prodrug
   - Haplotype
   - Pharmacogenetics
   - Pharmacokinetics
   - Pharmacodynamics
2. Describe the difference between type A and type B ADRs and provide an example of each.
3. Provide examples of pharmacogene-drug pairs that are associated with each of these three drug response processes: pharmacokinetics, pharmacodynamics, and immune response.
4. Explain how the CYP2D6 phenotype could predict a poor response to codeine and other prodrugs that are substrates of CYP2D6.
5. Explain how multiple pharmacogenes affect the therapeutic dose of warfarin.
6. Explain how genetic testing is used to evaluate whether or not a patient is a good candidate for ivacaftor.

KEY WORDS AND DEFINITIONS

Adverse drug reaction (ADR) Any undesirable side effect or toxic reaction that is caused by the administration of a drug.

Allele One of the alternative versions of a gene at a given location (locus) along a chromosome. An individual typically inherits two alleles for each gene, one from each parent. If the two alleles are the same, the individual is homozygous for that locus. If the alleles are different, the individual is heterozygous. The term now also refers to variation among noncoding DNA sequences in addition to genes.

Cytochrome P450 (CYP) A large family of metabolic enzymes that catalyze the oxidation of organic substrates. They are important in the metabolism of drugs, steroid hormones, fatty acids, and other substances.

Drug metabolism The process by which drugs are chemically modified in the body, typically by drug-metabolizing enzymes. Most but not all drugs are metabolized.

Gene duplication A process by which a chromosome or a portion of DNA is duplicated, resulting in an additional copy of a gene.

Genotype The genetic makeup of an individual or a group of individuals. An individual's genotype codes for that individual's phenotype. The genotype for a specific gene requires knowledge of both *alleles* of the gene.

Genotyping Laboratory testing designed to detect specific genetic variants that are used to identify and classify variant alleles. The alleles are reported together as a genotype that may or may not be predictive of a phenotype.

Haplotype A combination of alleles on the same chromosome that are usually located close together and tend to be inherited together.

International normalized ratio (INR) A method of reporting prothrombin time results for patients receiving oral anticoagulant therapy.

Metabolite A chemically modified form of a compound. Drug metabolites may or may not exhibit pharmacological activity. There are often many metabolites of a single drug, some of which may be identical to independently administered drugs.

Metabolizer classification A classification of four drug-metabolism phenotypes as they relate to an individual's metabolic efficiency for processing a specific therapeutic drug. The four types are (1) poor metabolizers (PMs), (2) ultrarapid metabolizers (UMs), (3) extensive/normal metabolizers (EMs), and (4) intermediate metabolizers (IMs).

Pharmacodynamics A process in pharmacology that defines how a drug acts on its target and its mechanisms of action.

Pharmacogene A gene that is known to influence the pharmacokinetics and/or pharmacodynamics of a drug.

Pharmacogenetics Study of the variations in single genes or small groups of related genes that affect the response to a drug, including its metabolism.

Pharmacogenomics The study of how combinations of variations in several genes, potentially extended to the complete genome, influence the pharmacology of a drug or drugs.

Pharmacokinetics A process in pharmacology that describes how a drug is (1) absorbed, (2) distributed, (3) metabolized, and (4) eliminated.

Phenotype The observable physical or biochemical characteristics of an individual as determined by his or her genetic makeup and environmental influences, including drug-drug or food-drug interactions.

Prodrug A drug that is administered in an inactive or poorly active form and converted to an active drug by metabolism.

Pseudogene A defective segment of DNA that resembles a gene but cannot be transcribed.

Star (*)-allele nomenclature A nomenclature developed to standardize genetic polymorphism annotation for many pharmacogenes, such as cytochrome P450.

PRINCIPLES OF PHARMACOGENETICS

The term **pharmacogenetics** comes from merging the terms pharmacology and genetics. Pharmacogenetics can predict and/or explain how individuals respond to drugs and is a prominent component of personalized precision medicine. Initiatives exist around the globe to improve patient care with pharmacogenetics. For example, former US President Barack Obama announced his "Precision Medicine Initiative" at his January 2015 State of the Union Address. This initiative was intended to promote "patient-powered research" and "provide clinicians with new tools, knowledge, and therapies to select which treatments will work best for which patients," including pharmacogenetics. Related work associating drug response with **pharmacogenes**, and ultimately the whole genome, is known as **pharmacogenomics**, although the terms are commonly used interchangeably. Pharmacogenetics and pharmacogenomics can apply to the human germline genome, tumor genomes (e.g., somatic mutations), and pathogen genomes (e.g., viral genomes). The goal of this chapter is to explain concepts and provide examples of human pharmacogenetics, with an emphasis on germline variants. Targeted variants of some pharmacogenes and specific applications to drugs are described.

Drug Response

For simplicity, the term drug(s) is used throughout this chapter to reflect any xenobiotic (foreign compound absorbed by the human body) that is capable of evoking a physiologic or behavioral response. Responses to drugs depend on many variables such as drug formulation, route of administration, age, gender, clinical status (e.g., kidney function, liver function, protein status), comedications, and genetics. Most drugs are selected and initially dosed according to drug labeling, clinical experience, and institutional protocols that stem from population-based dosage and dose frequency recommendations. However, dose optimization still largely relies on trial and error. Minimizing this process of trial and error with pharmacogenetics can improve efficacy of drugs and prevent **adverse drug reactions**.

Pharmacogenetic testing is designed to predict specific aspects of the two major processes upon which drug response is based: pharmacokinetics and pharmacodynamics. **Pharmacokinetics** describes how the body acts on a drug, often called ADME, referring to absorption, distribution, metabolism, and elimination. Pharmacogenes that encode drug-metabolizing enzymes and transport proteins are involved in pharmacokinetics. Parent drugs that are administered as inactive compounds and require metabolism to convert the compound to an active drug are called **prodrugs**. Drug **metabolites** can be pharmacologically active or inactive. Therefore it is important to consider whether pharmacogenetic variation influences the concentrations or kinetics of pharmacologically active or inactive compounds. Fig. 46.1 illustrates the common metabolic relationship for prodrugs, active drugs, and metabolites (active and/or inactive). **Pharmacodynamics** describes how the body responds to drugs, both desirable (e.g., therapeutic) and undesirable (e.g., therapeutic failure and/or toxicity). Pharmacogenes that encode for mechanistic proteins such as enzymes, receptors, and ion channels are involved in pharmacodynamics. Sometimes drugs cause adverse effects due to mechanisms that are unrelated to the intended use of the drug. For example, carbamazepine, a drug used to treat seizures and neuropathic pain largely by inhibiting voltage-gated sodium channels, can stimulate the immune system, leading to a severe cutaneous adverse reaction that can be life-threatening. The risk of this adverse reaction is increased in the presence of the *HLA-B*15:02* **allele**, representing variation in the genes that code for the human leukocyte antigen (HLA) system, and is unrelated to the mechanisms responsible for managing seizures and pain. The adverse reaction is also unrelated to the pharmacokinetics of carbamazepine. Pretherapeutic testing to detect this allele in people being considered for carbamazepine therapy is recommended for vulnerable populations, particularly people of Asian descent.

Many different pharmacogenes are often associated with the pharmacokinetics and pharmacodynamics of a single drug. In selecting drugs and drug dosing for an individual, one should consider the clinically significant pharmacogenes in combination with relevant clinical, demographic and nongenetic factors. After drug and dose are selected, response to the drug should be monitored. If the response is desirable (therapeutic), the pharmacokinetics and pharmacodynamics are appropriate. Therapeutic failure may occur if the concentrations of active drug are insufficient (e.g., pharmacokinetic variability) or if the physiology required to elicit the response to the drug is absent or impaired (e.g., pharmacodynamic variability). If the response is not optimal or is undesirable, therapy may have to

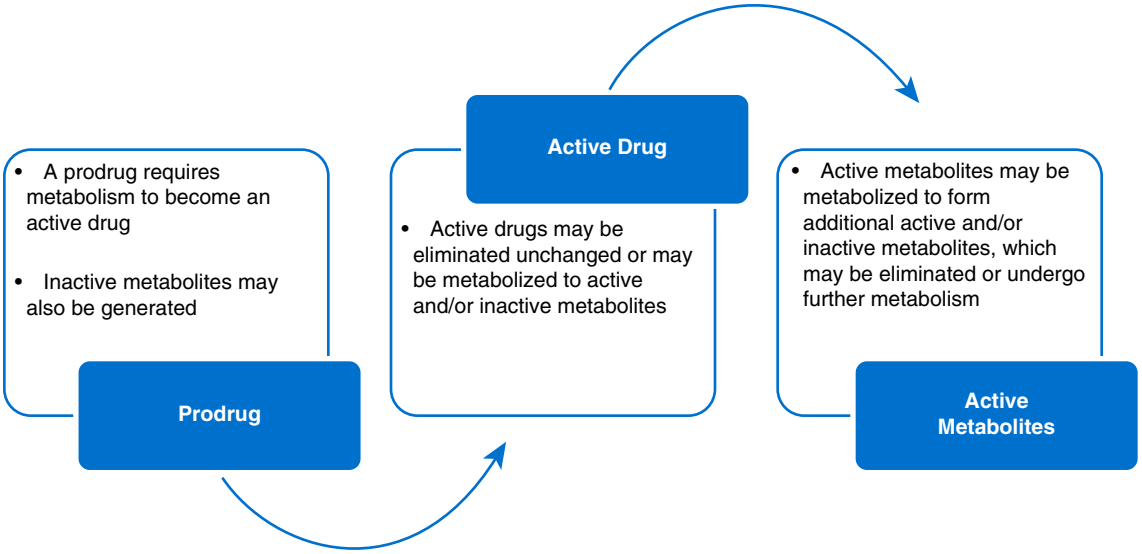


Fig. 46.1 Schematic relationships between prodrugs, active drugs, and metabolites.

be adjusted. The dose of a drug is adjusted based on results of clinical measurements (e.g., blood pressure for antihypertensive drugs), therapeutic drug monitoring (TDM), or monitoring of biochemical markers indicating response.

The goal of TDM is to adjust the dose to achieve blood concentrations of active drug that fall within an established therapeutic range at particular times after dose administration. This practice of dose adjustment is common to immunosuppressant drugs such as tacrolimus or anticonvulsant drugs such as carbamazepine. Thus blood specimens are collected at specific times after the drug is administered. The dose is adjusted to achieve concentrations that consistently fall within the therapeutic range selected for the patient population and clinical indication, noting that pharmacokinetic variation can contribute to differences in time to achieve stable drug concentrations, called achieving steady state. Pretherapeutic pharmacogenetic testing can guide drug selection, determine the initial drug dose, and predict the time required to achieve steady state.

An example of a biochemical marker of response is the **international normalized ratio (INR)**, which is calculated from the prothrombin time (PT) test (Box 46.1); it is used to guide and adjust doses of the common anticoagulant drug warfarin. As with TDM, blood samples are collected at specific times after the administration of warfarin. The dose is adjusted until the INR consistently falls within a target range, selected for the patient population and clinical indication. The target INR range is set to 2.0 to 3.0 for most indications. Pretherapeutic pharmacogenetic testing can guide the initial dose of warfarin and predict the time required to achieve steady state.

Adverse drug reactions that are dose-dependent are classified as type A. Such scenarios can be managed by adjusting the dose of a drug based on clinical or laboratory monitoring of drug concentrations and/or biomarkers. Inappropriate dosing of tacrolimus, carbamazepine, or warfarin can lead to type A reactions. Adverse drug reactions that occur independently of dose are classified as type B. As such, monitoring and the adjusting dose will be unsuccessful. In this case, an alternate drug is required. The carbamazepine-induced hypersensitivity example illustrates a type B reaction. Carbamazepine is an example of a drug that can produce both type A and type B adverse reactions. Pretherapeutic pharmacogenetic testing may predict vulnerability to type A and type B adverse reactions should these drugs be administered.

BOX 46.1 International Normalized Ratio (INR)

INR = [(PT result for the patient)/(PT result for a normal control)]^{ISI}

- Established by the World Health Organization and the International Committee on Thrombosis and Hemostasis.
- Standardizes reporting of the prothrombin time (PT) test, a common method of evaluating how many seconds it takes for a person's blood to clot.
- Includes an international sensitivity index (ISI) that compensates for variability in laboratory methods (usually 1.0–2.0).
- A typical therapeutic range of the INR for a patient treated with warfarin is 2.0–3.0.

POINTS TO REMEMBER

Phenotypes described for drug metabolizing enzymes are as follows:

- Ultrarapid or rapid metabolizer: more than normal enzyme activity and/or gene expression is expected/observed.
- Extensive/normal metabolizer: normal enzyme activity and/or gene expression is expected/observed.
- Intermediate metabolizer: less than normal enzyme activity and/or gene expression is expected/observed.
- Poor metabolizer: little or no enzyme activity is expected/observed.

Implementation and Logistics of Pharmacogenetics

Pharmacogenetic testing is intended to predict and/or explain discrete aspects of pharmacokinetics and pharmacodynamics to guide drug and dose selection. Once a drug is initiated, the response is monitored and dosing optimized with clinical and/or laboratory tools. It is not practical, medically indicated, or cost-effective to apply pharmacogenetics to every drug therapy situation. The pharmacogenetic tests that have proven to be most successful are those that produce actionable results. Many resources for labeling information, gene-drug associations, and clinical consensus guidelines are maintained and updated electronically. Some specific gene-drug relationships for which US Food and Drug Administration (FDA) labeling includes pharmacogenetic information are shown in Table 46.1. These gene-drug examples apply to medical disciplines including oncology, psychiatry, neurology, infectious disease, pain management, and cardiology. Other areas of medicine can also benefit from pharmacogenetic testing, and the number of clinically useful targets will only increase as comprehensive pharmacogenetic testing becomes more widely available with analytic techniques such as massively parallel sequencing.

Specimens

Most pharmacogenetic testing revolves around DNA extracted from blood, saliva, buccal cells, or other specimens from which DNA can be obtained. Specimens for DNA testing do not require special timing or patient preparation in most situations. Saliva or buccal cells may be preferred due to the noninvasive nature of collection, but this may not produce a sufficient quantity or quality of DNA for all applications.

Analytic Strategies

The analytic strategy used in pharmacogenetic testing depends on a variety of factors, such as the complexity of the gene; the extent, frequency, and type of genetic variation; and the time needed for return of the results. Most **genotyping** assays cannot detect all variants that have been identified in a gene. If the need for a rapid time to result is clinically indicated, an assay may be designed to limit complexity, as through targeted detection of the most common and clinically relevant variants. Commercially available in vitro diagnostics (IVDs) are available for some pharmacogenetic applications and may reduce the complexity of testing and data analysis in order to reduce the time to result. However, most pharmacogenetic testing is based on laboratory-developed approaches designed to target specific gene variants, and performed at a central or reference laboratory. Clinical laboratories that offer pharmacogenetic testing can be found through the voluntary National Institutes of Health Genetic Testing Registry.

Laboratory services may include a single-gene or multiple-gene panels where known variants are interrogated. Targeted genotyping will not detect any variant or alleles that are not directly interrogated, so a negative genotyping result does not rule out the possibility that a patient is carrying another variant not detected by the assay. Multiple gene panels may be achieved through exome or whole-genome sequencing. Exome sequencing is able to identify only those variants near to and including the coding regions of genes. This approach cannot identify intergenic differences, including structural and noncoding variants, which can be found by whole-genome sequencing. Massively parallel sequencing platforms available today have additional limitations

TABLE 46.1 Drug-Gene Examples With Published Guidelines, FDA Labeling Comments, and Graded Level of Evidence

		GRADES, GUIDELINES[a]			COMMENT IN FDA LABELING[a]			
Gene or Allele	Generic Drug Name	PharmGKB	CPIC	DPWG	Testing Required	Testing Advised	Actionable	Not Available
HLA-B*57:01	Abacavir	1A	Yes	Yes		X		
CYP2D6 CYP2C19	Amitriptyline	1A	Yes	Yes			X	
HLA-B*15:02	Carbamazepine	1A	Yes	No	X[b]			
CYP2C19	Clopidogrel	1A	Yes	Yes		X		
CYP2D6	Codeine	1A	Yes	Yes			X	
CFTR	Ivacaftor	1A	Yes	No	X			
TPMT	Mercaptopurine	1A	Yes	Yes		X		
G6PD	Rasburicase	1A	Yes	No	X			
SLCO1B1	Simvastatin	1A	Yes	No				X
CYP3A4	Tacrolimus	1A	Yes	Yes				X
CYP2C9 VKORC1	Warfarin	1A	Yes	No			X	

[a]Levels of evidence (grades), status of guidelines, and comments in FDA labeling are provided as examples but are time-sensitive and therefore may change. The PharmGKB grade of 1A indicates that the highest level of evidence to support the gene-drug association, for which CPIC guidelines are available and/or widespread clinical implementation is known. The PharmGKB grade of 2A indicates moderate evidence and lack of a CPIC guideline or widespread clinical implementation.

[b]In patients with ancestry in genetically at-risk populations, mainly Asians.

CPIC, Clinical Pharmacogenetics Implementation Consortium; *DPWG*, Dutch Pharmacogenomics Working Group; *PharmGKB*, pharmacogenomics knowledge base.

including the following: (1) decreased coverage in regions with high GC content (e.g., the 5' ends of genes), limited detection of copy number variants, limited detection of insertions/deletions, and interference from **pseudogenes**. These technical limitations are expected to improve if not resolve over time. However, rare and novel variants will likely be of uncertain clinical significance, causing difficulty in interpretation. Novel combinations of variants may also be identified by sequencing, which will necessitate the evolution of nomenclature and may affect **phenotype** predictions. Currently whole-genome sequencing is not practical for routine pharmacogenetics applications owing to the high costs and time associated with sequencing and interpretation.

Gene Dose (Copy Number)

Human DNA is normally associated with two copies of each gene, one copy inherited from each parent for autosomal regions on each chromosome. However, many genetic regions display a variation in the number of copies and are termed copy number variants (CNVs). CNVs can range in size from one kilobase to several megabases due to deletion, insertion, **duplication**, or complex recombination. CNVs in some pharmacogenes play a clear role in producing drug efficacy and toxicity. One example is the human CYP2D locus, which contains two pseudogenes, *CYP2D7* and *CYP2D8*, closely located and evolutionarily related to *CYP2D6,* and thus facilitates homologous crossovers and the formation of large gene conversions, deletions, duplications, and multiplications of CYP2D6. CNVs are not always included in analytic methods designed for genotyping.

Haplotyping

Many of the pharmacogenes exhibit combinations of variants—for example, single-nucleotide variants (SNVs) and insertion-deletions (indels) within a gene that are inherited together on the same chromosome and are referred to as **haplotypes**. Many nomenclature systems have been developed to describe pharmacogenetic haplotypes. In the most commonly used nomenclature system, combinations of pharmacogenetic sequence variants are designated **by star (*) alleles**, where *1 is designated as normal (commonly referred to as wild-type or fully functional), and numbered star alleles are assigned as new variants are identified. Pragmatically the *1 allele is assigned by default when none of the targeted variant alleles are detected. Therefore the true accuracy of a *1 allele designation depends on whether the assay detects all possible functional variants.

Reporting

The results of clinical pharmacogenetic testing should be reported using current recommended standard nomenclature. Pharmacogenetic nomenclature is constantly evolving, and laboratories or users of laboratory services might not be familiar with most current nomenclature. Therefore test results should be provided in current recommended standard nomenclature, which should include clarifications of commonly used terms and should indicate the **genotypes** detected.

In general, which variants/alleles should be interrogated and reported for the pharmacogenes is not standardized. In addition, some assays use different combinations of variants to define or infer the haplotypes that the assay detects, which ultimately can lead to discrepancies in star allele genotypes between platforms. Laboratories are required by the Clinical Laboratory Improvement Amendment (CLIA) to include interpretation of the test results on the test report. Clinical pharmacogenetic laboratories often provide an interpreted phenotype (e.g., poor **metabolizer**) based on the genotype results. Laboratories that are accredited by the College of American Pathologists (CAP) are required to include a summary of methods and the variants that can be detected on the report, but overall the lack of consensus for pharmacogenetic nomenclature (e.g., *alleles, rs#, or HGVS, diplotype vs. haplotype, etc.) and variable assay designs add to the complexity of analyzing and reporting results from pharmacogenetic assays.

Clinical Interpretation

Pharmacogenetic testing is clinically useful only when information is sufficient to allow clinical interpretation of the results. This information must be derived from human studies. Many examples of this type of information can be found in the peer-reviewed literature. One limitation of the existing literature, however, is that most data are based on retrospective studies, and there is no single printed source in which all of this information has been collated. The Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB) is a publicly available Internet research tool developed at Stanford University through a nationwide collaborative research consortium. Its aim is to aid researchers in understanding how genetic variation among individuals contributes to differences in reactions to drugs. This regularly updated database is a source of genetic and clinical information. In addition, the Pharmacogene Variation (PharmVar) Consortium was introduced in September 2017 as a central electronic repository and resource.

Many ongoing clinical trials designed to study the efficacy and toxicity of new pharmaceutical products or new indications for previously developed pharmaceuticals have employed pharmacogenetic testing. Thus it is anticipated that pharmacogenetic guidelines and new diagnostic products will be available in the coming years. In addition, drugs that were previously removed from development because of adverse drug reactions or poor therapeutic response may be reconsidered if a genetic test can be demonstrated to identify individuals who are likely to respond favorably to the drug. For each of the major pharmacogenes discussed in this chapter, Table 46.1 provides a list of common drugs with pharmacogenetic associations that have resulted in several FDA-revised drug labels. Because many other drug-gene combinations are recognized that may lead to additional drug label revisions, the discussion provided here is in no way comprehensive. In addition, many of the specific genes discussed later have applications to other areas of medicine, some of which are mentioned briefly. The European Medicines Agency has

TABLE 46.2 Recommendations for Pharmacogenetics in the Summary of Product Characteristics Sections of the European Union Drug Labeling for Germline Variation, by Gene

Section	Section Title	Recommendation	Gene Examples
4.1	Therapeutic indications	State when a product indication depends on a particular genotype, phenotype, or expression of a gene	*HLA-B*57:01, CFTR*
4.2	Posology and method of administration	State dose adjustment recommendations linked to a particular genotype	*TPMT, CFTR*
4.3	Contraindications	State contraindications linked to a particular genotype	*DPYD, G6PD*
4.4	Special warnings and precautions for use	State when adverse drug reactions including therapeutic failure are linked to a specific genotype or phenotype	*HLA-B*57:01, TPMT, CFTR, CYP2C19, UGT1A1*
4.5	Interaction with other medicinal products	State when interactions with other medicinal products depend on specific genotype or phenotype	*CYP2D6, CYP2C19, UGT1A1*
4.8	Undesirable effects	State any clinically relevant differences in adverse drug reactions including therapeutic failure that are linked to a specific genotype	*HLA-B*57:01, CFTR, G6PD*
5.1	Pharmacodynamic properties	State relevant clinical studies that show a difference in benefit or risk depending on a specific genotype or phenotype	*HLA-B*57:01, CFTR, G6PD*
5.2	Pharmacokinetic properties	State variations in metabolism and associated quantitative terms if clinically relevant	*CYP2D6, TPMT, DPYD, CYP2C19, UGT1A1*

Data from Ehmann, F., Caneva, L., Prasad, K., et al. (2015). Pharmacogenomic information in drug labels: European Medicines Agency perspective. *Pharmacogenomics J, 15*, 201–210.

included pharmacogenetic information in drug labeling as well. Although not all European countries participate, the European Union has promoted standardization in the labeling of pharmaceuticals and included pharmacogenetic information in the summary of product characteristics. Table 46.2 provides examples of the sections in a European drug label that may include pharmacogenetic information and examples of genes that are targeted in drug labels. The specific examples discussed further on can be used as a guide for translating the principles of pharmacogenetics to additional gene-drug pairs in this continually evolving field of laboratory medicine.

SPECIFIC EXAMPLES OF PHARMACOGENE ASSOCIATIONS

Associations With Pharmacokinetics

Drug-Metabolizing Enzymes

Most drugs are metabolized by at least one reaction mediated by **cytochrome P450 (CYPs)** enzymes. CYPs are classified into families and further into subfamilies based on amino acid homology. Using *CYP2D6* as an example, the naming convention for CYPs is shown in Fig. 46.2, illustrating that the core name of the gene and protein is shared, identifying the family, subfamily and polypeptide, also referred to as the isozyme. The variant allele designations follow the name of the protein. When referring to a gene, the name should be italicized, whereas the name of the associated the protein should not be.

Genetic variants of CYPs have been associated with changes in enzyme activity, stability, and/or substrate (e.g., drug or drug metabolite) affinity that can lead to clinically significant phenotypes. Alleles for all CYP genes are

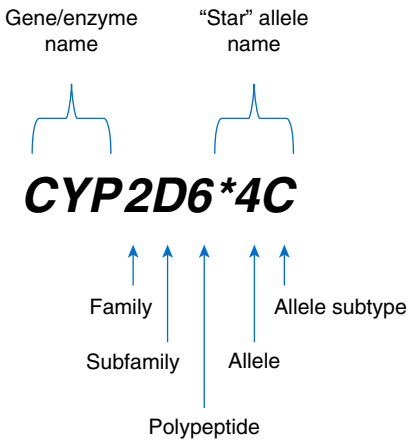


Fig. 46.2 Description of the nomenclature for the *CYP2D6*4C* allele.

described according to international consensus, based on the star (*) nomenclature described earlier, often followed by a letter to describe an allele subtype. Subtypes do not usually influence the clinical phenotype but could be important for haplotype assignment or to support research studies.

Cytochrome P450 CYP2D6 (CYP2D6). CYP2D6 is one of the most important drug-metabolizing enzymes as it is responsible for metabolizing approximately 25% of all drugs and is commonly involved in drug-drug interactions. Common alleles for the *CYP2D6* gene are shown in Table 46.3 with the predicted functional consequences. Although the definition of an allele may involve detection of many variants, a common SNV is frequently chosen as the primary analytic target used to detect each allele (bolded). In many cases, however, detection of other variants is required to accurately classify an allele. For

TABLE 46.3 Examples of Common Star (*) Allele Definitions for *CYP2D6*

Allele	Nucleotide Changes (cDNA)[a]	Effect	Enzyme Function
CYP2D6*1	None		Normal
CYP2D6xN	Gene amplifiers (multiple copies)	Depends on allele	Increased for functional alleles; no effect for nonfunctional alleles
CYP2D6*2A	−1584C>G; −1235A>G; −740C>T; −678G>A; CYP2D7 gene conversion in intron 1; 1661G>C; 2850C>T; 4180G>C	R296C; S486T	Normal
CYP2D6*3A	**2549delA**	Frame shift	Nonfunctional
CYP2D6*4A	100C>T; 974C>A; 984A>G; 997C>G; 1661G>C; **1846G>A**; 4180G>C	P34S; L91M; H94R; splicing defect; S486T	Nonfunctional
CYP2D6*5	Gene deletion		Nonfunctional
CYP2D6*10A	**100C>T**; 1661G>C; 4180G>C	P34S; S486T	Decreased function
CYP2D6*17	**1023C>T**; 1661G>C; **2850C>T**; 4180G>C	T1071; R296C; S486T	Decreased function
CYP2D6*36 (single)	−1426C>T; −1235A>G; −1000G>A; **100C>T**; 310G>T; 843T>G; 1039C>T; 1661G>C; 2097A>G; 3384A>C; 3582A>G; gene conversion to CYP2D7 in exon 9; 4180G>C	**P34S**; P469A; T470A; H478S; G479R; F481V; A482S; S486T	Nonfunctional
CYP2D6*41	−1235A>G; −740C>T; −678G>A; CYP2D7 gene conversion in intron 1; 1661G>C; 2850C>T; **2988G>A**; 4180G>C	R296C; splicing defect; S486T	Decreased function

[a]When known, *bold* nucleotide variants represent the major alterations responsible for the effect.
Data from Sim, S. C., & Ingelman-Sundberg, M. (2010). The Human Cytochrome P450 (CYP) Allele Nomenclature website: a peer-reviewed database of CYP variants and their associated effects. *Hum Genomics*, *4*, 278–281. <https://www.pharmvar.org/>.

TABLE 46.4 Summarized Assignment of CYP2D6 Phenotype

Predicted Function of Star (*) Allele	Star Allele (Determined from Genotype)	Diplotype	Predicted Phenotype Based on Diplotype	Predicted Activity Score Based on Diplotype
Increased function		More than two copies of normal functional alleles	Ultrarapid metabolizer (~2% of people[a])	>2.0
Normal function	*1, *2, *33, *35	Two copies of functional or decreased-function alleles One functional allele and one decreased-function or nonfunctional allele	Extensive/normal metabolizer (~80% of people[a])	1.0–2.0
Decreased function	*9, *10, *17, *29, *36, *41	One decreased-function allele and one nonfunctional allele	Intermediate metabolizer (~10% of people[a])	0.5
Nonfunctional	*3, *4, *5, *6, *7, *8, *11, *12, *13, *14, *15, *16, *18, *19, *20, *21, *36, *38, *40, *42, *44, *68, *92, *100, *101	Two or more copies of nonfunctional alleles	Poor metabolizer (~8% of people[a])	0

[a]Actual prevalence is population-dependent; percentages shown here represent published estimates for the general population.
Data from Kroetz, D. L., Yee, S. W., & Giacomini, K. M. (2010). The pharmacogenomics of membrane transporters project: research at the interface of genomics and transporter pharmacology. *Clin Pharmacol Ther, 87*, 109–116.

example, the *CYP2D6* c.100C>T variant (rs1065852) is present in three of the alleles shown in Table 46.3. Misclassification of *CYP2D6*4* or *CYP2D6*36* as *CYP2D6*10* would incorrectly predict the phenotype. Allele frequencies vary among populations. The most common variant allele in Caucasians is *CYP2D6*4*, whereas the most common variant alleles in African American, Middle Eastern, and East Asian populations are *CYP2D6*17*, *CYP2D6*41*, and *CYP2D6*10*, respectively.

The assignment of phenotype predictions from the *CYP2D6* genotype is described in Table 46.4. An alternate approach to CYP2D6 phenotype characterization is the assignment of *activity scores*. Drug-specific dosing guidelines have been published by the Clinical Pharmacogenetics Implementation Consortium (CPIC) and other organizations such as the Dutch Pharmacogenomics Working Group (DPWG) for *CYP2D6*, and many other pharmacogenes (see Table 46.1 for examples).

BOX 46.2 Definition of Drug Inhibitors as per the US Food and Drug Administration

Strong Inhibitors
≥5-fold increase in the area under the curve (AUC), wherein the AUC refers to a plot of drug concentration in blood versus time after drug administration *or* >80% decrease in clearance (CL) of a drug from the blood expressed in terms of volume cleared per unit of time.

Moderate Inhibitors
≥2 but <5-fold increase in AUC *or* 50%–80% decrease in CL.

Weak Inhibitors
≥1.25 but <2-fold increase in AUC *or* 20%–50% decrease in CL.

Drug-drug or food-drug interactions often occur at the level of **drug metabolism** and are of particular concern under circumstances when multiple medications are prescribed and used by a single patient at one time. Drugs can be classified as inducers as well as strong, moderate, or weak inhibitors (Box 46.2). Some drugs may be both substrates and inducers or inhibitors for the same enzyme. A well-recognized food-drug interaction is the inhibition of CYP3A4 by grapefruit juice. For a person who is genetically an intermediate, extensive/normal, or ultrarapid metabolizer, the metabolic phenotype may be modified by drug-drug interactions. For a person who is genetically a poor metabolizer, drug-drug and/or food-drug interactions are not expected to affect drug and dose selections.

Codeine application. Codeine is a widely used medication, prescribed for pain management and to suppress coughing. Codeine must be activated by metabolism to morphine. This occurs through a demethylation reaction mediated primarily by CYP2D6 (Fig. 46.3). Therefore a CYP2D6 poor metabolizer would not activate codeine and should avoid this drug owing to the lack of predicted efficacy. An intermediate metabolizer may require higher doses of codeine than an extensive/normal metabolizer. The guidelines published by the CPIC suggest that alternative analgesia be sought should a CYP2D6 intermediate metabolizer fail to respond to codeine. An ultrarapid metabolizer should avoid codeine owing to the potential for toxicity. In a typical extensive/normal metabolizer, approximately 10% of a codeine dose is converted to morphine. Administration of codeine to a person with the ultrarapid metabolizer phenotype is a significant safety concern because higher than expected concentrations of morphine can be produced, leading to the risk of unintentional overdose and opioid toxicity. For example, when nursing women with an ultrarapid metabolizer phenotype for CYP2D6 are prescribed codeine, their babies can be exposed to higher than expected morphine concentrations in the breast milk. Children who receive codeine for pain control after tonsillectomy and adenoidectomy are also vulnerable to potentially deadly opioid-induced respiratory depression. A similar pharmacogenetic vulnerability to opioid-induced toxicity exists for other opioids that utilize CYP2D6 for activation, such as tramadol.

Antidepressant application. Antidepressant dosing is challenging because several weeks are typically required to assess efficacy, and optimizing dose may require several months of trial and error. Many antidepressant medications available today are substrates of CYP2D6, and many are also inhibitors. For example, fluoxetine is a selective serotonin reuptake inhibitor (SSRI) that is demethylated to form an active metabolite, norfluoxetine. Although several CYP enzymes are involved in the metabolism of fluoxetine, metabolism by CYP2D6 is considered the major route. CYP2D6 is also inhibited by fluoxetine and norfluoxetine, which are classified as strong inhibitors that lead to prolongation in the half-life of the drug and the metabolite within the first weeks of therapy. The CPIC has published guidelines based on CYP2D6 phenotype for SSRIs. Poor and ultrarapid metabolizers should avoid these drugs due to risk of toxicity and lack of efficacy, respectively.

Tricyclic antidepressants (TCAs) such as nortriptyline are also metabolized by CYP2D6. The relationship between drug concentrations and toxicity for CYP2D6 phenotypes and genotypes has been extensively characterized for nortriptyline. As shown in Table 46.5, lower doses of nortriptyline are recommended for a CYP2D6 poor metabolizer than for an extensive/normal metabolizer in order to produce similar serum concentrations of active drug. Nortriptyline, which is also an active metabolite of amitriptyline, is hydroxylated by CYP2D6 to form an inactive metabolite (10-hydroxy-nortriptyline). Clearance of amitriptyline, nortriptyline, and other TCAs is reduced by at least 50% in poor metabolizers of CYP2D6. TCAs have a narrow therapeutic index and are associated with severe adverse drug reactions that may be life-threatening. The CPIC has published dosing guidelines based on CYP2D6 phenotype for TCAs.

Cytochrome P450 2C family. The CYP2C family includes four genes that code for enzymes with the same name: CYP2C8, CYP2C9, CYP2C18, and CYP2C19. CYP2C9 and CYP2C19 are commonly involved in drug metabolism. CYP2C9 is involved in the metabolism of vitamin K–dependent anticoagulants (e.g., warfarin), anticonvulsants, nonsteroidal anti-inflammatory drugs, antidiabetic agents, cholesterol-lowering drugs, angiotensin receptor blockers, and drugs used to treat infection. Genotype-based dosing guidelines have been proposed for some CYP2C9 substrates, sometimes in combination with other pharmacogenes. The most common variant in Caucasians, *CYP2C9*2* (c.430C>T, rs1799853), decreases enzyme activity to approximately 12% in homozygotes. *CYP2C9*2* has a lower allele frequency in individuals of African and Native American ancestry and is not detected in Asians. The second most common variant in Caucasians, *CYP2C9*3* (c.1075A>C, rs1057910), decreases enzyme activity to approximately 5% in homozygotes. The allele frequency of *CYP2C9*3* is lower in people of Asian, African, and Native American ancestry. A low activity variant mainly present in African populations is *CYP2C9*8* (c.449G>A, rs7900194).

The CYP2C19 enzyme is involved in the metabolism of a number of therapeutic drugs, such as citalopram, diazepam,

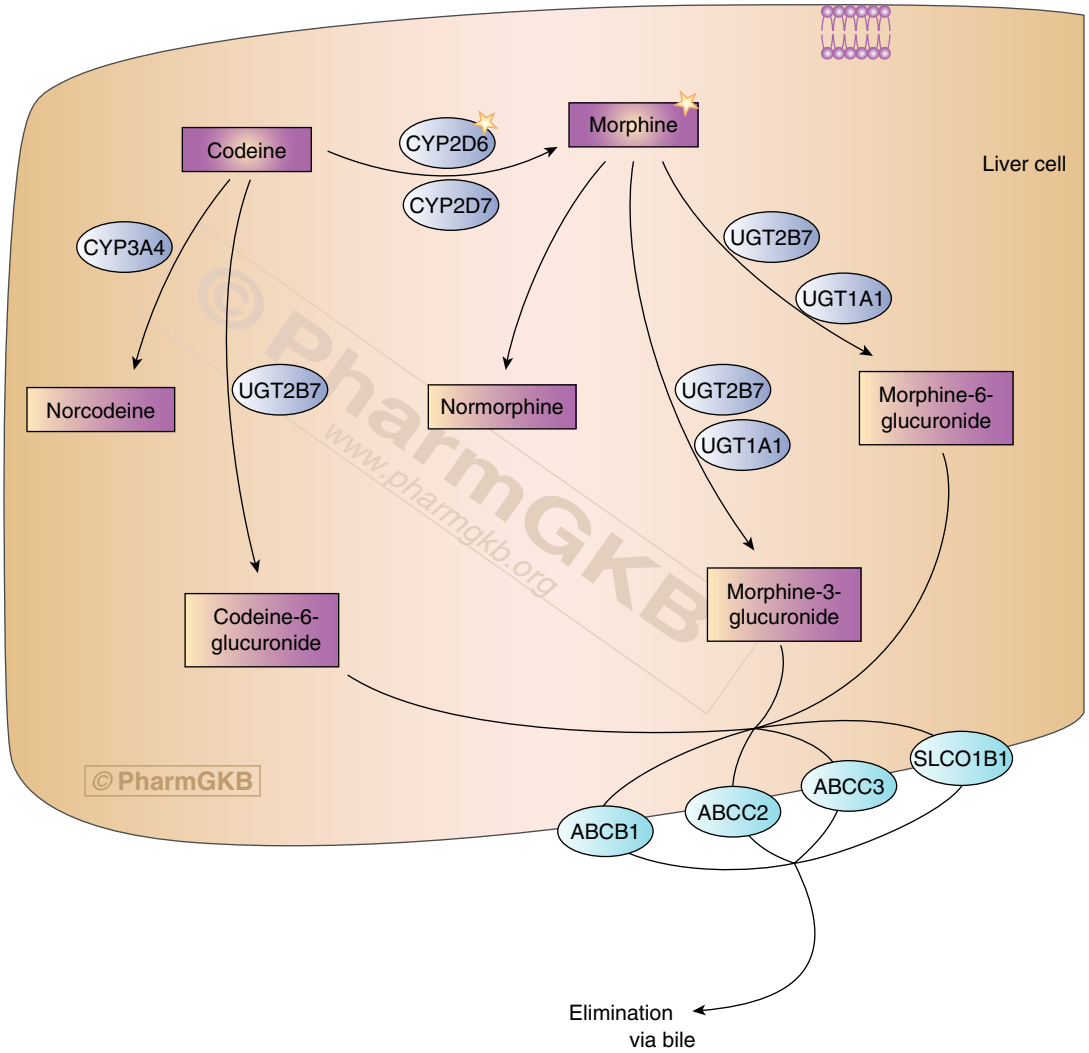


Fig. 46.3 Schematic illustration of codeine pharmacokinetics. *Purple boxes* indicate codeine and metabolites and the *blue ovals* are genes that code for drug metabolizing enzymes. Codeine is the prodrug, Morphine is the primary active metabolite, formed through a reaction mediated by CYP2D6. The *orange stars* indicate the importance of this metabolic activation. All other metabolites are thought to have little to no pharmacological activity. (From Thorn, C. F., Klein, T. E., & Altman, R. B. [2009]. Codeine and morphine pathway. *Pharmacogenet Genomics, 19*, 556–558. Reproduced with permission from the copyright holder, PharmGKB and Stanford University.)

TABLE 46.5 Examples of Dose Adjustments for Antidepressants Based on CYP2D6 or CYP2C19 Metabolizer Phenotype

			METABOLIZER PHENOTYPE			
Antidepressant	**CYP2C19**	**CYP2D6**	**Ultrarapid (%)**	**Extensive/Normal (%)**	**Intermediate (%)**	**Poor (%)**
Amitriptyline	X		141	105	87	70
		X	138	114	90	67
Citalopram	X		130	107	87	59
Desipramine		X	207	136	76	25
Doxepine		X	204	131	77	34
Fluvoxamine		X	147	117	89	68
Nortriptyline		X	195	133	72	50
Paroxetine		X	169	125	81	51
Trimipramine	X		154	107	76	45
		X	147	118	88	59
Venlafaxine	X		124	107	89	44
		X	172	109	92	77

Data from Stingl, J., & Viviani, R. (2015). Polymorphism in CYP2D6 and CYP2C19, members of the cytochrome P450 mixed-function oxidase system, in the metabolism of psychotropic drugs. *J Intern Med, 277*, 167–177.

omeprazole, and clopidogrel. Approximately 90% of CYP2C19 poor metabolizers are explained by *CYP2C19*2* (c.681G>A, rs4244285) and *CYP2C19*3* (c.636G>A, rs4986893). *CYP2C19*17* is responsible for rapid (one copy) and ultrarapid (two copies) phenotypes and is common in Caucasians and African Americans but less frequent in Hispanics and Asians. This allele arises from a promoter variant in the gene (c.-806C>T, rs12248560) that is associated with increased expression of the allele. However, the phenotype prediction is altered when multiple variants are detected. For example, the presence of both *CYP2C19*2* and *CYP2C19*17* is associated with an intermediate metabolizer phenotype.

Clopidogrel application. Clopidogrel inhibits platelet aggregation and is widely used in combination with aspirin to reduce the incidence of thrombotic and ischemic events in patients with coronary artery disease or acute coronary syndrome and/or after percutaneous coronary intervention with stenting. Yet many patients do not achieve adequate platelet inhibition; rates of drug resistance are estimated at 30% and are partly due to the fact that clopidogrel is a prodrug. Intermediate and poor CYP2C19 metabolism is associated with reduced formation of the active metabolite of clopidogrel, poor clopidogrel-induced platelet inhibition, and a higher incidence of major thrombotic events. The CPIC guidelines recommend drugs that do not require bioactivation by CYP2C19, such as prasugrel or ticagrelor, for CYP2C19 intermediate and poor metabolizers. No dose adjustment for clopidogrel is recommended for CYP2C19 ultrarapid or extensive/normal metabolizers.

Cytochrome P450 3A family and tacrolimus. The CYP 3A family includes four genes that code for enzymes with the same name: *CYP3A4, CYP3A5, CYP3A7* and *CYP3A43.* CYP3A7 is expressed primarily during the fetal period and the CYP3A43 is not significantly involved in drug metabolism. In contrast, CYP3A4 and CYP3A5 enzymes are responsible for the metabolism of approximately half of all currently used drugs. *CYP3A4*22* (g.15389C>T, rs35599367) is the most common loss-of-function allele for this gene; other variants are rare. However, the *CYP3A5*3* allele (c.6986A>G, rs776746), is a very common loss-of-function allele in many populations. Thus 70% to 90% of Caucasians and Asians are poor metabolizers for CYP3A5. The nonfunctional *CYP3A5*6* (c.14690G>A, rs10264272) and *CYP3A5*7* (c.27131_27132insT, rs41303343) alleles are relatively common in African populations but rare in Caucasians, Asians, and Middle Eastern populations. Approximately 50% of African Americans are predicted to be poor metabolizers.

The immunosuppressant drug tacrolimus is widely used to prevent rejection in solid organ transplantation. This drug has a narrow therapeutic index and doses are adjusted based on the results of routine TDM. Acute rejection may result if blood concentrations are insufficient, and life-threatening toxicity can ensue with excess dosing. Predicting a target dose for a patient pretherapeutically could reduce dosing errors, particularly in the critical early posttransplant period. Tacrolimus is the active drug that is inactivated by reactions mediated primarily by CYP3A4 and CYP3A5. *CYP3A5* poor metabolizer genotypes predict lower dose requirements than the extensive/normal metabolizer (e.g. *CYP3A5 *1/*1*) or intermediate metabolizer phenotypes (e.g., *CYP3A5 *1/*3*). The CPIC guideline suggests standard doses for poor metabolizers and 1.5 to 2 times higher doses for extensive/normal and intermediate metabolizers followed by TDM to optimize dose.

N-Acetyltransferases (NAT1 and NAT2). The *N*-acetyltransferase (NAT) polymorphism is one of the earliest pharmacogenetic targets recognized and characterized. NATs catalyze the transfer of an acetyl moiety from acetyl-CoA to homocyclic and heterocyclic arylamines and hydrazines. Substrates include drugs, carcinogens, toxicants, and possibly endogenous compounds. There are two forms of NAT, with 81% amino acid sequence identity, coded by genes of the same names (*NAT1* and *NAT2)*, and both enzymes have affinity for most substrates. NAT2 slow acetylators are common in many populations, particularly Egyptians, Europeans, and African Americans.

Genotyping can predict NAT2 phenotype, but testing has limited clinical utility. For many NAT drug substrates, the acetylator status is managed through nongenetic approaches. For example, the antiarrhythmic drug procainamide is routinely monitored along with the active metabolite *N*-acetylprocainamide in timed blood samples. The dose of procainamide is adjusted based on the parent/metabolite ratio. For isoniazid, the neuropathic adverse drug reactions are linked to pyridoxine deficiency and can be avoided by the coadministration of pyridoxine to all patients. Also, isoniazid dosing intervals can be changed from once a week to twice a week to compensate for the fact that rapid acetylators are less likely to respond to the conventionally administered dose.

Glucose-6-phosphate dehydrogenase and rasburicase application. Glucose-6-phosphate dehydrogenase (G6PD) is one of the most common human enzymatic defects that, triggered by oxidative stress, can lead to neonatal jaundice and hemolytic anemia. G6PD catalyzes the first step of the pentose phosphate pathway, resulting in the production of nicotinamide adenine dinucleotide phosphate (NADPH). NADPH is essential to cope with oxidative stress in red blood cells. Factors such as infections, certain drugs, and ingesting fava beans can increase the levels of reactive oxygen species, thus causing red blood cells to undergo hemolysis faster than the body can replace them. The loss of red blood cells causes the signs and symptoms of hemolytic anemia, such as dark urine, enlarged spleen, fatigue, rapid heart rate, shortness of breath and jaundice, which are the characteristic features of G6PD deficiency

More than 400 mutations in the *G6PD* gene have been identified; most are missense mutations that affect protein stability. The nomenclature for *G6PD* variants is based on the location where the variant was first described. All *G6PD* variants are broadly divided into five classes according to the resulting level of enzyme activity, with class I being the most severely dysfunctional and class V having the highest enzyme activity (Table 46.6). Targeted *G6PD* genotyping can establish

TABLE 46.6 Dosing for Rasburicase and *G6PD* as Recommended by the Clinical Pharmacogenetics Implementation Consortium

Phenotype/Genotype	Examples of Diplotypes	Implications for Outcome	Dosing Recommendations
Normal. A male carrying a nondeficient (class IV) allele or a female carrying two nondeficient (class IV) alleles.	Male: B, Sao Boria. Female: B/B, B/Sao Boria.	Low or reduced risk of hemolytic anemia.	Use label-recommended dosage and administration.
Deficient or deficient with CNSHA. A male carrying a class I, II, or III allele or a female carrying two deficient class I–III alleles.	Male: A–, Orissa, Kalyan-Kerala, Mediterranean, Canton, Chatham, Bangkok, Villeurbanne. Female: A–/A–, A–/Orissa, Orissa/Kalyan-Kerala, Mediterranean/Mediterranean, Chatham/Mediterranean, Canton/Viangchan, Bangkok/Bangkok, Bangkok/Villeurbanne.	At risk of acute hemolytic anemia.	Select alternate drug; avoid use of rasburicase.
Variable. A female carrying one nondeficient (class IV) and one deficient (class I–III variants) allele.	B/A–, B/Mediterranean, B/Bangkok.	Unknown risk of hemolytic anemia.	G6PD enzyme activity must be measured to determine phenotype or rasburicase use should be avoided; consider allopurinol.

Rasburicase and *G6PD* data from Relling, M. V., McDonagh, E. M., Chang, T., et al. (2014). Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines for rasburicase therapy in the context of G6PD deficiency genotype. *Clin Pharmacol Ther*, *96*, 169–174.

a molecular diagnosis of G6PD deficiency; however, prediction of drug response can be difficult without testing G6PD enzyme activity levels.

Rasburicase is a drug approved by the FDA for the prophylaxis and treatment of hyperuricemia during chemotherapy in adults and children with lymphoma, leukemia, and solid tumors. When chemotherapy is administered, cancer cells are destroyed, releasing large amounts of uric acid into the blood. Rasburicase is a recombinant urate oxidase enzyme that works by breaking down uric acid to allantoin and hydrogen peroxide, which are eliminated from the body by the kidneys. The pegylated form of urate oxidase, pegloticase, is also FDA-approved for the treatment of refractory gout. Notably, both rasburicase and pegloticase carry an FDA boxed warning and are contraindicated for use in patients with known G6PD deficiency due to mutations in the *G6PD* gene. Additionally, the CPIC guidelines (see Table 46.6) recommend an alternative therapy (e.g., allopurinol) for G6PD-deficient patients.

Thiopurine S-methyltransferase and thiopurines application. Thiopurine S-methyltransferase (TPMT) is a metabolic enzyme that catalyzes the inactivation of thiopurine drugs (i.e., azathioprine, mercaptopurine, and thioguanine) by S-methylation, thus preventing the formation of thioguanine nucleotides (TGNs). These drugs are analogs of the nucleic acid guanine and are incorporated into RNA and DNA by phosphodiester linkages, ultimately inhibiting several metabolic pathways and inducing apoptosis. In addition, mercaptopurines are metabolized to methyl-thioinosine monophosphate, which inhibits de novo purine synthesis and cell proliferation, thus adding another mechanism of cytotoxicity. However, approximately 10% of many populations have intermediate levels of TPMT activity and approximately 0.3% have low or undetectable enzyme activity, which results in significantly increased risks for thioguanine nucleotides toxicity and life-threatening myelosuppression.

Over 30 variant alleles of the *TPMT* gene have been identified, many of which are missense mutations associated with decreased in vitro activity. The star (*) nomenclature is used to describe *TPMT* haplotypes. The first variant allele described, *TPMT*2* (c. 238G>C; rs1800462), is rare in all populations studied. *TPMT*3A* contains two missense variants *in cis*, c.460G>A (rs1800460) and c.719A>G (rs1142345), while *TPMT*3B* and *TPMT*3C contain only one, c.460G>A (rs1800460) and c.719A>G (rs1142345), respectively. *TPMT*3A is the most common variant allele associated with low TPMT activity in Caucasians. The assumed diplotype is *1/*3A when both variants are present in an individual of Caucasian descent, although *3B/*3C cannot be ruled out. If *3B/*3C is suspected, phenotypic testing should be used to distinguish between *1/*3A and *3B/*3C. *TPMT*3C* is the most common variant allele in East Asian and African American populations.

Thiopurines are used for the treatment of childhood acute lymphoblastic leukemia, autoimmune diseases, inflammatory bowel diseases, lupus, and transplantation. With conventional doses of thiopurines, individuals who inherit two loss-of-function *TPMT* alleles universally experience severe myelosuppression; a high proportion of heterozygous individuals show moderate to severe myelosuppression, whereas individuals in whom no variant is detected have a low risk of myelosuppression. The CPIC guidelines suggest dose reductions for both intermediate and poor metabolizers (Table 46.7).

Testing TPMT enzyme activity directly (phenotyping) is an alternative to *TPMT* genotyping that is challenged by storage and stability concerns. TPMT enzyme activity depends

TABLE 46.7 Dosing for Thiopurines and *TPMT* as Recommended by the Clinical Pharmacogenetics Implementation Consortium

Phenotype/Genotype	Examples of Diplotypes	Implications for Outcome	Dosing Recommendations
Homozygous wild-type or normal, high activity (two functional *1 alleles)	*1/*1	"Normal" (low) concentrations of thioguanine nucleotide metabolites. Note that thioguanine nucleotide metabolite concentrations with thioguanine are 5–10x higher than after mercaptopurine or azathioprine.	Use label-recommended dosage and administration. Adjust based on degree of myelosuppression and disease-specific guidelines. Allow 2–4 weeks to reach steady state after each dose adjustment.
Heterozygote or intermediate activity (one functional allele—*1, plus one nonfunctional allele)	*1/*2, *1/*3A, *1/*3B, *1/*3C, *1/*4	Moderate to high concentrations of thioguanine nucleotide metabolites.	Reduce dose by 30%–50% and adjust based on degree of myelosuppression and disease-specific guidelines. Allow 2–4 weeks to reach steady state after each dose adjustment.
Homozygous or compound heterozygous variant, mutant, low, or deficient activity (two nonfunctional alleles)	*3A/*3A, *2/*3A, *3C/*3A, *3C/*4, *3C/*2, *3A/*4	Extremely high concentrations of thioguanine nucleotide metabolites; fatal toxicity possible without dose decrease.	Reduce daily dose by 10-fold and dose thrice weekly instead of daily; adjust based on degree of myelosuppression and disease-specific guidelines. Allow 4–6 weeks to reach steady state after each dose adjustment. For nonmalignant conditions, consider nonthiopurine alternative.

Thiopurines and *TPMT* data from Relling, M. V., Gardner, E. E., Sandborn, W. J., et al. (2011). Clinical Pharmacogenetics Implementation Consortium guidelines for thiopurine methyltransferase genotype and thiopurine dosing. *Clin Pharmacol Ther*, *89*, 387–391.

on stable enzyme activity between the times of blood collection and analytic testing. In addition, because the TPMT enzyme is expressed in red blood cells, testing is limited to patients who have not received a blood transfusion over the previous weeks and who have healthy red blood cells, which is often not the case when acute lymphoid leukemia is diagnosed. Coadministered drugs such as ibuprofen and thiazide diuretics may inhibit enzyme activity such that an inaccurate phenotype prediction may be assigned. Nonetheless, as with other pharmacogenetics applications designed to predict metabolic phenotype, dose optimization requires clinical and/or therapeutic drug monitoring.

Solute Carrier Organic Anion Transporter Family Member 1B1 and Simvastatin Application

Drug transporters are membrane proteins that mediate the influx and efflux of chemicals using active and passive mechanisms. They can be divided into two major families: the adenosine triphosphate (ATP) binding cassette subfamily B member 1 (e.g., ABCB1) and the solute carrier family (e.g., SLCO1B1). The solute carrier organic anion-transporting polypeptide 1B1 (OATP1B1) is one of the main influx transporters expressed in hepatocytes. OATP1B1 substrates include several drugs as well as endogenous compounds such as bile acids, thyroid hormones, and estrogens. OATP1B1 is encoded by the solute carrier organic anion transporter gene *SLCO1B1*. Of more than 40 known *SLCO1B1* variants, two have been well characterized: 388A>G (c.492A>G, rs2306283) and 521T>C (c.625T>C, rs4149056). These two variants are in partial linkage disequilibrium and lead to important haplotypes that contain neither, either, or both polymorphisms. The star (*) nomenclature has been applied. Clinical relevance of haplotypes that contain 521C (*5, *15, and *17) has been shown especially for simvastatin and high-dose methotrexate. In addition, OATP1B1 transport activity is inhibited by several drugs including atorvastatin, rifampicin, immunosuppressants (cyclosporine and tacrolimus), and the HIV drugs (ritonavir, lopinavir, and nelfinavir).

Statins, such as simvastatin, are safe and effective in the majority of patients but are associated with muscle toxicity in approximately 10%. The clinical spectrum of muscle toxicity ranges from mild forms to pain with evidence of muscle degradation (myopathy) and finally to severe muscle damage with acute kidney injury (rhabdomyolysis). Homozygous carriers of *SLCO1B1* 521C are exposed to higher levels of the active simvastatin acid due to decreased hepatic uptake and have an increased risk of myopathy during treatment with simvastatin. The *SLCO1B1* 512C variant does not explain all simvastatin myopathy; other genetic contribution is currently sought.

Associations With Pharmacodynamics

Vitamin K Epoxide Reductase Complex Subunit 1 and Warfarin Application

Vitamin K epoxide reductase (VKOR) performs the rate-limiting step in the vitamin K cycle. Impairment of VKOR

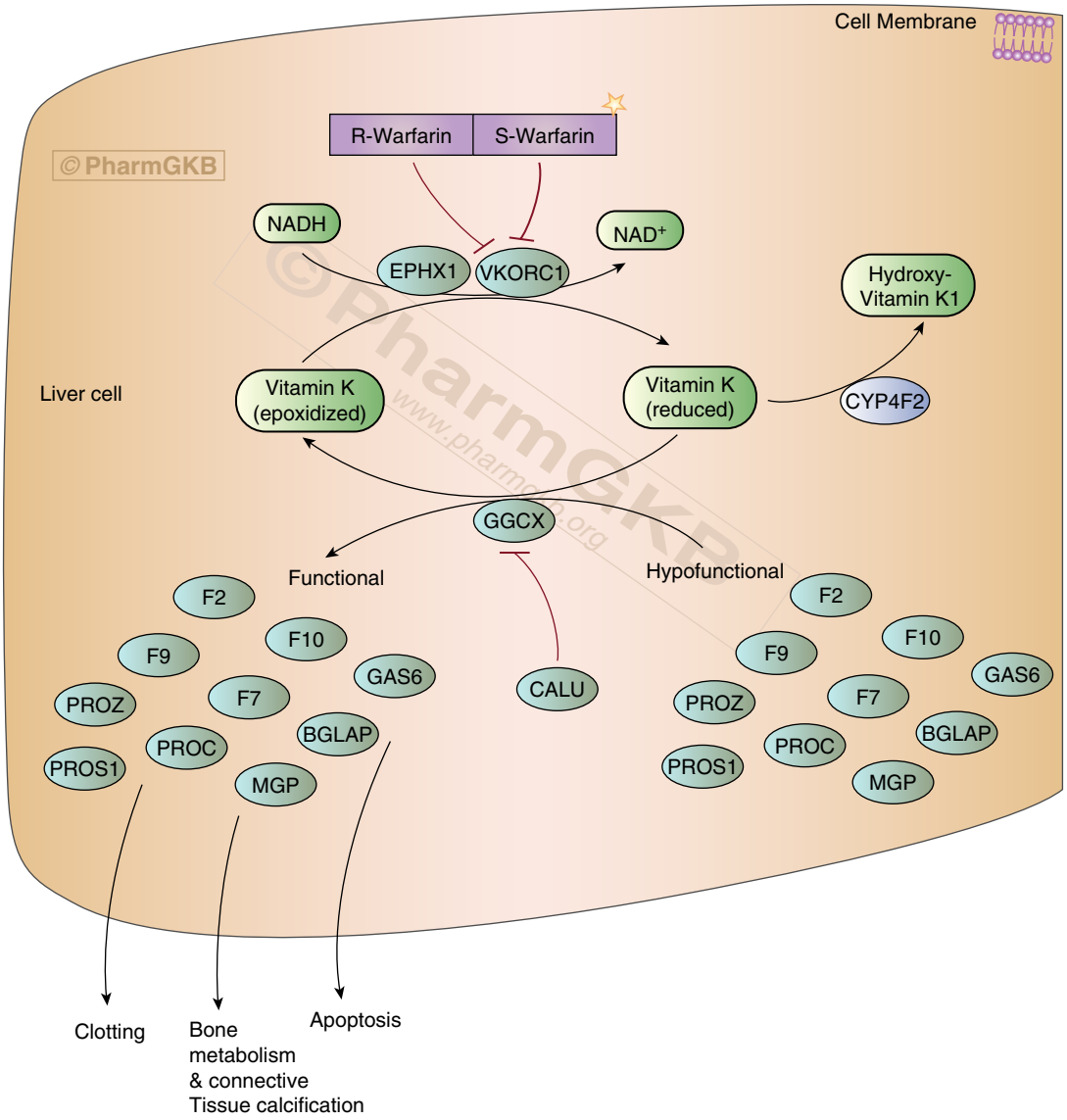


Fig. 46.4 Schematic illustration of warfarin pharmacodynamics. *Purple boxes* show the two enantiomers of warfarin. *Green rounded boxes* are cofactors important for reductive metabolism. *Blue ovals* are genes that code for enzymes or coagulation factors. Warfarin is administered as a racemic mixture of R- and S-enantiomers with S-warfarin being the more potent inhibitor of vitamin K epoxide reductase (VKORC1) in the vitamin K cycle. Other enzymes involved in the vitamin K cycle are epoxide hydrolase 1 (EPHX) and gamma-glutamyl carboxylase (GGCX). The vitamin K cycle is further regulated by calumenin (CALU) that inhibits GGCX and by oxidation of reduced vitamin K by cytochrome P450 4F2. The vitamin K cycle activates the coagulation factors FII, FVII, FIX and FX, and proteins C, S and Z into their functional forms. Other vitamin K–dependent proteins activated are the apoptotic growth arrest specific 6 (Gas-6); the bone gamma-carboxyglutamate (Gla) protein (BGLAP), which regulates bone remodelling; and the matrix Gla protein (MGP), which inhibits osteogenic factors. (From Whirl-Carrillo, M., McDonagh, E. M., Hebert, J. M., et al. [2012]. Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther*, *92*[4], 414–417, PMID: 22992668. Reproduced with permission from the copyright holder, PharmGKB and Stanford University.)

limits activation of coagulation factors II, VII, IX, and X and explains the anticoagulant effect of warfarin (Fig. 46.4). Warfarin is a widely used oral anticoagulant used to treat deep venous thrombosis and to prevent thromboembolic disease. Common *VKORC1* variants are associated with warfarin sensitivity. The allele frequency of the warfarin-sensitive *VKORC1* haplotype is low in African Americans, moderate in whites, and high in Asians. It is not known which polymorphism on the haplotype mediates the effect. Most clinical tests detect the upstream *VKORC1* variant c.-1639G>A (rs9923231), typically along with *CYP2C9*2* and *CYP2C9*3*. Additional variants in these and other genes, such as *CYP4F2*, appear to have little importance for warfarin dosing in Caucasians but may be important in other populations.

Individual response to warfarin is highly variable and is monitored by the INR (see Box 46.1). It is influenced by

clinical factors such as age, body size, diet, comorbidities, and interacting medications as well as by genetic variation. Carriers of variant alleles *CYP2C9* and *VKORC1* are more sensitive to warfarin and risk being overdosed by standard doses. Patients who are sensitive to warfarin spend a higher proportion of time above target range in the first 90 days of treatment and have an increased risk of bleeding compared with normal responders.

Several algorithms that predict individual warfarin dose requirements have been developed. These algorithms usually include *CYP2C9 *2* and *3*, a *VKORC1* genetic variant, and clinical factors such as age, body size, and interacting drugs. Dose revision algorithms also include previous warfarin doses and existing INR data. The performance of these algorithms varies between populations partly due to different frequencies of the included variables.

It has been predicted that dosing guided by *CYP2C9* and *VKORC1* would improve the management of warfarin, particularly for warfarin-sensitive patients. In 2013, two large randomized clinical trials on genotype-guided warfarin dosing, EU-PACT and COAG, evaluated the primary outcome time in the therapeutic INR range. In EU-PACT, time in range improved with genotype-guided dosing, whereas no difference was seen in COAG. Different dosing strategies and allele frequencies due to continental ancestry explain part of the divergent results. In 2017, a third large randomized controlled trial, the Genetics-InFormatics Trial (GIFT), showed a 27% relative reduction in the composite outcome for the genotype-guided arm, including venous thromboembolism, major hemorrhage, and INR equal to or greater than 4. CPIC has also published an update on genotype-guided warfarin dosing that takes continental ancestry into account.

Cystic Fibrosis Transmembrane Conductance Regulator and Ivacaftor Application

Cystic fibrosis (CF) is one of the most common lethal autosomal recessive diseases in Caucasians, with an estimated incidence of 1 in 2500 to 3300 live births. Worldwide, approximately 70,000 individuals have CF, the majority of whom are diagnosed before the age of 1 year. The gene mutated in CF is the cystic fibrosis transmembrane conductance regulator *(CFTR)*. More than 1800 *CFTR* mutations have been identified. The Human Genome Variation Society (HGVS) nomenclature is used to describe these mutations leading to a complex multisystem disease that affects the respiratory tract, pancreas, intestine, male genital tract, hepatobiliary system, and exocrine system. There is, however, wide variability in clinical presentation, severity, and the rate of disease progression between patients, which can be influenced by the underlying *CFTR* genotype as well as other genetic modifiers and environmental factors.

Traditionally the treatment of CF has focused on ameliorating symptoms (e.g., fighting infection, thinning mucus, and reducing inflammation) rather than directly targeting the genetic cause. Ivacaftor is the first FDA-approved therapeutic developed to target a specific *CFTR* defect by improving the transport of sodium through the CFTR ion channel. It was originally indicated in CF patients 6 years of age and older who have a specific variant (c.1652G>A), but the indication now includes additional CFTR gating defect variants. In all these variants, CFTR is trafficked correctly to the epithelial cell surface, but once there, the protein cannot transport chloride through the channel. Coadministration of CYP3A inducers (e.g., rifampin, St. John's wort) that substantially decrease exposure to ivacaftor is not recommended.

Human Leukocyte Antigen Complex and Applications

The HLA complex is a cornerstone of the immune system because of its involvement in the identification of foreign proteins. The HLA proteins produced from the associated genes are expressed on the surface of nearly all cells, where they bind to peptides that are displayed to the circulating cells of the immune system. If the immune system recognizes the peptides as foreign (e.g., viral or bacterial peptides), it responds by triggering the infected cell to self-destruct. The HLA complex is highly variable and its classification is very complicated. The *HLA-B* gene is best characterized from the perspective of pharmacogenetics, based on the association of variants with drug hypersensitivity, specifically severe and life-threatening cutaneous reaction syndromes, such as Stevens-Johnson syndrome and toxic epidermal necrolysis. Because these hypersensitivity reactions are not dose-dependent, they are classified as type B adverse drug reactions. Pretherapeutic testing has been recommended for select applications.

Published clinical consensus guidelines exist for *HLA-B*15:02* and the anticonvulsants carbamazepine and phenytoin and for *HLA-B*57:01* and the HIV drug abacavir. Typing of *HLA-B*15:02* in patients of East Asian origin, where this allele is common, and of *HLA-B*57:01* in all patients, substantially reduces risk of the associated adverse drug reaction. A summary of CPIC guidelines for carbamazepine or abacavir based on HLA alleles is shown in Table 46.8.

FUTURE DIRECTIONS

The future success of pharmacogenetics depends on understanding the following: (1) the interrelationship of the various proteins involved in pharmacokinetics and pharmacodynamics, (2) the balance between major versus minor pathways, and (3) the mechanisms of action of specific drugs. Standardization of test content and interpretation is needed. Although massively parallel sequencing will increase the complexity of pharmacogenetics, it will also improve our understanding and the utility of genotype-phenotype relationships. Additional clinical studies, particularly prospective and outcome-related studies, are needed to guide implementation of pharmacogenetic findings in the clinic. Finally, closer integration of pharmacy professionals with clinical laboratories and clinicians and increased availability of cost-effective commercial testing will improve the success of pharmacogenetic applications.

TABLE 46.8 Dosing for Carbamazepine and Phenytoin with *HLA-B*15:02* and Abacavir with *HLA-B*57:01* as Recommended by the Clinical Pharmacogenetics Implementation Consortium

Phenotype/Genotype	Implications for Outcome	Dosing Recommendations
HLA-B*15:02		
Absence of **15:02* alleles (may be reported as "negative" on a genotyping test)	Low or reduced risk of carbamazepine and phenytoin associated hypersensitivity	Use label-recommended dosage and administration.
Presence of at least one **15:02* allele (may be reported as "positive" on a genotyping test)	Increased risk of carbamazepine or and phenytoin hypersensitivity	Select alternative drug; avoid use of carbamazepine and phenytoin.
HLA-B*57:01		
Absence of **57:01* alleles (may be reported as "negative" on a genotyping test)	Low or reduced risk abacavir hypersensitivity	Use label-recommended dosage and administration.
Presence of at least one **57:01* allele (may be reported as "positive" on a genotyping test)	Increased risk of abacavir hypersensitivity	Select alternate drug; avoid use of abacavir.

See original references for complete details.
HLA-B*15:02 data from Leckband, S. G., Kelsoe, J. R., Dunnenberger, H. M., et al. (2013). Clinical Pharmacogenetics Implementation Consortium guidelines for HLA-B genotype and carbamazepine dosing. *Clin Pharmacol Ther*, *94*, 324–328; Caudle, K. E., Rettie, A. E., Whirl-Carrillo, M., et al. (2014). Clinical pharmacogenetics implementation consortium guidelines for CYP2C9 and HLA-B genotypes and phenytoin dosing. *Clin Pharmacol Ther*, *96*, 462–468.
HLA-B*57:01 data from Martin, M. A., Klein, T. E., Dong, B. J., et al. (2012). Clinical pharmacogenetics implementation consortium guidelines for HLA-B genotype and abacavir dosing. *Clin Pharmacol Ther, 91*, 734–738.

REREVIEW QUESTIONS

1. How a drug elicits responses, whether the responses are desirable or undesirable, is described by which one of the following?
   a. Pharmacodynamics
   b. Phenocopying
   c. Pharmacokinetics
   d. Pharmacogenetics
   e. Pharmacogenomics
2. A drug that is administered in an inactive form that must be metabolized to an active form is referred to as a(n)
   a. Cytochrome P450 (CYP450) enzyme
   b. Antimetabolite
   c. Prodrug
   d. Poor metabolizer
   e. Drug metabolite
3. Which of the following is an active drug that gets inactivated by metabolism?
   a. Clopidogrel
   b. Codeine
   c. Tamoxifen
   d. Warfarin
   e. Irinotecan
4. What is the drug-metabolizing phenotype when an individual has two nonfunctional alleles?
   a. Extensive/normal metabolizer
   b. Intermediate metabolizer
   c. Rapid metabolizer
   d. Poor metabolizer
   e. Ultrarapid metabolizer
5. What characterizes a type A adverse drug reaction?
   a. It is caused by the drug's pharmacological effects
   b. It is independent of dose
   c. It is an unexpected effect
   d. The drug class can never be given again
   e. It can be caused by the immune system
6. A pathologic disorder linked to the ingestion of fava beans, commonly described as favism, is caused by an enzyme deficiency that also has pharmacogenetic implications. Which gene is involved?
   a. *CFTR*
   b. *TPMT*
   c. *G6PD*
   d. *NAT2*
   e. *CYP2C19*

SUGGESTED READINGS

Alfirevic, A., & Pirmohamed, M. (2017). Genomics of adverse drug reactions. *Trends in Pharmacological Sciences*, *38*(1), 100–109.

Caudle, K. E., Dunnenberger, H. M., Freimuth, R. R., Peterson, J. F., Burlison, J. D., Whirl-Carrillo, M., et al. (2017). Standardizing terms for clinical pharmacogenetic test results: Consensus terms from the Clinical Pharmacogenetics Implementation Consortium (CPIC). *Genetics in Medicine*, *19*(2), 215–223.

Caudle, K. E., Klein, T. E., Hoffman, J. M., Muller, D. J., Whirl-Carrillo, M., Gong, L., et al. (2014). Incorporation of pharmacogenomics into routine clinical practice: The Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline development process. *Current Drug Metabolism*, *15*(2), 209–217.

Collins, S. L., Carr, D. F., & Pirmohamed, M. (2016). Advances in the pharmacogenomics of adverse drug reactions. *Drug Safety*, *39*(1), 15–27.

Ehmann, F., Caneva, L., Prasad, K., Paulmichl, M., Maliepaard, M., Llerena, A., et al. (2015). Pharmacogenomic information in drug labels: European Medicines Agency perspective. *The Pharmacogenomics Journal*, *15*(3), 201–210.

Luzum, J. A., Pakyz, R. E., Elsey, A. R., Haidar, C. E., Peterson, J. F., Whirl-Carrillo, M., et al. (2017). Pharmacogenomics Research Network Translational Pharmacogenetics Program. *Clinical Pharmacology and Therapeutics*, *102*(3), 502–510.

Maggo, S. D., Savage, R. L., & Kennedy, M. A. (2016). Impact of new genomic technologies on understanding adverse drug reactions. *Clinical Pharmacokinetics*, *55*(4), 419–436.

Kalman, L. V., Agúndez, J., Appell, M. L., Black, J. L., Bell, G. C., Boukouvala, S., et al. (2016). Pharmacogenetic allele nomenclature: International workgroup recommendations for test result reporting. *Clinical Pharmacology and Therapeutics*, *99*(2), 172–185.

Maranville, J. C., & Cox, N. J. (2016). Pharmacogenomic variants have larger effect sizes than genetic variants associated with other dichotomous complex traits. *Pharmacogenomics J*, *16*(4), 388–392.

Mathias, P. C., Hendrix, N., Wang, W. J., Keyloun, K., Khelifi, M., Tarczy-Hornoch, P., & Devine, B. (2017). Characterizing pharmacogenomic-guided medication use with a clinical data repository. *Clinical Pharmacology and Therapeutics*, *102*(2), 340–348.

Pratt, V. M., Everts, R. E., Aggarwal, P., Beyer, B. N., Broeckel, U., Epstein-Baak, R., et al. (2016). Characterization of 137 genomic DNA reference materials for 28 pharmacogenetic genes: A GeT-RM collaborative project. *The Journal of Molecular Diagnostics*, *18*(1), 109–123.

Relling, M. V, & Evans, W. E. Pharmacogenomics in the clinic. *Nature*, 526(7573):343–350.

Sim, S. C., & Ingelman-Sundberg, M. (2013). Update on allele nomenclature for human cytochromes P450 and the Human Cytochrome P450 Allele (CYP-allele) Nomenclature Database. *Methods in Molecular Biology*, *2013*(987), 251–259.

Thorn, C. F., Klein, T. E., & Altman, R. B. (2013). PharmGKB: The pharmacogenomics knowledge base. *Methods in Molecular Biology*, *1015*, 311–320.

van der Wouden, C. H., Cambon-Thomsen, A., Cecchin, E., Cheung, K. C., Dávila-Fajardo, C. L., Deneer, V. H., et al. (2017). Implementing pharmacogenomics in Europe: design and implementation strategy of the ubiquitous Pharmacogenomics Consortium. *Clinical Pharmacology and Therapeutics*, *101*(3), 341–358.

Wang, B., Canestaro, W. J., & Choudhry, N. K. (2014). Clinical evidence supporting pharmacogenomic biomarker testing provided in US Food and Drug Administration drug labels. *JAMA Internal Medicine*, *174*(12), 1938–1944.

Zhang, G., Zhang, Y., Ling, Y., & Jia, J. (2015). Web resources for pharmacogenomics. *Genomics Proteomics Bioinformatics*, *13*(1), 51–54.

47

Molecular Principles

John Greg Howe, Elaine R. Mardis, Jason Y. Park, Carl T. Wittwer

OBJECTIVES

1. Compare and contrast the structure and function of DNA and RNA.
2. Identify five different types of RNA in the cell and describe their functions.
3. Detail how the information in DNA is used to make protein through transcription and translation.
4. Describe the genetic code and why it is important.
5. Compare and contrast the human genome with bacterial, viral, and fungal genomes.
6. Identify different kinds of sequence variants and discuss how they can cause disease.
7. Compare and contrast the different types of single nucleotide variants.
8. Describe three reversible reactions on histones that can increase accessibility for transcription, repair, or replication.
9. Learn how to name single nucleotide variants.
10. Describe the spectrum of genomic variations found in cancer.

KEYWORDS AND DEFINITIONS

Alleles Different variants of a gene.
Annotation Biologic information attached to a genomic sequence.
Assembly Reconstruction of short-sequence reads on a scaffold of reference DNA.
Autosome A non–sex chromosome.
Base pair Complementary base pairing of guanine and cytosine and adenine and thymine through hydrogen bonding.
Centromere The region of a chromosome between the small and long arms that is used during mitosis to attach the spindle fibers.
Chromatin A structure in the eukaryotic nucleus composed of DNA and histone proteins.
Chromosome A nucleic acid structure bound by protein containing all or a portion of an organism's genetic information.
Codon A three-base code found in messenger RNA, used in translation to specify the amino acid to be incorporated in the growing polypeptide chain.
Copy number variant (CNV) A structural variant of a large region of the genome that has been deleted or duplicated.
Deletion A DNA sequence that is missing in one sample compared with another. Deletions may be as small as one nucleotide or as large as an entire chromosome.
Deoxyribonucleic acid (DNA) Repository of the genetic material in an organism.
DNA methylation Addition of a methyl group to a cytosine, used to regulate gene expression.
Epigenetics Modifications that affect DNA packaging and accessibility without changing the DNA sequence.
Eukaryote An organism that contains a nucleus and whose DNA is in chromosomes.
Exon Portion of a gene that codes for protein.
Exonuclease An enzyme that cuts in the middle of a stretch of DNA at a specific sequence of bases.
FASTQ file A text output file of sequencing reads in a run, along with the quality scores of each position.
Fusion A translocation, inversion, large deletion, or large duplication resulting in a hybrid gene formed from originally separate genes.
Gene DNA structure that codes for the production of a protein.
Gene expression Synthesis of a gene product in the form of an RNA or protein from a gene.
Genetic code Collection of three base sequences called codons used to translate the nucleic acid information into amino acids during protein synthesis.
Genome The complete genetic content of an organism.
Genotype The unique genetic information of an organism or cell.
Heteroplasmic A mixture of more than one type of mitochondrial sequence in one cell.
Histones Basic proteins that bind to DNA to form nucleosomes, basic units of chromatin.

Indels Originally, designating a unique class of sequence variants including both insertions and deletions, resulting in an overall change in the number of base pairs. Today the term more commonly refers to either insertions or deletions or a combination thereof.

Insertion An extra DNA sequence that is present in one sample compared with a reference sequence.

Intergenic DNA sequence between genes.

Intron DNA between exons that do not code for protein and are removed by RNA splicing.

Micro RNA (miRNA) Small RNAs that are not translated and function to regulate translation.

Missense variant A nucleotide substitution that changes a codon to the code for a different amino acid. Although these sequence changes are commonly referred to as missense "mutations," this is strictly a misnomer because missense variants may be benign and cause no disease. The preferred nomenclature for nucleotide changes in clinical testing is now "variants" rather than mutations.

Mitochondrial DNA Genetic information found uniquely in mitochondria that codes for proteins involved with energy production.

Mutation A disease-causing sequence variation. Historically the term has been interchangeable with *variant* to describe any change in a DNA sequence regardless of its relation to disease causation. For current clinical descriptions or reporting, the use of *mutation* is reserved for the scenario when disease causation is known. The preferred nomenclature for a disease-causing variation is now "pathogenic" or "likely pathogenic" variant. There is also an evolving tendency to describe whether a DNA variant causes disease within a specific clinical context.

Nonsense variant A nucleotide substitution that results in a stop codon, prematurely terminating the protein. The preferred nomenclature for nucleotide changes in clinical testing is now "variants" rather than mutations.

Nucleases Enzymes that cleave DNA to smaller fragments.

Nucleic Acid A polymer made of nucleotides consisting of a sugar, a phosphate group, and a nitrogenous base.

Nucleosome A unit of DNA that contains 146 bases and is wound around 8 histone proteins.

Nucleotide A basic unit of DNA and RNA consisting of a sugar, a phosphate group, and a nitrogenous base.

Phenotype The physical and biologic characteristics of a genotype.

Phred score Estimate of the error probability for a base called in DNA sequencing. It is represented as a Q score; the higher the number, the higher the probability of a correct call.

Plasmid An extrachromosomal ring of double-stranded closed DNA found in bacteria.

Polymerase A protein that is able to connect individual molecules into a single strand.

Promoter A DNA sequence usually upstream from the start of transcription of the gene that regulates its expression.

Pseudogene A genetic element that does not code for a functional gene product, usually because of accumulated sequence variations.

Purine A double-ring structure that constitutes the foundation for adenine and guanine, two of the bases in DNA.

Pyrimidine A single ring structure that constitutes the foundation for thymine and cytosine, two of the bases in DNA and uracil and one of the bases of RNA.

Recombination The rearrangement of DNA through the breaking and rejoining of DNA strands to create a different sequence.

Replication The process that synthesizes a copy of an organism's DNA.

Ribonucleic acid (RNA) A polymer made of ribonucleotides consisting of ribose, a phosphate group, and four bases—guanine, adenine, cytosine, and uracil.

Ribosome A large structure that functions to convert messenger RNA to protein.

Short tandem repeat (STR) A simple sequence repeat that is 1 to 13 bases long.

Simple sequence repeat (SSR) A sequence from 1 to 500 bases that is repeated end to end. If the repeat unit is 1 to 13 bases, it is a microsatellite or STR. If the repeat is 14 to 500 bases, it is a minisatellite.

Single nucleotide polymorphism (SNP) A benign single nucleotide variant (substitution, deletion, or insertion) that occurs in a population at a frequency of at least 1%.

Single nucleotide variation (SNV) A single nucleotide variant (substitution, deletion, or insertion). SNVs may be benign or may cause disease.

Spliceosome A large structure that functions to remove introns from a heterogeneous RNA and splice the remaining exons together to form a messenger RNA.

Structural variation A region of DNA greater than 1000 bases in size that is inverted, translocated, inserted, or deleted.

Synonymous variant A nucleotide change that results in no change to the predicted amino acid sequence. Although synonymous variants are typically considered to be benign, since there is no protein coding change, there is the possibility of pathogenicity by changes in splicing, gene expression, or mRNA stability.

Telomere A repetitive DNA sequence at the end of a chromosome.

Transcription The biologic process that copies the information in DNA into RNA.

Translation The biologic process that uses ribosomes to convert RNA to protein.

Variant A disease-causing sequence variation. Historically the term has been interchangeable with *mutation* to describe any change in a DNA sequence regardless of its relation to disease causation.

Variant call format (VCF) After aligning all reads onto a reference sequence, variants that are different from the reference genome at a given nucleotide position are stored in a text file in a specific format.

Molecular diagnostics and its parent field, molecular pathology, examine the origins of disease at the molecular level, primarily by studying **nucleic acids**. **Deoxyribonucleic acid (DNA)**, which contains the blueprint for constructing a living organism, is the centerpiece for research and clinical analysis. Molecular pathology is an outgrowth of the enormous amount of successful research in the field of molecular biology that has discovered the basic biologic and chemical processes of how a living cell functions. Molecular biology is now used for clinical diagnosis and the development and use of therapeutics. In this chapter the fundamentals of molecular biology are reviewed, followed by a focus on **genomes** and their variants as well as an introduction to solid tumor genomics.

MOLECULAR BIOLOGY ESSENTIALS

Whether it is a bacterium, virus, or eukaryotic cell, the genetic material located in these organisms dictates their form and function. For the most part the genetic material is DNA, which is composed of two strands of a sugar-phosphate backbone that are bound together by hydrogen bonds between **purines** and **pyrimidines** attached to the sugar molecule, deoxyribose, in a double helix (Figs. 47.1 and 47.2). DNA in human cells is wrapped around histone proteins and packaged into **nucleosome** units, which are compacted further to form **chromosomes** (Fig. 47.3). Each chromosome is a single length of DNA with a stretch of short repeats at the ends called **telomeres** and additional repeats in the **centromere** region. In humans, there are two sets of 23 chromosomes that are a mixture of DNA from the mother's egg and father's sperm. Each egg and sperm is therefore a single or haploid set of 23 chromosomes and the combination of the two creates a diploid set of human DNA, allowing each individual to possess two different sequences, genes, and **alleles** on each chromosome, one from each parent. Each child has a unique combination of alleles because of **recombination** between homologous chromosomes during meiosis in the development of gametes (egg and sperm cells). This creates genetic diversity within the human population. If a child has a random DNA sequence change, the child's **genotype** will be different from that inherited from either of the parents (de novo variant).

DNA is composed of **genes** that code for proteins and **ribonucleic acid (RNA)**. For DNA to convert its store of vital information into functional RNA and protein, the DNA strands must separate so that RNA **polymerase** can bind to the gene. With the help of transcription factors that bind upstream to **promoters**, an RNA polymerase produces single strands of RNA that are further processed to remove the **introns** and retain the protein-encoding **exons**. The mature, processed RNA molecule, the messenger RNA (mRNA), migrates to the cytoplasm, where it is used in the production of protein.

To start the process of protein synthesis or translation, the mRNA is bound by various protein factors and a **ribosome** containing ribosomal RNA (rRNA) and additional protein. The mRNA-bound ribosome begins to produce a polypeptide chain by binding a methionine-bound transfer RNA (tRNA) to the mRNA's initiating AUG **codon** or triplet code. The conversion of the nucleic acid triplet code to a polypeptide is accomplished by additional specific tRNAs, which contain nucleic acid triplet codes (anticodons) and specific amino acids that are transferred to the growing polypeptide chain. After synthesis is complete, the protein migrates to its functional location and eventually is removed and degraded.

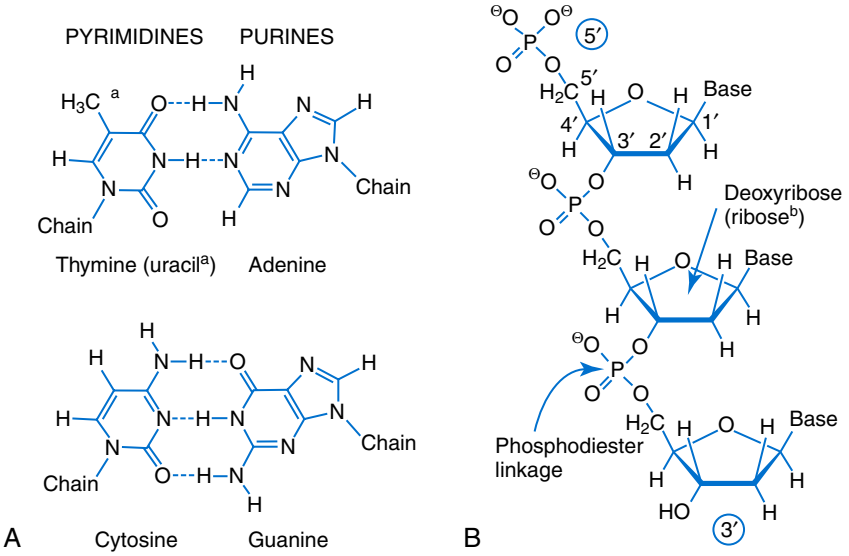


Fig. 47.1 (A) Purine and pyrimidine bases and the formation of complementary base pairs. Dashed lines indicate the formation of hydrogen bonds. ([a]In RNA, thymine is replaced by uracil, which differs from thymine only in its lack of a methyl group.) (B) A single-stranded DNA chain. Repeating nucleotide units are linked by phosphodiester bonds that join the 5′ carbon of one sugar to the 3′ carbon of the next. Each nucleotide monomer consists of a sugar moiety, a phosphate residue, and a base. ([b]In RNA, the sugar is ribose, which adds a 2′-hydroxyl to deoxyribose.)

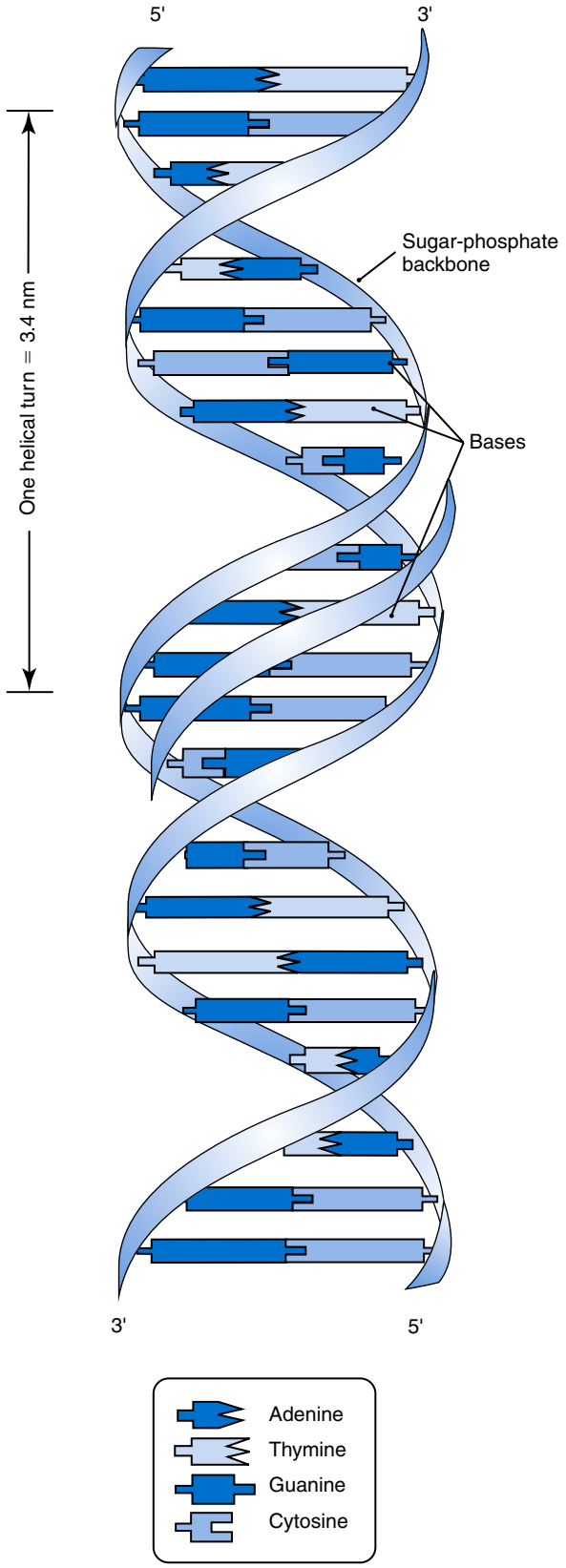


Fig. 47.2 The DNA double helix, with a sugar-phosphate backbone and pairing of the bases in the core-forming planar structures. (From Jorde, L. B., Carey, J. C., & Bamshad, M. J. [Eds.]. [2010]. *Medical genetics* [4th ed.]. Philadelphia: Mosby.)

NUCLEIC ACID STRUCTURE AND FUNCTION

DNA is a rather simple molecule with a limited number of components compared with those of proteins. DNA is composed of a deoxyribose sugar, phosphate linkages, and four nitrogen-containing bases. Deoxyribose contains five carbon atoms that are numbered from 1′ to 5′, starting with the carbon that will be attached to the base in DNA and progressing around the ring until the last carbon, which is attached to a phosphate. The bases consist of the purines, adenine and guanine and the pyrimidines, cytosine, and thymine; an additional base, uracil, replaces thymine in RNA. The basic building block is the **nucleotide**, which consists of a deoxyribose sugar with an attached base at the 1′ carbon and a triphosphate group at the 5′ carbon. The triphosphate nucleotide is the building block for making newly synthesized DNA, which then forms a polynucleotide chain that connects the individual nucleotides through the 5′ and 3′ carbons of each deoxyribose sugar via phosphodiester bonds.

Structure of Deoxyribonucleic Acid

DNA is double-stranded, and the two strands bind to one another through hydrogen bonds between the bases on each strand. Hydrogen bonding is augmented by hydrophobic attraction (stacking) between bases on adjacent rungs of the DNA ladder. Both hydrogen bonds and base stacking are not covalent but are weak bonds that can be broken and reestablished. This important property is exploited by many of the methods that are used in molecular diagnostics. There are two hydrogen bonds between adenine (A) and thymine (T) and three hydrogen bonds between cytosine (C) and guanine (G); because of this difference in the number of hydrogen bonds, separating a guanine-cytosine (G-C) pair takes more energy than an adenine-thymine (A-T) pair (see Fig. 47.1).

Each of the two DNA strands is formed by a phosphate sugar backbone that starts at the 5′ phosphate and ends at a 3′ hydroxyl group, with the complementary bases binding to one another between the two phosphate sugar backbones (see Fig. 47.2). When the two strands are bound to one another, they run in opposite 5′ to 3′ directions called *an antiparallel configuration*. By convention, the DNA sequence is written in a 5′ to 3′ direction. As discussed later, both the **replication** of new DNA and the **transcription** of DNA to RNA progress in the 5′-to-3′ direction. In addition, the conversion of RNA to protein, a process called **translation**, proceeds from the 5′ end of the RNA to the 3′ end. Double-stranded DNA in living cells is generally found as a right-handed, B-DNA helix, which has specific dimensions. Each turn of the helix is 3.4 nm long and consists of 10 bases. The DNA sugar-phosphate backbone is on the outside of the helix with the internal bases bound to their complement on the other strand by hydrogen bonds. Non-B DNA also occurs and can affect transcriptional control and genetic instability; it can also be disease-associated.

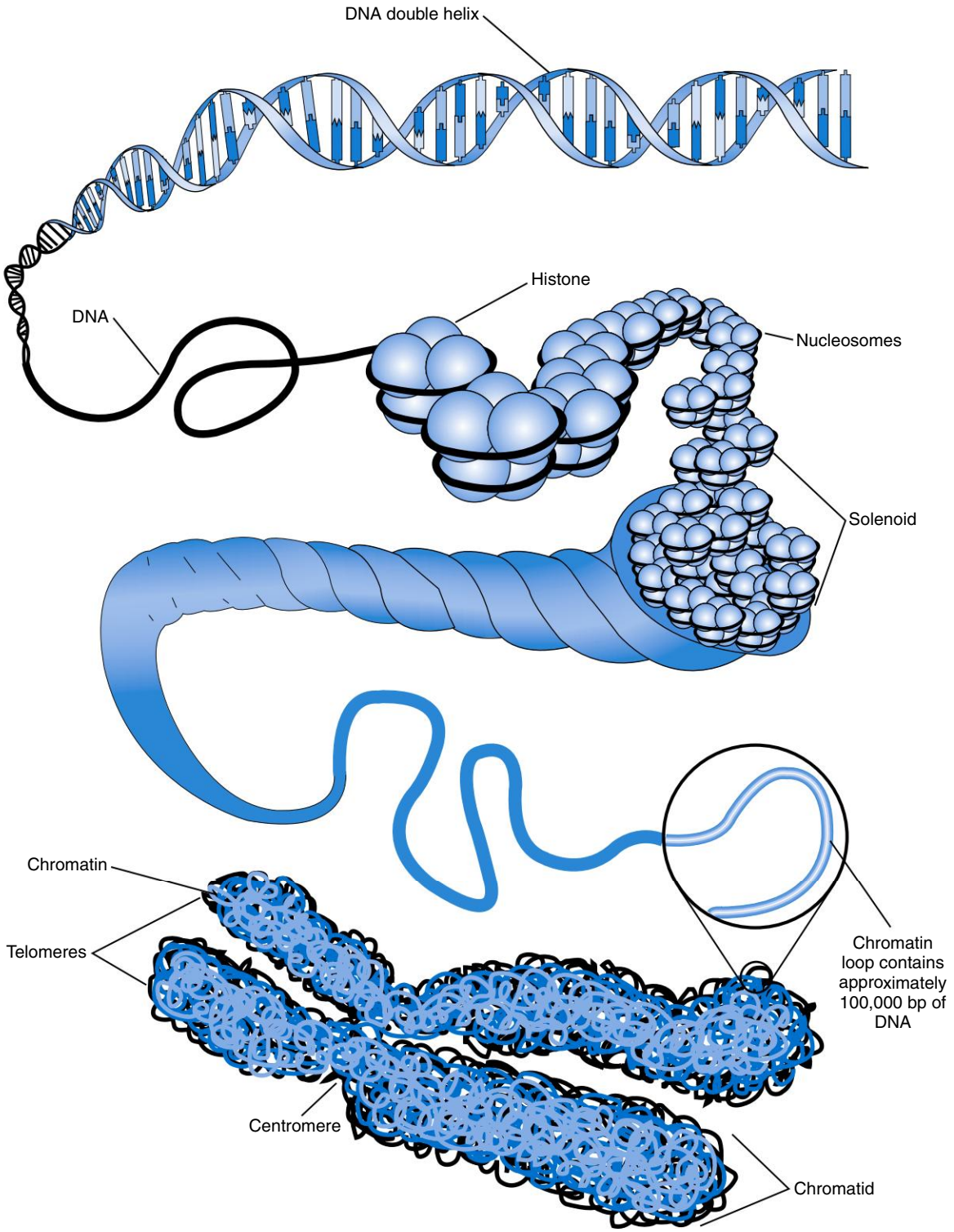


Fig. 47.3 Structural organization of human chromosomal DNA. Double-stranded DNA is wound around the octamer core of histone proteins to form nucleosomes, which are further compacted into a helical structure called a *solenoid*. Nuclear DNA in conjunction with its associated structural proteins is known as *chromatin*. Chromatin is the most compact state of chromosomes. The primary constriction of a chromosome is the centromere, and the chromosome's ends are the telomeres. (From Jorde, L. B., Carey, J. C., & Bamshad, M. J. [Eds.]. [2010]. *Medical genetics* [4th ed.]. Philadelphia: Mosby.)

Composition and Structure of Ribonucleic Acid

The composition of RNA is similar to that of DNA because it contains four nucleotides linked together by a phosphodiester bond, but with several important differences. RNA consists of a ribose sugar with a hydroxyl group at the 2′ carbon instead of the hydrogen atom in DNA. The bases attached to the ribose sugar are adenine, cytosine, and guanine but not thymine, because RNA uses another

pyrimidine—uracil—as a substitute for thymine. Another difference between DNA and RNA is that RNA does not normally exist as two strands bound to one another, although a single strand can bind internally to itself, creating functionally important secondary structures. Although in the past several decades the complexity and number of different RNAs has greatly expanded, the most important cellular RNAs include mRNA, rRNA, and tRNA.

Ribonucleic Acids Associated With Protein Production

mRNA is the most diverse group of the three major types of RNAs, but it constitutes only a small percentage of the total RNA. mRNAs are transcribed from DNA that codes for proteins and therefore are used as the template for the translation of proteins. mRNA is often modified by the addition of a 5′ end cap, which protects the mRNA from degradation, and a polyadenosine (polyA) sequence at the 3′ end. The production and processing of mRNA takes place in the nucleus, followed by transport to the cytoplasm, where it is translated.

rRNA is associated with ribosomes, which are the primary structures that produce protein through the biologic process of translation. rRNA, unlike mRNA, does not code for proteins. In **eukaryotes**, the ribosome is composed of two structures, the 60S and 40S subunits, where the "S" stands for Svedberg units and is determined by the centrifugal sedimentation rate. The ribosome is a mixture of RNA and protein. rRNAs have secondary and tertiary structures that are well conserved with various loops, stem loops, and pseudoknots that contribute to their function. rRNA and protein, as the components of ribosomes, function to carry out the translation of proteins. The structure of the ribosome is now known, and the rRNA is more important than ribosomal proteins in ribosome functioning. The RNA acts as a catalytic agent called a *ribozyme*.

Another important group of RNAs are the tRNAs. These are molecules with a unique cloverleaf structure that translate the **genetic code** from nucleic acid triplets into amino acids that make up proteins. The 3′ ends of tRNAs are covalently attached to specific amino acids. In the middle of the tRNA structure is the anticodon sequence that binds to a specific codon in the mRNA. Therefore the codon directs the binding of a specific tRNA linked to its corresponding amino acid. The genetic code, which consists of 64 three-base codes ($4^3 = 64$), specifies the appropriate amino acid in the growing polypeptide chain.

CENTRAL DOGMA OF MOLECULAR BIOLOGY

The "central dogma" of molecular biology describes the transfer of genetic information into functional macromolecules. Genetic information moves from DNA to RNA via transcription using RNA polymerase and is further translated into protein via ribosomes. Since the initial introduction of the central dogma, a number of other postulated transfers have been described. DNA can enzymatically replicate itself by DNA polymerase, and RNA can be made into DNA using reverse transcriptase.

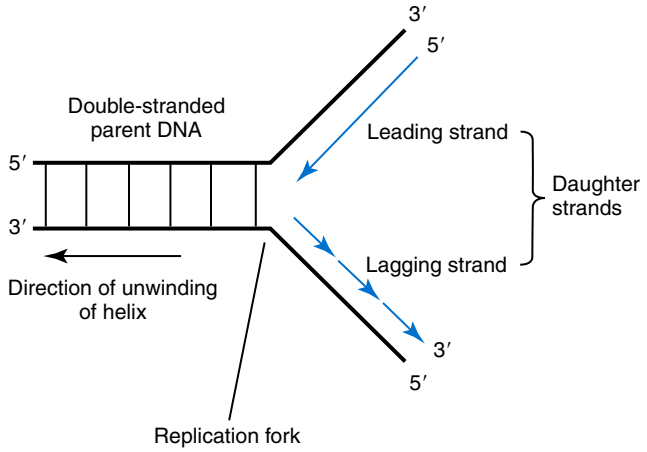


Fig. 47.4 DNA replication. Double-stranded DNA is separated at the replication fork. The leading strand is synthesized continuously, whereas the lagging strand is synthesized discontinuously but is joined later by DNA ligase.

Deoxyribonucleic Acid Replication

Synthesis or replication of new DNA uses one of the two original DNA strands as a template to make a new homologous strand and is termed *semiconservative replication*. Since DNA can be supercoiled into complex structures, a topoisomerase is required to first unfold the **chromatin** and nucleosomes so that the DNA is accessible. A DNA helicase then binds to the double-stranded DNA and separates the two strands, providing two single-stranded DNA templates. Replication progresses in a 5′-to-3′ direction; therefore one strand, the leading strand, is synthesized as one continuous strand. The other strand, called the *lagging strand*, is synthesized discontinuously in small segments. The small fragments on the lagging strand are then linked by a ligase (Fig. 47.4).

DNA polymerases of various types have been identified; they function in many different roles, the most important being the replication of new DNA and the repair of existing DNA. Using the template strand as a guide, the DNA polymerase binds a nucleotide triphosphate to the primer at a free 3′ hydroxyl group, releasing pyrophosphate. The specific nucleotide selected depends on the base on the template strand; for example, an adenine nucleotide is used if a thymine nucleotide is in the template strand. In summary, a complementary sequence is synthesized opposite the template strand. Mistakes occur approximately every 100,000 nucleotides and a major function of some DNA polymerases is error correction or proofreading, accomplished by an intrinsic 3′-to-5′ **exonuclease** activity. DNA polymerases are important in molecular diagnostics because they are used in the polymerase chain reaction (PCR) and DNA sequencing.

DNA replication is part of the cell cycle and defines the synthesis (S) phase. Before the S phase is a first growth phase (G_1). After the S phase is a second growth phase (G_2), followed by a mitosis (M) phase. The M phase, which involves the splitting of one cell into two cells, occurs after the G_2 phase and before the next G_1 phase. Important control points

in the cell cycle are between the G_1 and S phases, just before it begins DNA replication, and between G_2 and M phases, just as the cell commits to creating two cells from one. The G_1/S boundary control point is disrupted in many cancers.

Deoxyribonucleic Acid Repair

DNA can be damaged in a variety of ways that culminate in changes or **mutations** in the DNA sequence. DNA bases may be damaged, removed, cross-linked, or incorrectly paired with one another; single- or double-stranded breaks may also occur. When the cell senses that its DNA has become damaged, it stops the progression of its cell cycle and initiates DNA repair processes. Cells repair these lesions by employing multiple repair mechanisms that are specific for the type of DNA lesion and include base excision repair, nucleotide excision repair, mismatch repair, and homologous recombination repair.

Base excision repair removes bases that are chemically altered or damaged. Deamination of guanine, cytosine, and adenine transforms them into bases that cause incorrect base pairing, creating transition mutations, which are changes between similar nitrogenous bases such as a purine to a purine. A transversion mutation is a change from a purine to a pyrimidine or vice versa. DNA glycosylases, such as uracil-DNA-glycosylase, cleave the damaged base, and a 5′-deoxyribose phosphate lyase removes the nucleotide upstream of the removed base. DNA polymerase and ligase then add a new nucleotide, repairing the damage.

Nucleotide excision repair removes base modifications that change the helical structure of DNA, including bulky DNA distortions and covalently bound structures that may be created by ultraviolet radiation and certain cancer drugs. After the repair is initiated, DNA is unwound and the gap is filled by DNA polymerase and finally sealed by DNA ligase.

Mismatch repair recognizes base incorporation errors and base damage. The mismatched nucleotides must be repaired on the newly synthesized strand of DNA. These mutations are corrected with DNA mismatch repair proteins that excise the surrounding sequence and then repair the excision with new sequence.

Double-stranded breaks may destabilize the genome, resulting in gross chromosomal changes such as translocations, which are frequently found in cancer. Double-stranded breaks are caused by ionizing radiation and chemotherapy drugs and are repaired by either homologous recombination or nonhomologous end joining. The homologous recombination repair pathway is initiated by recognition of a double-stranded break, followed by resection using exonucleases to create a 3′ single-stranded overhang. With the assistance of many proteins, the overhang invades the intact homologous double-stranded DNA of the sister chromatid and uses it as a template for new double-stranded DNA repair.

DNA repair mechanisms operate independently to repair simple lesions. However, the repair of more complex lesions involves multiple DNA processing steps regulated by the DNA damage-response pathway. When single- and double-stranded DNA breaks occur, a cascade of responses is initiated that culminate in either DNA repair, stopping the cell cycle, or programmed cell death.

Deoxyribonucleic Acid Modification Enzymes

There are two groups of **nucleases**, the endonucleases that cut through the sugar-phosphate backbone and exonucleases that digest the ends of DNA. The commercially important restriction endonucleases, which bacteria have acquired to protect themselves from viral infections, are used to cleave DNA at a specific nucleotide sequence or restriction sites. Several thousand restriction endonucleases have been characterized and are used extensively to manipulate DNA in molecular biology and molecular diagnostics. Recent work has described new nucleases, such as the RNA-guided engineered nuclease CRISPR/Cas system, which can precisely cleave genomic DNA.

DNA glycosylases are a family of enzymes associated with base excision repair; they are used in the first step of DNA repair to remove the damaged base without disrupting the sugar-phosphate backbone. An important member of that family, uracil DNA glycosylase, repairs the most common mutation found in humans, the spontaneous deamination of cytosine to uracil, by removing the uracil base.

Gene Structure

The structure of prokaryotic genes is straightforward; almost all of the gene sequence is used to make protein; however, this is not the case with eukaryotic genes. One of the unique hallmarks of eukaryotic genes is that the protein-coding DNA is interspersed with regions that do not code for DNA. A mature mRNA retains only the protein-coding sequences, called *exons*; the sequences between the exons are non–protein-encoding sequences called *introns*, which are removed during mRNA maturation (Fig. 47.5).

Ribonucleic Acid Transcription and Splicing

RNA transcription involves synthesizing an RNA strand using DNA as a template. This requires many different proteins, the most important being a RNA polymerase. Additional proteins function in combination to recognize and regulate the transcription of different genes. The synthesis of RNA proceeds in a 5′-to-3′ direction using DNA as a template, and a specific DNA sequence acts as a transcription start site. Transcription progresses through three phases: initiation, elongation, and termination. The initiation phase includes the binding of transcription factors to promoters upstream from the start site and includes the core promoter immediately upstream and the ancillary promoters further away. Transcription factors binding to upstream promoters act as regulators of the transcription of genes.

Important recurring sequences are found in some promoters. For example, some core promoters contain a TATAAA sequence, called a *TATA box*, located 25 to 40 nucleotides upstream from the transcriptional start site. Transcription factors bind to the TATA box, which in turn promotes the binding of RNA polymerase and additional proteins. A functional transcription complex forms when the double-stranded DNA in the promoter region separates and the transcription complex moves away from the core promoter region. Once started, the

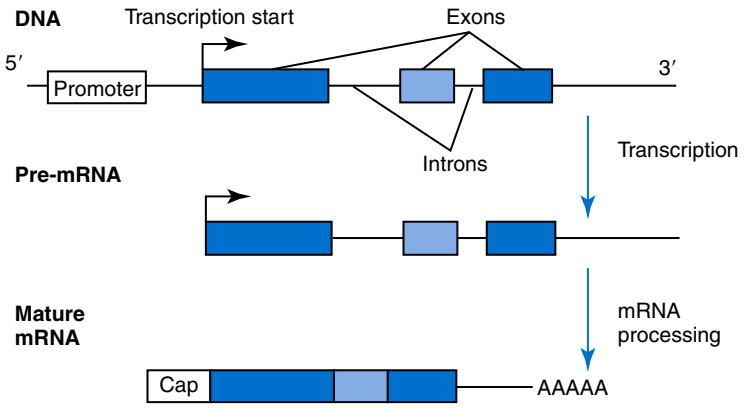


Fig. 47.5 DNA transcription and messenger RNA *(mRNA)* processing. A gene that encodes for a protein contains a promoter region and variable numbers of introns and exons. Transcription commences at the transcription start site. Pre–mRNA is processed by capping, polyadenylation, and intron splicing to become a mature mRNA.

RNA polymerase adds nucleotides to the free 3′-hydroxyl group in a manner similar to that of DNA replication. Transcription is eventually terminated by a polyadenylation step (see Fig. 47.5).

Transcription initially produces RNA that contains both exons and introns. This pre-RNA transcript must be processed or spliced into mature mRNA for it to be properly translated into protein. RNA splicing involves the cleavage and removal of intron RNA segments and splicing of exon RNA segments. Splicing requires the effort of a number of proteins and small RNAs that come together to form a **spliceosome**, which directs the splicing of exons and removal of introns.

An important modification of the splicing process, alternative splicing, allows for the generation of different mRNAs from the same primary RNA transcript by the cutting and joining of the RNA strand at different locations. Among the types of alternative splicing are exon skipping, alternative 3′ and 5′ splice sites, and intron retention. Approximately 92% to 95% of all human genes are alternatively spliced, resulting in additional protein complexity that may be tightly regulated and specific to different cells and tissues.

Translation

The final phase of the transfer of information from DNA is to proteins, the structural and functional molecules that make up most of a living organism. Proteins are long single strands of various amino acids and are synthesized by a process called *translation*, which requires the functioning of many protein factors, tRNAs, and ribosomes.

Amino acids have a common structure consisting of a carbon atom bound to amino and carboxylic acid groups with a side chain. There are 20 amino acids, each with a different side chain that gives them their unique properties. The side chains can be divided into four types: nonpolar (hydrophobic), polar (hydrophilic uncharged), negatively charged, and positively charged. A protein's amino acid makeup and sequence in the polypeptide chain determine the overall structure and function of the protein. Some amino acids commonly induce structural changes, such as proline, which disrupts secondary structure, and cysteine, which can cross-link to another cysteine through disulfide bonds. For additional information on protein structure, see Chapter 18.

The genetic code, which was deciphered in the early 1960s, is required to convert a nucleic acid sequence into an amino acid sequence. Humans have four unique nucleotides, which are arranged into sets of three per codon. Theoretically there are 64 combinations; however, rather than encoding 64 amino acids, humans have 20. A hallmark of this genetic code is that it is redundant, meaning that there are several codes for one amino acid. That is the case for most amino acids but not all; for example, methionine and tryptophan have only one code. The redundancy is usually in the third base of the code. All of the 64 three-nucleotide codon possibilities code for an amino acid except for the three that serve as stop codons (UAA, UGA, and UAG) (Fig. 47.6).

Protein synthesis or translation occurs in the cytoplasm and proceeds in three steps: initiation, elongation, and termination. The process requires tRNA and rRNA molecules as well as ribosomes and initiation, elongation, and termination factors. The first codon (AUG) always codes for methionine; therefore, to initiate translation, a methionine tRNA first binds to the ribosome. The tRNA specific for the next three-base codon—for example, lysine—binds to the ribosome, and a peptide bond is created between the amino group of one amino acid and the carboxyl group of the next amino acid. The ribosome moves forward one codon and the next tRNA specific for the next codon binds in the ribosome; this process is repeated until a termination codon is reached (Fig. 47.7). Protein synthesis occurs in the endoplasmic reticulum, where multiple ribosomes called *polyribosomes* are involved in translating an individual mRNA.

POINTS TO REMEMBER

- The two strands of DNA are bound together by hydrogen bonds and stacking forces that can be broken and reformed without permanent damage to the DNA.
- The conversion of DNA information into protein is facilitated by aminoacyl tRNA synthetases and their ability to create amino acid–specific tRNAs.
- The genetic code is redundant; the three-base code can have 64 different combinations, but only 20 amino acids are recognized.

First Letter	Second Letter: U	Second Letter: C	Second Letter: A	Second Letter: G	Third Letter
U	UUU, UUC Phenylalanine UUA, UUG Leucine	UCU, UCC, UCA, UCG Serine	UAU, UAC Tyrosine UAA Stop Codon UAG Stop Codon	UGU, UGC Cysteine UGA Stop Codon UGG Tryptophan	U C A G
C	CUU, CUC, CUA, CUG Leucine	CCU, CCC, CCA, CCG Proline	CAU, CAC Histidine CAA, CAG Glutamine	CGU, CGC, CGA, CGG Arginine	U C A G
A	AUU, AUC, AUA Isoleucine AUG Methionine	ACU, ACC, ACA, ACG Threonine	AAU, AAC Asparagine AAA, AAG Lysine	AGU, AGC Serine AGA, AGG Arginine	U C A G
G	GUU, GUC, GUA, GUG Valine	GCU, GCC, GCA, GCG Alanine	GAU, GAC Asparatic Acid GAA, GAG Glutamic Acid	GGU, GGC, GGA, GGG Glycine	U C A G

Fig. 47.6 Genetic code. Translation of messenger RNA to amino acids during protein synthesis.

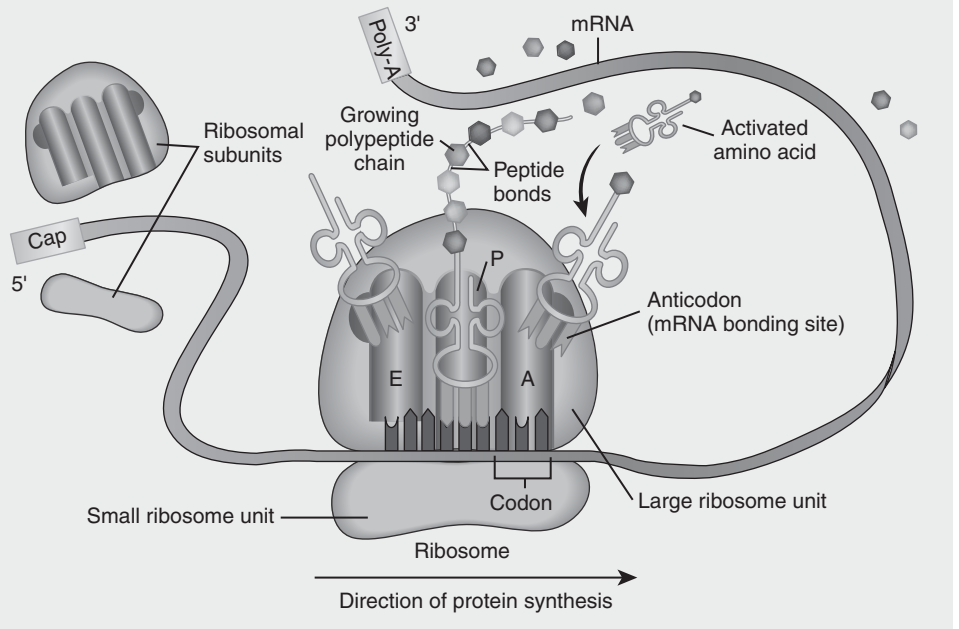


Fig. 47.7 Translation. Shown is a ribosome bound to a messenger RNA *(mRNA)* converting the mRNA triplet code (codon) via a specific amino acid–bound transfer RNA containing a complementary anticodon sequence. There are three transfer RNA positions. A new amino acid–bound transfer RNA first arrives on the ribosome at the *A* or acceptor site, at the front of the moving ribosome, and then moves to the *P* or peptidyl site, where the amino acid on the newly arrived transfer RNA combines with the growing polypeptide chain. Finally the now empty transfer RNA moves to the *E,* or exit site, where it prepares to leave the ribosome. (Modified from Huether, S. E., & McCance, K. L. [2017]. *Understanding pathophysiology* [6th ed.]. St. Louis: Elsevier.)

EPIGENETICS

Global and mRNA-specific regulation of translation often occurs by **epigenetic** changes—that is, changes that do not modify the DNA sequence. Currently there are three major types of epigenetic modifications: (1) DNA methylation, (2) histone modifications, and (3) noncoding RNAs.

DNA methylation is a well-known epigenetic change. Methylation of cytosine to 5-methylcytosine occurs frequently; about 70% of CpG dinucleotides in the human genome are methylated. CpG islands are about 1000 bases in length and are often found near the 5′ end of genes. These regions consist of clusters of CG dinucleotides that are usually not methylated in normal cells. However, CpG methylation

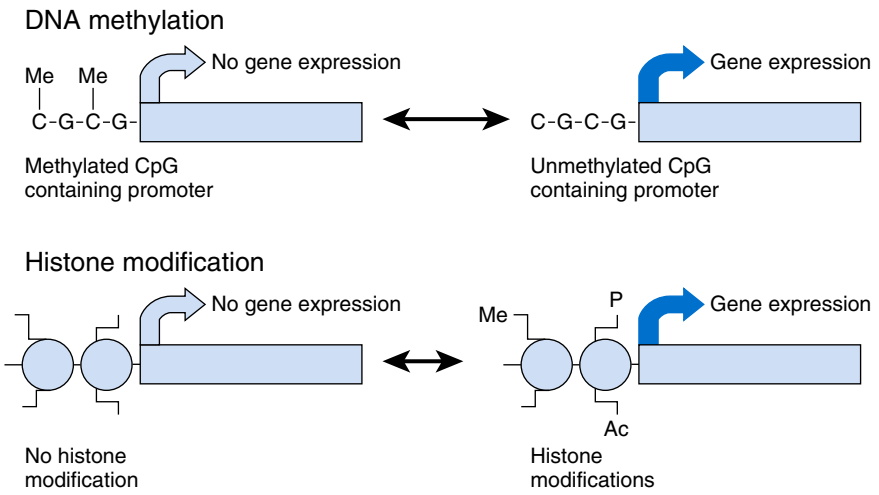


Fig. 47.8 Epigenetics. *Top,* DNA methylation of CpG island regions indicated by *Me* in and around gene promoters is associated with loss of gene expression and silencing of the gene. When *CpG* islands are unmethylated, shown by absence of Me, gene expression is unaffected. *Bottom,* Modifications of the tails of histone proteins, such as methylation, acetylation, and phosphorylation, shown as *Me, Ac,* and *P,* respectively, can increase gene expression. (Modified from Zaidi, S. K., Young, D. W., Montecino, M., van Wijnen, A. J., Stein, J. L., Lian, J. B., et al. [2011]. Bookmarking the genome: maintenance of epigenetic information. *Journal of Biological Chemistry,* 286, 18355–18361.)

correlates with condensed chromatin structure and promoter inactivation (Fig. 47.8). Cancer is the most common human disease associated with aberrant DNA methylation.

DNA replication, repair, and transcription require free access to DNA. However, DNA is usually wound around **histones** to form nucleosomes and is further condensed into chromatin. Therefore the conformation of chromatin is tightly regulated and is very dynamic; at any one point in time portions of the DNA are being exposed and other portions are being covered. For example, reversible phosphorylation at serine, threonine, and tyrosine residues adds negative charges to histones that can repel them from the negatively charged DNA. Similarly, specific histones may be reversibly modified by acetylation to remove the positive charge on lysine residues and thus to lower histone attraction to DNA. Histone, lysine, and arginine residues can also be mono-, di-, and trimethylated. Although the charges are unchanged, methylation further serves to promote DNA transcription, replication, and repair.

Noncoding RNAs also function in the regulation of translation. In addition to noncoding ribosomal and tRNAs, there are many forms of noncoding RNA, including long (>200 nucleotides) and short (20 to 200 nucleotides) noncoding RNAs. Long noncoding RNAs are diverse and highly regulated during the development of an organism. At least 93% of the genome is transcribed, producing more than 10 times the amount of RNA than is produced from the coding segments of genes. Both strands of DNA may be transcribed, and long noncoding transcripts may overlap coding regions, producing a complex transcriptome of functional RNA molecules that may variably regulate transcription, RNA processing, mRNA stability, translation, protein stability, and secretion.

Small noncoding RNAs include **microRNAs (miRNAs)** that are particularly interesting as potential markers for disease. miRNAs are functional single-stranded RNAs 21 to 22 bases long that are expressed in a tissue-specific manner. There are thousands of different miRNAs that inhibit mRNA expression by either mRNA degradation or by blocking of translation.

Human Chromosomes

Cellular DNA is bound by many proteins to form chromatin (see Fig. 47.3). The proteins in chromatin consist of histones, which are bound in precise amounts per length of DNA, and other proteins called *nonhistone proteins* that are bound more irregularly. The histone proteins consist of two copies of four proteins that bind as a unit to 147 **base pairs** (bp) of DNA to make up a nucleosome. One additional protein binds between the nucleosomes (Fig. 47.9). Access to DNA by transcription factors is controlled by proteins that remodel the histone proteins through phosphorylation, acetylation, and methylation. The nucleosomes are condensed into filaments and even more compact structures to form a chromosome (see Fig. 47.3). The DNA in chromosomes is continuous for each chromosome and can be as much as several hundred million base pairs in length.

Chromosome regions can be classified by their transcriptional activity. The ends of chromosomes, called *telomeres*, contain the repeat sequence TTAGGG, which shortens with age. The centromeres, at the centers of most chromosomes, are important during mitosis and contain mostly repeated segments of DNA.

THE HUMAN REFERENCE GENOME

One of the defining achievements of the early 21st century is the sequencing and **assembly** of more than 90% of the human genome. Of course there is not a single human genome; individuals differ from each other by about 0.1%, based on differences in **single nucleotide polymorphisms (SNPs)**.

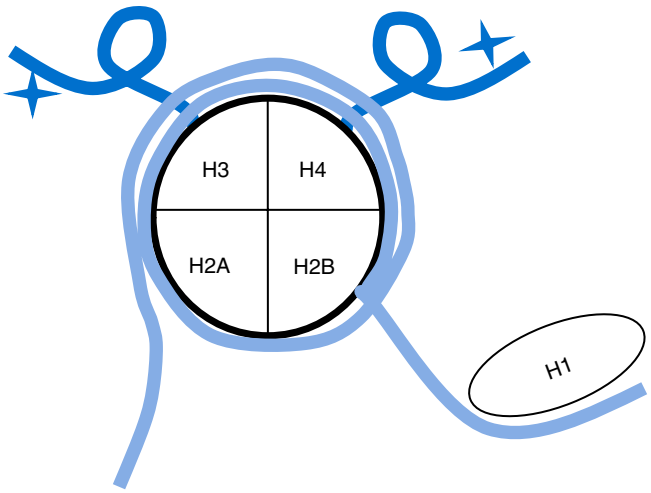


Fig. 47.9 Schematic illustration of a nucleosome unit. A segment of DNA is wound around a nucleosome core particle consisting of an octamer containing two copies of histone proteins H2A, H2B, H3, and H4. Tails with modifications (indicated by a *blue star*) are shown to protrude from H3 and H4. Adjacent nucleosomes are separated by a segment of linker DNA and the linker histone H1.

Variation comes in many different forms, including single base changes and copy number changes in large segments of DNA. Even more challenging than sequencing the whole genome is documenting and understanding the clinical significance of human sequence variation.

Throughout the 1990s, there was an international effort to sequence the human genome. The first draft was released in 2001, followed by a more complete version in 2004. The 2004 version contains 2.85 billion nucleotides (bases) and was considered 99% complete for actively transcribed DNA. The overall size of the genome is estimated to be 3.08 billion nucleotides including highly repeated regions such as telomeres and centromeres. Within the genome are 19,438 known genes and an additional 2188 predicted genes. The portion of the genome encoding proteins, known as the *exome*, comprises only 1.2% of the genome. The 2004 genome has been used as the reference sequence for most subsequent sequencing projects, although periodic release of improved versions has occurred since its initial release. A reference is required because commonly used next-generation DNA sequencing technologies produce relatively short read lengths and hence require a scaffold onto which these short sequences are aligned prior to variant identification.

Statistics for the human genome and its variations are summarized in Table 47.1. Three-quarters of human DNA is **intergenic**, or between genes. More than 60% of this intergenic sequence consists of "parasitic" DNA regions of mostly defective transposable elements 100 to 11,000 bases in length. Between 2 and 3 million of these "retrotransposons" are present in each copy of the genome. They contribute to genetic recombination and chromosome structure and provide an evolutionary record of sequence variation and selection.

Segmental duplications constitute 5.3% of the human genome. They are more than 1 kilobase (1000 bases or 1 kb) in length, have a sequence identity of at least 90%, and are

TABLE 47.1 The Human Genome and Its Sequence Variation

The Human Genome	
3.08 billion base pairs in 24 chromosomes (counting X and Y separately)	
23 chromosome pairs per cell (46–244 million base pairs per chromosome)	
75% Intergenic Sequences	
Transposable elements	45%
Segmental duplications	5%
Simple sequence repeat	3%
Structural (centromeres, telomeres)	2%
Other	20%
25% Genes That Code for Proteins	
Introns	23%
Exons	1.9%
• Coding segments	1.2%
• Untranslated regions	0.7%
Number of genes	19,438 known 2,188 predicted
Average gene	27,000 base pairs 10.4 exons 9.1 transcribed exons 1340 exonic bases 446 amino acids
Sequence Variants	
99.9% identity (one difference every 1250 bases between randomly selected haploid genomes)	
Single Nucleotide Variants (SNVs): Identified Every 75 Bases on Average	
Noncoding	97%
Average number within a gene	126
Average number within the coding region of a gene	5
Copy Number Variants (CNVs): Involves 5%–12% of the Genome	
Disease-Causing Variants	
SNVs	68%
• Missense (amino acid substitution)	45%
• Nonsense (termination)	11%
• Splicing	10%
• Regulatory	2%
Small insertions or deletions (or both)	24%
Structural variants (copy number variations, inversions, translocations, rearrangements, repeats)	8%
Epigenetic Alterations	
Variable initiation and alternative splicing	
Cytosine methylation	
Histone phosphorylation, methylation, acetylation	

not transposable elements. Segmental duplications are common in the human genome and are prone to **deletion** or rearrangement (or both), often with phenotypic consequences. Intergenic DNA also carries most of the **simple sequence repeats (SSRs)** present in the genome. A subset of SSRs, the **short tandem repeats (STRs)**, have repeat units of one to several bases that may tandemly repeat up to thousands of times. STRs have played a large role in genetic linkage studies and in forensic and medical identity testing. They are formed by slippage during replication and are highly polymorphic among individuals. The most common STRs are dinucleotide repeats, such as ACACAC and ATAT. On average, one STR occurs every 2000 bases.

Approximately 2% of the genome is required to maintain the structure of chromosomes and is located at chromosome centers (centromeres) and ends (telomeres). Centromeric DNA includes many tandem copies of nearly identical 171-bp repeats encompassing 0.24 to 5.0 megabytes (MB) per chromosome. Each chromosome end is capped with several kb of the telomeric 6-base repeat TTAGGG. Although intergenic DNA does not code for protein and was originally considered "junk," much of this DNA is transcribed to RNA, producing a complex "transcriptome" network of RNA control elements whose function and mechanics are active areas of investigation.

One-quarter of the human genome consists of genes. The average gene covers 27,000 bases, but only about 1300 of these bases code for amino acids. The primary RNA transcript is processed by splicing to retain exons that are interspersed throughout the gene and have a higher GC content than noncoding regions. On average, 95% of a gene is excised as introns, retaining a mean of 10.4 exons, of which on average 9.1 are translated into proteins. Exons make up only 1.9% of the total genome, with 1.2% of the genome coding for proteins. Some important genes are present in many copies, so that overall protein expression is not affected if a chance variation occurs in one copy. If extra copies of genes lose their function, they are known as **pseudogenes.** At least as many pseudogenes as functional genes are present in the human genome. It is important to distinguish pseudogenes from functional genes because variants in pseudogenes are seldom of clinical importance and often complicate DNA diagnostic assays.

PROJECTS INSPIRED BY THE HUMAN GENOME PROJECT

Genome-wide association studies (GWAS) use microarray tests that type large numbers of **single nucleotide variations (SNVs)** to find associations between genetic variations and diseases. DNA variants are often clustered into regions that are inherited as a unit, such that a unique SNVs pattern or haplotype can be passed from generation to generation. The International HapMap Project used SNVs to investigate haplotype associations with disease.

The 1000 Genomes Project sequenced a large number of diverse human samples from around the world. The goal was to build a comprehensive catalog of the most common human genetic variants, which included SNVs as well as **insertions**, deletions, copy number and structural variants that are found in the populations studied. The Exome Aggregation Consortium (ExAC) has amalgamated sequencing data for over 60,000 exomes to delineate common genetic variation within human exomes. The SNP database, International HapMap Project, 1000 Genomes Project, ExAC, and GWAS have helped to define genetic variability within individuals and populations to understand the basis of many genetic diseases.

A more fundamental biology project is the encyclopedia of DNA elements, or ENCODE, whose goal is to catalog the functional elements of the genomes of humans and other species. The functional elements include promoter and enhancer regions, all expressed RNA forms, and epigenetic modifications.

NONHUMAN GENOMES

The genomes of different organisms vary greatly in size and complexity. There is overlap in genome size among eukaryotes (animals, plants, fungi), viruses, and bacteria (Fig. 47.10). Human disease is caused by fungi, bacteria, and viruses with smaller and less complex genomes.

Bacterial genomes have only one chromosome, usually a circular DNA double helix of 2 to 5 million base pairs, or ~1000 times less than the amount of DNA in a human cell. About 90% of the DNA in bacteria codes for protein. There are no introns, but there are multiple small intergenic regions of repetitive sequences that are dispersed throughout the genomes. *Escherichia coli*, a common bacterium in the human intestinal tract, has about 4300 genes. In addition to the large circular chromosome that carries essential genes, bacteria also carry accessory genes in smaller circles of double-stranded DNA (dsDNA) known as **plasmids**. These range in size from 1000 to more than 1 million base pairs. Plasmids are important in the molecular diagnosis of bacterial infections because they often encode pathogenic factors and antibiotic resistance.

Viral genomes are considerably smaller and less complex than bacterial genomes. Common viruses that infect humans vary in size from about 5000 to 250,000 bases, or 20 to 1000 times less than the *E. coli* genome. Because viruses use the host's cellular machinery, they do not need as many genes as bacteria. Small viruses may encode only several genes, but the larger viruses can encode hundreds. The viral genome consists of either DNA or RNA, and the nucleic acid may be single- or double-stranded, linear or circular, with one or multiple fragments or copies per viral particle. As in bacteria, viral genes have no introns. In fact, in some viruses the genes overlap, with different reading frames coding different products from the same nucleic acid sequence. Noncoding regions are usually present at the terminal ends of linear viral genomes. Repeat segments are often found as terminal or internal repeats and may be inverted.

Sequence alterations in viruses are common. Areas of high sequence variation may be interspersed between

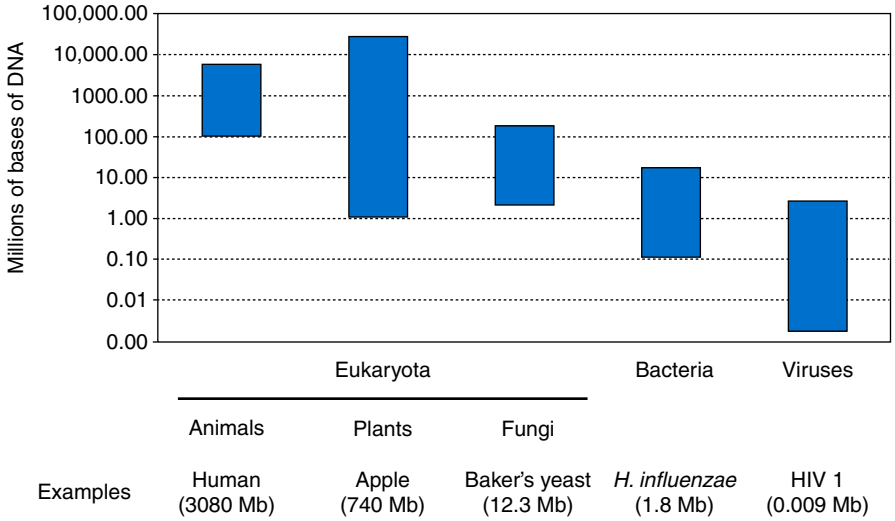


Fig. 47.10 Range of genome sizes. There is a wide variation in genome size among different organisms. In this plot of publicly available genomes, the *y*-axis is in megabases and the *x*-axis lists various organisms: Eukaryota (animals, plants, fungi), bacteria, and viruses. On average, Eukaryota have larger genomes compared with bacteria and viruses; however, there are exceptions, in which viral genomes are larger than those of bacteria or Eukaryota. The difference between the smallest and largest known genomes is more than six orders of magnitude. Several specific genome sizes are illustrated in Mb (megabases, million). (Data extracted from the National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/genome.)

conserved domains. Higher frequencies of variation correlate with lower polymerase fidelity and may allow escape from antibody recognition and antiviral drugs. Common sequence variants in viruses include single base changes, insertions, and deletions. Sequence diversity within a viral species may be so great that consensus sequences for molecular assays are difficult to find.

VARIATION IN THE HUMAN GENOME

About 99.9% of human genome sequence is identical between randomly chosen individuals in a population in considering the simplest SNPs. However, **copy number variants (CNVs)** greater than 50 kb involve up to 0.5% of the genome. That is, between individuals, at least five times as many bases are affected by copy number changes as by small sequence differences.

Most human genetic material is present in two copies with the exception of the unpaired sex chromosome in males and **mitochondrial DNA**. The presence of only single gene copies on the X and Y chromosome in males leads to well-known sex-linked disorders. In contrast, the 16,500-bp mitochondrial genome is present in multiple copies per cell, constituting about 0.3% of human DNA depending on the tissue source. Allele fractions may vary over a wide range when all mitochondria in a cell are considered. That is, sequence variations in mitochondrial DNA are **heteroplasmic**, meaning that the ratio of the wild-type allele to a variant allele can vary almost continuously, sometimes resulting in a wide range of symptoms even when only one sequence variant is involved.

Any sequence change from a reference sequence is called a *sequence variant* or *variation*. Many variations do not affect human health and are benign or silent. For example, most (1) CNVs, (2) SNVs, and (3) STRs found between genes are seldom associated with disease. Sequence alterations that are known to cause disease are often called *mutations, pathogenic variants,* or *disease-causing variants.* About 68% of known disease-causing variants involve only a single base change. Most of the remaining disease-causing variants (24%) are small insertions or deletions. The remainder (8%) includes more complex **structural variations** (see Table 47.1).

Single Nucleotide Variants

The most common sequence variations are single base variants, also known as SNVs. More than 40 million SNVs have been described, and many new SNVs continue to be reported. Some SNVs are common in the population, with allele frequencies of 0.1 to 0.5 (i.e., present in 10 to 50 of every 100 copies studied), but other single base changes are very rare. The vast majority of SNVs (97%) occur in noncoding regions; only 3% of SNVs are associated with exons. Similarly, most of the SNVs within introns except for splicing and regulatory variants are not known to affect gene function. In addition, some of the SNVs within exons are silent alterations that do not code for a change in amino acid sequence because of the redundancy in the genetic code. Still other SNVs in exons code for amino acid changes that do not affect protein function.

Most SNVs that cause disease are missense and result in an amino acid substitution; significantly fewer are nonsense variants that result in a termination codon and premature polypeptide chain termination. Approximately 10% of disease-causing variants are SNVs that affect splicing sites and result in altered concatenations of coding sequences or intron retention. Finally, fewer than 2% of known disease-causing variants are SNVs that affect the regulatory efficiency of transcription by altering promoter and/or enhancer regions in introns or the stability of the RNA transcript.

Small insertion and/or deletion variants account for 24% of variants that cause disease. An insertion refers to the presence of extra bases, and deletion implies the absence of certain bases in comparison with a reference sequence. Insertions and deletions **(indels)** often cause a shift of the codon reading frame, resulting in altered amino acid sequence downstream of the variation—commonly followed by chain termination from a nonsense codon.

The remaining 8% of variants that relate to health and disease are mostly structural variants, including (1) duplications or deletions of entire exons or genes; (2) gene fusions, including chromosomal translocations and inversions; (3) STR expansions (e.g., an increased number of trinucleotide repeats); (4) gene rearrangements (e.g., rearrangements of immunoglobulin genes in B cells); (5) complex polymorphic loci related to health and disease (e.g., human leukocyte antigens); and (6) CNVs.

Copy Number Variation (Gains and Losses)

Although SNVs are the most common sequence variant, CNVs cover more of the genome than SNVs. An average of 12.4 large CNVs are present per individual. Some CNVs affect up to 2 million bases of DNA in phenotypically normal individuals. CNVs may be duplicated in tandem or may involve complex gains or losses of homologous sequences at multiple sites in the genome. CNVs can involve single genes or contiguous sets of genes. When the normal dosage of the gene is two but more than two functional copies of a gene are present, then the gene is "amplified." If a dosage-sensitive gene is amplified, it usually leads to overexpression of mRNA and protein, resulting in cellular abnormalities and possible progression to diseases such as cancer. When the normal gene dosage is two and loss of one of the functional copies of the gene occurs, disorders such as mental retardation and developmental delay may result. Loss of both copies is also deleterious and may lead to cancer development or result in embryonic lethals.

Fusions

Gene **fusions** from deletions, duplications, inversions, and translocations are commonly found in cancer. Often they arise by balanced translocations, whereby a chimeric protein is created by the fusion of two coding regions. Gene fusions promote tumor proliferation by either activating an oncogenic driver or inactivating a tumor suppressor. By identifying gene fusions that act as primary oncogenic drivers, targeted therapies may become available for precision treatment.

NAMING OF CHROMOSOMES, GENES, AND VARIANTS

Autosomes are numbered from 1 to 22 in descending order of length. The sex chromosomes are named X and Y. The symbols p and q designate the short and long arms of the chromosomes, respectively. A chromosome band is the part of the chromosome that is clearly distinguishable from adjacent segments that are darker or lighter in appearance. G bands are the bands resulting from Giemsa dye staining. Additional rules for naming chromosomal rearrangements such as inversions, deletions, and translocations are also standardized.

Genes are named descriptively, along with a gene symbol which is a short abbreviation in upper case Latin letters and Arabic numbers that are both italicized (e.g., collagen type II, alpha 1 chain, *CT2A1*). Sequence variants are described at the DNA level without indication of disease potential. Preferred terms include sequence variant, CNV, and SNV. Basic naming recommendations for SNPs are introduced in Box 47.1.

BOX 47.1 Nomenclature for Naming Small Sequence Variants

The position of single nucleotide substitutions can be referenced to the genome (g.), amino acids in the protein (p.), or nucleotide bases in the cDNA transcript (c.). In all cases, the genome build (e.g., Genome Reference Consortium human genome build 38 or "GRCh38") should be specified.

For genomic coordinates, specify the chromosome and the base position followed by the base change. For example: GRCh38 Ch17: g.3424566C > T. Genomic coordinates are the only option for most intergenic variants, as well as deep intronic variants.

When the variant is within the protein coding region of a gene and protein coordinates (p.) are preferred, give the gene name followed by the number of the amino acid affected starting with the initiating methionine as 1. The wild-type amino acid is listed to the left of the number, and the variant amino acid is listed to the right. For example, *CYBB* p.T42R (or *CYBB* p.Tyr42Arg) are listed for the one- and three-letter codes of amino acids, respectively. The protein coordinates best specify the phenotype, but it is a challenge to convey the results of frameshifts and the exact nucleotide change is often ambiguous.

When the variant is in or near the exons of a gene, cDNA coordinates (c.) can be used. List the gene name followed by the base number using the A in the ATG initiation codon as 1 and then the base change. For example: *CYBB* c.125C > G. If the base change is in an intron (e.g., a splice site mutation), the number is counted from the nearest intron-exon boundary. For example, *CYBB* c.141 + 2T > C or c.142 − 12C > T. If the variant is 5′ of the ATG initiation sequence, use negative numbers (*CYBB* c.−64C > T), or if it is 3′ of the last exon, use an asterisk (*CYBB* c.*67G > A). With c. coordinates, you know the exact base change even though it may not affect the amino acid sequence.

VARIANT DATABASES

Some commonly used variant databases are detailed in Box 47.2; they include dbSNP (Database of Single Nucleotide Polymorphisms), OMIM (Online Mendelian Inheritance in Man), HGMD (Human Gene Mutation Database), and ClinVar (Clinical Variant). dbSNP is a systematic catalog of SNVs including small insertions and deletions is dbSNP. The reference identifier for variants in dbSNP begins with the prefix "rs." OMIM is a catalog of genes and variants related to disease. As of April 2017, some 24,000 diseases or syndromes were described in OMIM, along with 15,600 genes.

BOX 47.2 Human Genomic Databases

Comprehensive
NCBI (National Center for Biotechnology): ncbi.nlm.nih.gov-/genome
Ensembl: ensembl.org/index.html
UCSC (University of California Santa Cruz): genome.ucsc.edu/

Genes and Disease
OMIM (Online Mendelian Inheritance in Man): ncbi.nlm.nih.gov/omim
COSMIC: http://cancer.sanger.ac.uk/cosmic
ClinVar: ncbi.nlm.nih.gov/clinvar/
Decipher: decipher.sanger.ac.uk/index

Sequence Databanks
NCBI GenBank: ncbi.nlm.nih.gov/Genbank/
EMBL (European Molecular Biology Laboratory)-Bank: www.ebi.ac.uk/embl
DDBJ (DNA Data Bank of Japan): ddbj.nig.ac.jp/

MicroRNAs
miRBase: mirbase.org

General Variation Databases
Leiden Open Variation Database: lovd.nl/3.0/home
Human Genome Mutation Database: www.hgmd.cf.ac.uk/ac/index.php
Short Genetic Variations (dbSNP): ncbi.nlm.nih.gov/projects/SNP/
1000 Genomes: www.1000genomes.org/
Exomes (ExAC): exac.broadinstitute.org/

Specialized Variant Databases
Database of Genomic Structural Variation (dbVar): ncbi.nlm.nih.gov/dbvar
Database of Genomic Variants (DGV): dgv.tcag.ca/dgv/app/home
Retrotransposons: dbrip.brocku.ca/
Haplotypes (HapMap): hapmap.org/

Nomenclature
HUGO (Human Genome Organization) Gene names: genenames.org
HGVS (Human Genome Variation Society) Sequence variants: www.hgvs.org/mutnomen

In contrast to OMIM, HGMD seeks to be a comprehensive database of all variants with reported disease association. As of 2017, there were 203,885 variants from 8271 genes cataloged in HGMD. The types of variants included 114,703 missense or nonsense variants, 18,386 splicing variants, 30,169 small deletions, 12,557 small insertions, and 15,272 gross deletions. The chief limitation of databases such as OMIM and HGMD is that they rely on published reports. As new variants of known genes are discovered in research or clinical laboratories, they are rarely published. This has led to the ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), which includes annotated variants by clinical laboratories, research laboratories, and the literature within a publicly available database.

BIOINFORMATICS

Although a single human genome is approximately 3 billion bases, the amount of massively parallel sequencing (MPS) data needed to accurately determine an individual's genotype may exceed 90 billion bases. This is true because short reads require a specific depth of read coverage across the genome in order to confidently genotype each variant and variant type. The data storage requirements are typically more than 0.5 terabytes per human genome. Fortunately, as the scale of sequencing has increased, tools have been developed to manage and analyze the information derived from these analyses. Often both publicly available tools and commercial software are assembled into what is referred to as a pipeline. In the pipeline, the aligned sequence data are processed in a serial manner (Fig. 47.11). Raw sequencing data from millions of short sequencing reads are aligned to the reference genome. Depending on the coverage required, there may be tens to thousands of sequencing reads at a single nucleotide position. After alignment, the bases are "called" or determined by a software algorithm. The base call is dependent on a variety of factors, including the quality of the reads (Q score; Box 47.3) and the percentage of reads at the nucleotide position that are in agreement. Postcall filtering removes sources of known false positivity such as variants called only on one strand or only at the ends of reads, where data quality is poorer. After the bases are called for each nucleotide position, a variant call format file (VCF) is generated. The VCF includes the nucleotide positions of the variants that differ from the reference genome and are tabulated by the genomic coordinate of the variant. The VCF is then examined by another set of software algorithms that query multiple databases; the examination by databases is sometimes referred to as filtering or interpretation. A standard filter is to further examine only variants that are predicted to change protein coding (e.g., missense or nonsense variants). Other filters may be based on the frequency of specific variants in various populations. If the variant occurs in a gene implicated in a rare disease that occurs in fewer than 1 in 100,000 individuals, that variant should occur much less frequently than 1% in any population. In addition to population databases, there are databases that catalog variants with clinical **phenotype** information or based on reports in the published literature. Finally, rather than an examination of databases, tools that predict the importance of variants to protein structure or function may be used. These predictive tools may be helpful to prioritize the examination of a long queue of variants from a sequencing study.

CANCER GENOMICS

An important clinical application of genomics is cancer diagnostics. The first whole genome sequencing (WGS) of a cancer was an acute myeloid leukemia (AML) in 2008; many others have subsequently followed. The combination of phenotypic information and molecular profiling provides an opportunity for new therapeutic targeting of specific cancer-associated genes.

Modern-day cancer diagnostics are based on the cumulative knowledge gained from large-scale cancer genomics

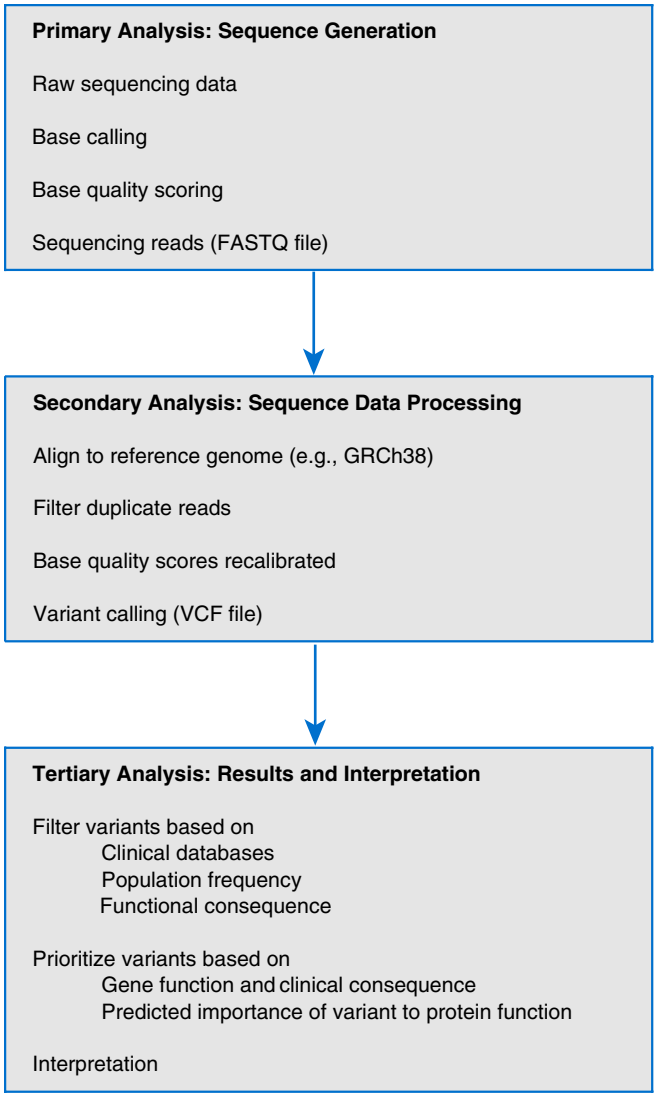


Fig. 47.11 Bioinformatics pipeline. The analysis of data from massively parallel sequencing can be broadly considered to occur in three phases. *Primary analysis:* The raw output (e.g., optical or electronic signals) from the sequencing instrument is transformed into data that describe the individual bases of DNA as well as the quality and confidence of the base call at each position. These reads of DNA are assembled into a FASTQ data file. *Secondary analysis:* The data file is then assembled onto a reference sequence. For human DNA sequencing, this is typically a reference genome such as *GRCh38*. If the fragments of DNA were prepared by randomly sheared fragments, then the sampling of a wide diversity of fragments improves quality and is ensured by sequencing that is from exact duplicates. When the fragments are assembled against the reference genome, the quality of each of the base calls at specific nucleotide positions can be determined. The variant at each position is then determined and reported in a single variant call format file *(VCF). Tertiary analysis:* The variants are then queried against multiple databases that have information on population frequency and clinical significance. Based on these queries, the variants can be prioritized in terms of importance to the given scientific question or clinical scenario. (Modified from Oliver, G. R., Hart, S. N., & Klee, E. W. [2015]. Bioinformatics for clinical next generation sequencing. *Clinical Chemistry, 61,* 124–135.)

discovery efforts. These studies of tens to thousands of cancers across many tissues of origin have cataloged the genomic landscape of human cancers by MPS and often include data from RNA expression and DNA methylation.

BOX 47.3 Phred Quality Score (Q Score)

In the 1990s, Phil Green at the University of Washington developed software to automatically read the fluorescent sequence chromatograms generated from Sanger sequencing. The original software, **Phred** (**Ph**il's **r**ead **ed**itor), used the following basic parameters:

1. Find the predicted location of peaks.
2. Find the observed location of peaks.
3. Match predicted and observed peaks.
4. Find missing peaks.

A component of Phred was an estimator of the error probability of a base call. A quality value (Q score) was generated from the formula:

$$q = -10 \times \log_{10}(p)$$

q = quality value
p = estimated error for a base call

Some representative examples of quality value (q) scores:

Q score of 30 (Q30): The probability *(P)* is 1/1,000 of being incorrect.

Q score of 20 (Q20): The probability *(P)* is 1/100 of being incorrect.

Q score of 10 (Q10): The probability *(P)* is 1/10 of being incorrect.

Although massively parallel sequencing does not generate a Sanger sequencing type of chromatogram, the convention of a Phred Q score is still used to calculate the quality (and accuracy) of a sequenced base. Under ideal conditions, current massively parallel sequencing can achieve more than 90% of bases at Q30.

Sampling and Preservation Methods for Solid Tumor Genomics

Solid tumors can be sampled using either fine-needle aspiration or core biopsy procedures. These are often done in advance of surgery or in patients with unresectable tumors as determined by imaging to provide a sample for pathology-based examination and diagnosis. Fine-needle aspiration (using a 21- to 25-gauge needle) generates a minimal amount of tissue consisting of a few tumor cells in a fluid that is coaspirated into the needle. Core biopsies (using an 18- to 21-gauge needle) can obtain intact and solid cores from the tumor mass. If imaging is used to guide the biopsy needle to the tumor mass, several passes are made for better sampling. Alternatively, in patients with resectable cancer, the tumor mass is removed at surgery and may be preserved in bulk or sampled with core needle biopsies.

Tissue fixation by soaking in formalin or formaldehyde stabilizes cellular proteins and preserves tissue structure for microscopic examination. Subsequent embedding in paraffin wax creates an impervious, room temperature–stable substrate that can be cut for subsequent characterization and diagnosis. However, formalin cross-links proteins to cytosine residues in DNA and RNA. Subsequent oxidation results in breaks in the nucleic acid sugar-phosphate

backbone, thereby degrading the nucleic acids. Because diagnostic assays of solid tumors include nucleic acids, alternative preservatives and preservation methods are now being developed. These methods include flash freezing tissue at −80°C (in a dry ice acetone bath) or immersion in nucleic acid stabilizers.

Staining and Selection of Tumor Cells

A hematoxylin and eosin stain (H&E) is often used to identify tumor and normal cells in a tissue section under a light microscope and to estimate the percentage of tumor nuclei present. Because solid tumors are a mixture of tumor cells and various normal cells, this estimate of tumor "cellularity" indicates how tumor-rich the sample is relative to other needle biopsies or other portions of the bulk tumor. Tumor features such as necrotic areas may be also identified in the microscopic evaluation. In sections with evident necrosis, a process of macrodissection to cut away the nonnecrotic sections for subsequent nucleic acid isolation can be used. In tumor types such as prostate and pancreatic adenocarcinomas, there is a low proportion of tumor nuclei relative to the surrounding normal cells (e.g., stroma and immune cells). In these tumor types, laser capture microdissection can be used to isolate tumor cells from surrounding normal cells before nucleic acid isolation in order to enhance the yield of tumor-derived nucleic acids.

Cancer Is a Genomic Disease

Although we have known for many decades that cancer's origins lie in changes to the cellular genome, only recently have technologies become available to profile the mutations in genes. Initial gene cloning efforts in the early 1980s identified the chromosomal locations and sequences of many oncogenes and tumor suppressors in the human genome. The decoding of the human genome opened the door to genome-wide studies of cancer. For example, learning the sequences of human genes enabled the construction of microarrays to query RNA expression in tumors. By using advanced bioinformatic analysis approaches (e.g., clustering algorithms), similarities and differences in **gene expression** across tumors of a single tissue type (e.g., lung adenocarcinoma) were revealed. Clusters of gene expression correlated with other pathology-based categories revealed subtypes within a given cancer, such as the intrinsic subtypes of breast cancer. Similarly, SNP microarrays can be used to identify chromosomal aberrations by comparing normalized signal strength between tumor and normal DNA. Research-based inquiries of cancer genomes with rich clinical data such as information regarding treatments and outcomes have led to the clinical utility (diagnosis, prognosis, treatment decisions) of the resulting classifications.

The first application of MPS to decode a cancer genome compared the WGS data from a bone marrow aspirate of a patient's AML to a matched normal skin sample. This comparison identified several somatic point mutations that were unique to the tumor genome and were termed single nucleotide variants, or SNVs, to distinguish them from constitutional SNPs. The second WGS case published was also on a single patient's AML sample; it revealed an unexpected somatic mutation in the gene *IDH1* (a metabolic enzyme in the glycolytic pathway). This *IDH1* mutation was then characterized as a recurrent mutation in a panel of 188 AML samples and also associated with a poor prognosis. From these single-patient beginnings, the field of cancer genomics has exploded to characterize tens of thousands of cases across the predominant adult and pediatric cancer types.

Large-Scale Discovery Efforts

One consequence of rapid, high-throughput, and inexpensive sequencing has been the production of large data sets that explore the mutational, transcriptional, and methylation landscape of thousands of tumors representing major human cancer types. Examples of these large-scale discovery projects include The Cancer Genome Atlas (http://cancergenome.nih.gov), the International Cancer Genome Consortium (https://icgc.org), the Pediatric Cancer Genome Project (http://www.pediatriccancergenomeproject.org/site). Unexpectedly, many commonly mutated genes were identified in multiple tumor types and different tissue sites had widely different mutational burdens. New classes of genes—including those involving the spliceosome complex, cellular metabolism, and DNA packaging in histones—were identified that contained previously unknown mutations. Many of these mutations discovered by genomic methods have yet to be functionally studied to determine their effects on the resulting protein.

Meta-analyses of these catalogs were pursued, further reinforcing the similarities and differences between tissue sites by virtue of integration across various "omic" data sets. The similarities, in particular, challenged long-held notions about tissue site specificity of therapeutic approaches and introduced the concept that oncogenic "drivers" present in cancers from different tissues could respond similarly to targeted therapeutic interventions. The impact of gene alterations on protein function results in activated or repressed cellular pathways. Hence cancer is a disease of pathway alterations rather than a specific gene and/or protein alteration, and therapeutic decision making should occur using this framework.

Cancer "Hot Spot" Characterization

The discovery of oncogenes and their activating mutations (also known as "hot-spot" or "gain of function" mutations) emerged from gene cloning efforts in the 1980s and 1990s. By combining PCR and sequencing, mutations in these hot spots could be detected in a few days' time. The clinical usefulness of hot-spot characterization to identify a drug target arose from studies of patients with lung cancer who carried mutations in the tyrosine kinase domain of the epidermal growth factor receptor *(EGFR)* gene and who responded to a new class of targeted therapies called tyrosine kinase inhibitors. The correlation of these mutations to treatment response initiated a paradigm in which the clinical assay for hot-spot mutations became the companion diagnostic for the targeted therapy.

MASSIVELY PARALLEL SEQUENCING OF SOLID TUMOR SAMPLES

Clinical assays of solid tumors using MPS to identify somatic mutations in multiple genes ("gene panels") or in exomes have been developed because (1) large-scale, MPS-based discovery efforts have identified mutated genes in the major adult and pediatric cancers; (2) specific genes and/or mutations have been studied in clinically annotated sample sets and their mutational status correlates with outcome; and (3) novel small molecule- and antibody-based targeted therapies have been developed to address specific somatic alterations; therefore the mutational profile of a patient's somatic genome can indicate a therapy based on predicted gene-drug interactions.

Although testing multiple genes from a single tumor DNA isolate could be accomplished using combinations of individual clinical assays to evaluate individual gene and/or hot-spot mutation status, often tumor material is limited owing to the size of a biopsy and the need to use the material for other, more conventional pathology assays. Hence MPS-based assays make more efficient use of diagnostic material, especially when so many genes that are potentially mutated can provide important information regarding patient care.

Because of the heterogeneity of solid tumors and the increasing knowledge that certain mutations in genes, if present at low levels in the tumor cell population, can predict early onset of acquired resistance to targeted therapies, MPS-based assays must provide a high depth of tumor sequence coverage for each gene in the assay. High coverage can be defined as 300- to 1000-fold read depth and can be challenging to obtain when MPS libraries are constructed from low amounts of input DNA. In general, with 500-fold or higher coverage, mutations can be detected even from samples with low percentage tumor cellularity (10% to 20%), but detecting low-prevalence mutations becomes less likely with decreasing tumor cell percentages. High coverage is also important to attain because not all probes capture complementary sequences with the same avidity. When this deficiency is combined with representation bias from MPS processes downstream of capture, lowered coverage at certain sites may result. Hence higher coverage overall may reduce or eliminate low and/or no coverage sites, thereby reducing the incidence of false-negative results, wherein a mutation is missed due to lack of coverage.

Whole Genome Sequencing

Ideally the most unbiased characterization of each cancer sample by MPS-based methods should be pursued to optimize the likelihood of identifying all drug-targetable portions of the genome. This is best accomplished using WGS, which can deliver the totality of point mutations, small insertions and deletions, structural variants (including copy number alterations), and other large chromosomal rearrangements in a tumor-to-normal genome comparison. However, WGS is seldom performed as a clinical assay for several reasons. First, the genome is very large, and although new MPS instrumentation can produce the data rapidly (up to 30-fold coverage for 48 human genomes in 2 days), the postsequencing efforts to analyze a whole genome comparison of tumor tissue to matched normal tissue and compile a clinical report is in conflict with the time frame needed to inform the oncologist's decision-making process (2–4 weeks). Second, owing to the need for high coverage of the tumor genome, generating the necessary 100-fold or higher WGS coverage needed for confident variant detection remains prohibitive in terms of cost. Third, most of the somatic variants will be uninterpretable in terms of their impact on tumor progression because they will lie outside of genes. Fourth, many variants in the coding sequences will not be interpretable in terms of their impact on the protein's function as a potential driver of disease development (i.e., gain or loss of function).

Enrichment by Hybridization Capture

Locus-selective or hybridization capture methods use synthetic probes designed to hybridize to human loci and/or genes of interest as a means of selectively isolating sequences from a whole human genome library. MPS of the captured sequences is followed by alignment of the reads to the genes of interest for mutational analysis. Depending on the assay, either a select set of genes or hot spots, or the exome, can be captured. As the number of loci assayed by this approach increases, so do both the analysis time and the complexity of the mutational signature. The decision to include drug-targetable genes that carry novel, noncanonical mutations with untested drug interactions is one challenge that can arise in producing the clinical interpretation of these assays.

Enrichment by Multiplex Polymerase Chain Reaction

Because hybrid capture works best when many genes are analyzed, multiplex PCR is often used when the number of targets is low. In this type of assay, careful design of PCR primer pairs that target the selected loci for mutation detection achieves the coincident amplification of all loci in a single reaction. Multiplex PCR has the distinct advantages of (1) selectively amplifying smaller target regions of the genome below that which is optimal for selective hybridization capture, (2) requiring only 5 to 10 ng of DNA for the PCR, and (3) focusing mutation detection on a relatively small portion of the genome in an efficient manner. These assays are also relatively inexpensive to perform, and by post-PCR addition of DNA barcoded linkers for MPS, they can be multiplexed with other patients onto a single sequencing instrument run. There are obvious limitations to multiplex PCR that include (1) the ability to design PCR primers with similar T_m, uniqueness and lack of interprimer complementarity to diminish amplification artifacts (e.g., primer dimers, hybrid amplicons) and (2) the presence of pseudogenes and highly conserved regions of gene families may preclude unique amplification of specific gene sequences. PCR primers are often designed with formalin-fixed paraffin-embedded (FFPE)–associated degradation in mind to produce small products of less than 100 bp; therefore, these require a large number of primers to cover a large gene, which further complicates the design process and/or

complicates obtaining complete coverage from the resulting amplicons. Despite the limitations, multiplex PCR assays are widely used in the clinical testing of solid tumors.

Ribonucleic Acid Sequencing of Solid Tumor Isolates

RNA isolates from cancer samples can be assayed using MPS technologies—with a variety of preparatory methods generally referred to as RNA-seq—by converting the RNA to DNA by reverse transcription before library construction. Reasons to assay transcribed genes in a solid tumor include the following: (1) DNA sequencing assays do not reflect higher order changes in the tumor genome, such as methylation or chromatin packaging, that can influence the amount of RNA produced from a gene; (2) in high-mutation-load tumors such as melanomas and lung squamous or adenocarcinomas, RNA-seq can indicate which of the thousands of mutations identified from DNA sequencing assays are actually being transcribed and could therefore be targeted with a specific therapy; (3) alternatively spliced transcript (isoform) expression can be detected only at the level of RNA and may be important in a diagnostic, prognostic, or therapeutic context; and (4) predicted gene fusions from DNA assays can be verified by the presence of the predicted fusion transcript in the RNA-seq data. RNA is an unstable molecule, and preservation techniques such as FFPE degrade RNA more noticeably than DNA. Despite the disadvantages of RNA as an analyte, RNA-seq is emerging as a clinical tool for identifying gene fusions in cancer.

Interpreting Somatic Alterations in a Clinical Context

The use of MPS-based solid tumor diagnostics is increasing across clinical laboratories, both commercial and academic. This is largely due to the increasing knowledge of somatic alterations in cancers, the increasing numbers of targeted therapies designed to interact with known cancer-relevant alterations, and the demands of patients and oncologists to add this information to the evidence being considered in determining the treatment regimen. Progress in this area is confounded by the need to relate somatic alterations to other clinical data, such as the diagnosis of a specific tumor type or subtype, the range of possible patient outcomes, and therapeutic response. Additional studies of the functional impact of a somatic alteration, how that functional impact may contribute to the onset or progression of the cancer, or how and/or whether the alteration interacts with a variety of potential therapies based on putative drug–gene interactions are often needed. Drug studies typically involve preclinical testing in cell lines or other disease models, followed by clinical trials that require a year or more to compare patient outcomes when patients carrying the gene alteration(s) are treated with a specific therapy as compared with patients treated with the standard of care. In essence, there is often a significant delay between the research-based discovery of a somatic alteration and when its full clinical implication can be established. The clinical sequencing pipeline and how to determine which of the identified mutations in a cancer biopsy should be evaluated further in terms of their prognostic, diagnostic, or therapeutic value is an area of active investigation. Publicly available curated cancer mutation databases permit look-up of known cancer genes and mutations and offer prognostic, diagnostic, and therapeutic information. These include TARGET (www.broadinstitute.org/cancer/cga/target) and CIViC (www.civicdb.org), cBioportal, and GENIE.

Prognostic Interpretation

Understanding the prognostic or outcome-related value of a specific gene alteration depends on the prevalence of the disease, the time to establish outcome, the specific outcome measure, and the uniformity of patient care in each group. The time to establish outcome can vary widely; for example, patients with lung squamous cell carcinoma or glioblastoma typically present with metastatic disease or progression in 6 to 18 months after a primary diagnosis, whereas ER+ breast cancer may relapse after 15 to 20 years. The defined outcome may be overall survival, disease-specific survival, complete pathologic response, or other response metrics. Some examples of tumors where mutations provide prognostic information include gliomas as well as colorectal, endometrial, and ovarian cancers.

Diagnostic Interpretation

Multigene capture panels that detect alterations in known cancer susceptibility genes such as *TP53, APC, BRCA1/2*, and others in germline DNA can be used to identify inherited cancer susceptibility. Diagnostic characterization of solid tumor DNA based on somatic alterations is taking on increased significance owing to the availability of targeted therapies that correspond to these somatic (or germline) alterations. Non–small cell lung cancers, especially adenocarcinomas, perhaps provide the richest example of sequencing-based diagnosis and therapeutic option guidance among the solid tumors. In 2012, the spectrum of genomic alterations in genes that drove the development of non–small cell lung adenocarcinomas encompassed approximately 50% of the disease, of which approximately 40% were therapeutically targeted (Fig. 47.12). Increasingly, a gene panel assay is pursued to sequence these genes or gene fusions, and can indicate one or several targeted therapies that represent the first line of therapy. This paradigm shift represents a remarkable transition in medical practice over just a 10-year period from when lung adenocarcinoma mutations that predicted response to tyrosine kinase inhibitors were first described and is indicative of the directions in diagnostic molecular pathology for other solid cancers. However, there are several challenges to this paradigm that merit further discussion.

Challenges of Variants of Uncertain Significance

As somatic gene panel assays expand to generate sequence data across the length of each gene rather than from focused hot spots, more variants are identified. In particular, the resulting breadth of variant discovery permits the detection of variants for which no clear indication of response for a

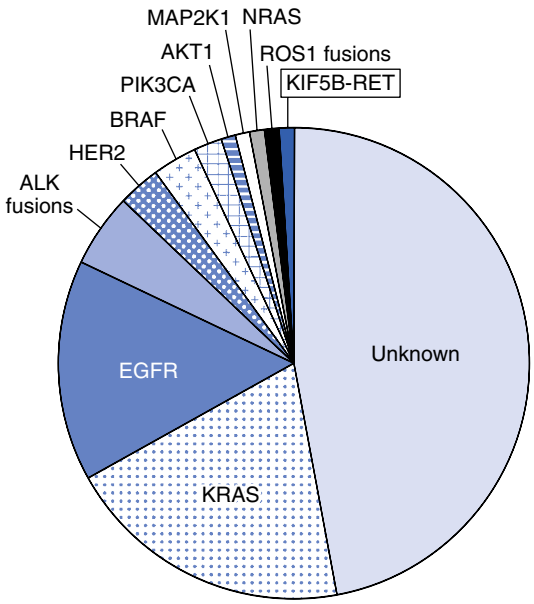


Fig. 47.12 Molecular subsets of lung adenocarcinoma. The clinically important driver mutations and fusions identified to date from molecular studies of lung adenocarcinoma. The width of the chart wedge for each driver event corresponds to its percentage distribution in the disease. (Reprinted by permission from Macmillan Publishers Ltd. From Pao, W., & Hutchinson, K. E. [2012]. Chipping away at the lung cancer genome. *Nature Medicine*, 18, 349–351.)

given therapeutic indication might be available. These variants of uncertain significance (VUS) raise issues with regard to whether the result should be reported because the therapeutic indication is uncertain. Although basic science experiments may eventually characterize the functional impact of all variants in all cancer-related genes, most molecular pathology groups that are conducting selective hybridization capture panel assays of cancers are instead developing local databases to catalog VUS in an effort to initially provide a record of whether the variant has been seen in previously completed assays. Over time, data mining of electronic medical records for patients with variants, targeted therapy, and response data may transition VUS to therapeutic indications.

Challenges of Off-Label Indications

The current process for clinical trials of new therapies in cancer toward approval by the US Food and Drug Administration (FDA) rests on a paradigm in which trials compare each new therapy in a specific tumor type to the standard of care. Hence, cancer therapies are tested and approved in a disease site–specific manner. By contrast, cancer genomics efforts have demonstrated that known cancer-initiating gene alterations occur in multiple disease and/or tissue sites. This fact, in turn, raises the possibility that a given targeted therapy might be effective for multiple solid tumor types that all carry the same somatic alteration. There is support for multiple-tissue-site therapeutic effectiveness, including the FDA-approved use of the kinase inhibitor imatinib (Gleevec) in both chronic myeloid leukemia targeting *ABL* kinase and gastrointestinal stromal tumors targeting *KIT*. Another example is EGFR-targeting therapies with FDA approval for use in lung adenocarcinomas, head and neck squamous cell carcinomas, and colorectal cancers with somatically altered *EGFR*. These approved tissue sites each required a separate series of clinical trials to gain FDA approval. However, in solid tumor assays with gene panels, off-label indications will arise due to the shared nature of cancer-initiating mutations across different tissue sites. This challenge occurs when a gene with a functional mutation is detected that indicates a therapy not yet approved in the specific tissue site of origin.

TRENDS IN CANCER GENOMICS

Solid tumor gene-based assays are in a dramatic transition to multigene testing. This trend is driven by (1) the increasing use of selective hybridization capture or other targeting approaches coupled with MPS, (2) the discovery of somatic alterations in the genomes of cancer cells, and (3) the emerging number of targeted therapies that address specific cancer-initiating genes or gene families. The comprehensive nature of multigene assays creates specific challenges, mainly due to the impedance mismatch between cancer mutation discovery and the functional characterization of those alterations that either contribute to the development and progression of cancer or do not contribute to it. The need to accelerate testing of new therapeutics across multiple tissue sites at once is slowly being addressed by new clinical trial designs to identify the most effective therapies for a given gene, alteration, and tissue site. Ultimately the widespread use of multigene testing for cancer diagnosis, prognosis, and therapeutic decision making rests on the demonstration of clinical utility and clinical benefit, which are established for some tissue sites but as yet not for all.

POINTS TO REMEMBER

- Gene expression microarrays were initially used to characterize large numbers of human cancers.
- Cancer gene cloning efforts and the completed Human Genome Reference Sequence enabled the use of PCR and sequencing efforts to catalog cancer mutations.
- Massively parallel sequencing enables the identification of somatic alterations in targeted gene panels in all known genes or the whole genome.

SUMMARY

The field of molecular diagnostics is an important and exciting area that is going to have an even greater impact on medicine in the future. As an increasing number of diseases are characterized at the molecular (e.g., nucleic acid and protein) level, new therapeutics and diagnostics specifically targeting these molecular changes will continue to emerge.

REVIEW QUESTIONS

1. The number of human proteins is greater than the number of genes because
   a. alternative splicing and posttranslational modifications occur.
   b. of the 3′ polyadenylation of mRNA.
   c. only some genes are expressed.
   d. the ribosome adds additional noncoded amino acids to some mRNAs.
2. How much of the human genome is transcribed into RNA?
   a. 0.1%
   b. 1%
   c. 10%
   d. 80% to 100%
3. If there were only 12 amino acids in proteins, what would be the minimal length of nucleotides needed for a complete genetic code?
   a. 1
   b. 2
   c. 3
   d. 4
4. The number of DNA bases in a human cell is about
   a. 3 billion
   b. 6 trillion
   c. 20,000
   d. 64
5. How would the sequence GAATATGCCAGCATATTAG translate?
   a. MetProAlaTyr
   b. GluTyrAlaSerIle
   c. It would not translate because the number of bases is not divisible by 3.
   d. It would not translate because it will not be transcribed.
6. Nucleotides that encode protein account for what percentage of the genome?
   a. 1%
   b. 10%
   c. 50%
   d. 90%
   e. 8%
7. What approximate percentage of the human genome comprises transposable elements?
   a. 1%
   b. 10%
   c. 20%
   d. 30%
   e. 40%
8. When massively parallel sequencing generates data with a Phred quality score of Q20, the probability of being incorrect is which of the following?
   a. 1 in 10
   b. 1 in 100
   c. 1 in 1000
   d. 1 in 10,000
   e. 1 in 100,000
9. What is the correct nomenclature for the variant in the following sequence if the base in *blue boldface text* is converted to a C? Hint: The sequence includes the initiation codon of the gene.

ACAGCATAGCATATGACGCATCAGCACATT

   a. GRCh37/hg19 g.32335623
   b. p.W4A
   c. c.12G > C
   d. c.-7G > C
   e. p.T12P
10. Hydrogen bonding between which DNA elements is required for most molecular diagnostic assays?
   a. Bonds between ribose sugars on different nucleotides of a strand of DNA
   b. Bonds between ribose and nucleotide bases
   c. Bonds between two cysteine residues
   d. Bonds between the complementary nucleotide bases
11. What nucleotide base is methylated in DNA and at what specific location?
   a. Adenine and polyadenylation
   b. Guanine and RNA transcription start site
   c. Thymine and the production of thymine dimers
   d. Cytosine and upstream CpG islands
12. Mutations can be generated by incorporating the wrong base into DNA during replication, and the mechanism to repair this error is
   a. mismatch repair
   b. nucleotide excision repair
   c. base excision repair
   d. homologous recombination repair
13. The structure of a human mRNA encoding gene generally contains promoters that are upstream from the transcriptional start site; however, there are other promoters that can be both upstream and downstream of the gene. Which of the following is one of these promoters?
   a. The TATA box
   b. The polyadenylation site
   c. The enhancer
   d. The proximal core promoter
14. An important group of DNA modification enzymes, the restriction endonucleases, are used to manipulate DNA fragments. What is their function in the organisms from which they are isolated?
   a. They digest the ends of DNA to prepare them for ligation
   b. They are used to protect themselves from viral infections
   c. They are used in the base excision repair found in bacteria
   d. They are RNA-guided nuclease that can precisely cleave genomic DNA
15. Solid tumors may consist of
   a. 100% tumor cells
   b. a mixture of variable proportions of tumor and normal cells

c. a mixture of blood vessels, tumor, and normal cells of different types
d. only chromosomal DNA

16. Cancer is considered a disease of the genome because
a. cancer cell nuclei contain no chromosomal DNA
b. alterations to genes in cancer cells disrupt cellular pathways
c. the genome in cancer cells does not express any genes
d. the DNA of cancer cells is cross-linked to proteins

17. Multiple gene testing of cancer DNA may determine
a. tumor cellularity
b. mutations in driver genes that predict a response to therapy
c. prognostic mutations that predict outcome
d. b and c

SUGGESTED READINGS

Abecasis, G. R., Auton, A., Brooks, L. D., et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*, 56–65.

Cech, T. R., & Steitz, J. A. (2014). The noncoding RNA revolution-trashing old rules to forge new ones. *Cell*, *157*, 77–94.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, *227*, 561–563.

Djebali, S., Davis, C. A., Merkel, A., et al. (2012). Landscape of transcription in human cells. *Nature*, *489*, 101–108.

Green, E. D., Guyer, M. S., & National Human Genome Research Institute (2011). Charting a course for genomic medicine from base pairs to bedside. *Nature*, *470*, 204–213.

International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*, 931–945.

Jackson, R. J., Hellen, C. U., & Pestova, T. V. (2010). The mechanism of eukaryotic translation initiation and principles of its regulation. *Nature Reviews Molecular Cell Biology*, *11*, 113–127.

Kandoth, C., McLellan, M. D., Vandin, F., et al. (2013). Mutational landscape and significance across 12 major cancer types. *Nature*, *502*, 333–339.

Landrum, M. J., Lee, J. M., Riley, G. R., et al. (2014). Clinvar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, *42*, D980–D985.

Ley, T. J., Mardis, E. R., Ding, L., et al. (2008). DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, *456*, 66–72.

Redon, R., Ishikawa, S., Fitch, K. R., et al. (2006). Global variation in copy number in the human genome. *Nature*, *444*, 444–454.

Schubeler, D. (2015). Function and information content of DNA methylation. *Nature*, *517*, 321–326.

Sorlie, T., Perou, C. M., Tibshirani, R., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceeding of the National Academy of Sciences of the United State of America*, *98*, 10869–10874.

Van Allen, E. M., Wagle, N., Stojanov, P., et al. (2014). Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nature Medicine*, *20*, 682–688.

Venter, J. C., Adams, M. D., Myers, E. W., et al. (2001). The sequence of the human genome. *Science*, *291*, 1304–1351.

Watson, J. D., & Crick, F. H. (1953). Genetical implications of the structure of deoxyribonucleic acid. *Nature*, *171*, 964–967.

Wheeler, D. A., & Wang, L. (2013). From human genome to cancer genome: the first decade. *Genome Research*, *23*, 1054–1062.

48

Molecular Techniques

Carl T. Wittwer, G. Mike Makrigiorgos

OBJECTIVES

1. Define the following:
   - Amplification
   - Hybridization
   - Melting analysis
   - Primer
   - Probe
   - Polymerase chain reaction (PCR)
   - Massively parallel sequencing
   - Real time PCR
   - Hydrolysis probe
   - Hybridization probe
2. Describe the steps involved in the PCR. Reason why a short defined PCR product is produced.
3. State the difference between target amplification and signal amplification.
4. Describe the different ways that fluorescence can be generated during real-time PCR.
5. Describe the differences between transcription mediated amplification (TMA) and the PCR.
6. Describe how digital PCR is different from regular PCR.
7. Detail the steps required for massively parallel sequencing using reversible terminators.
8. Describe one method for direct quantification of nucleic acids without amplification.
9. Detail how to perform high-resolution melting analysis.
10. Describe the kind of controls needed for the PCR.

KEYWORDS AND DEFINITIONS

Adapter Oligonucleotides that are ligated to library fragments in order to provide consensus priming sites.

Allele-specific PCR A version of polymerase chain reaction (PCR) in which only one allele at a locus is amplified. Specificity is achieved by designing one or both PCR primers so that they overlap the site of sequence variance between alternative alleles.

Asymmetric PCR A version of PCR that preferentially amplifies one strand of the target DNA.

cDNA Complementary DNA, usually refers to the DNA formed by reverse transcription; that is, DNA that is complementary to the original RNA.

Comparative genomic hybridization A technique in which reference and test DNA are respectively labeled with green and red fluorochromes. Abnormalities are detected by changes in the green-to-red ratio.

Deoxyribonucleoside triphosphates (dNTPs) Usually dATP, dCTP, dGTP, and dTTP; the building blocks of DNA.

Dideoxy-termination sequencing (Sanger sequencing) A method of DNA sequencing based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication.

Digital polymerase chain reaction (digital PCR or dPCR) A modification of PCR with the sample separated into many partitions so that some partitions have no template (0) and others have one or more (1).

DNA microarray An array of microscopic spots of different DNA molecules attached to a solid surface. Also known as DNA chips or biochips.

Electrophoresis Movement caused by an electrical field, often through a gel matrix. Polyacrylamide and agarose are common matrices used to separate DNA and RNA under an electric field.

Flow cytometry A technique for counting cells or particles suspended in fluid as they flow one at a time past a focus of exciting light.

Fluorescence A physical property of some molecules to emit light at a longer wavelength when excited at a shorter wavelength.

Fluorescence in situ hybridization (FISH) A genetic mapping technique using fluorescent tags for analysis of chromosomes and genetic abnormalities. FISH can also be referred to as chromosome painting.

Heteroduplex A DNA duplex with internal mismatches or loops.

High-resolution melting analysis (HRMA) A simple method for genotyping that compares fluorescence against temperature using a dye that detects double-stranded but not single-stranded DNA.

Homoduplex A perfectly matched DNA duplex.
Hybridization The annealing or pairing of two DNA strands.
Insert Part of the original DNA that has been fragmented before ligation to adapters.
Loop-mediated isothermal amplification (LAMP) A single tube technique for the amplification of DNA that uses a single temperature incubation.
Massively parallel sequencing (MPS) Sequencing of many fragments of DNA simultaneously.
Melting temperature (Tm) The temperature at which half of the DNA strands are in the duplex form.
Multiplex ligation-dependent probe amplification (MLPA) A variation of the multiplex PCR that assesses the copy number of several targets by permitting multiple targets to be amplified with only a single primer pair.
Oligonucleotide A short single-stranded polymer of nucleic acid.
Oligonucleotide ligation assay (OLA) A technique for determining the presence or absence of a specific nucleotide pair within a target gene, often indicating whether the gene is wild type (normal) or mutant (defective).
Polymerase chain reaction (PCR) An in vitro method for exponentially amplifying DNA.
Primer An oligonucleotide that serves to initiate polymerase-catalyzed addition of nucleotides by annealing to a template strand.
Probe A nucleic acid used to identify a target by hybridization.
Pyrosequencing A method of DNA sequencing based on the "sequencing by synthesis" principle that relies on the detection of pyrophosphate release on nucleotide incorporation.
Read The nucleotide sequence inferred by sequencing of a template.
Real-time PCR Observation of PCR during amplification at least once each cycle.
Restriction fragment length polymorphism (RFLP) A genetic polymorphism revealed by changes in the sizes of DNA fragments after restriction enzyme digestion and electrophoresis.
Rolling circle amplification (RCA) A probe amplification method with a linear probe that is ligated to form a circle in the presence of template. The circle is replicated continuously by a polymerase and one or more primers.
Sanger sequencing See dideoxy-termination sequencing.
Single nucleotide variant (SNV) Any change in a nucleic acid that involves only a single base, including substitutions and single base insertions/deletions. Favored over the related term "single nucleotide polymorphism" (SNP) and not limited to variants present at a frequency of greater than 1% as is SNP.
Strand displacement amplification (SDA) An amplification technique that uses two types of primers, DNA polymerase, and a restriction endonuclease to exponentially produce single-stranded amplicons asynchronously.
Transcription-mediated amplification (TMA) An amplification method that uses RNA polymerase and DNA reverse transcriptase to produce RNA amplicon from a target nucleic acid. TMA is used to amplify both RNA and DNA.
Whole-genome amplification (WGA) A nonspecific amplification technique that produces an amplified product representative of the initial starting material (whole genome).

The expansion and power of molecular diagnostics is enabled by techniques that modify, amplify, detect, discriminate, and sequence nucleic acids (NAs). Molecular diagnostic techniques are getting faster, better, and less expensive as a rapidly progressing and highly competitive field, both in academics and in industry. This growth is reflected in the number of publications in molecular diagnostics in recent years (Fig. 48.1). In this chapter, our intent is to focus on understanding the mechanisms of NA techniques in use today.

We begin by considering the amplification techniques that are often necessary to observe or quantify NA sequences of interest. Next, the tools used to detect or visualize NAs are discussed, along with methods that allow identification, quantification, and segregation of individual NA species. Finally, we end with the extraordinary power of **massively parallel sequencing (MPS)**.

NUCLEIC ACID PREPARATION

Conventional NA testing requires (1) isolation of DNA, RNA, or both from the clinical sample; (2) amplification of the NA; and (3) detection, analysis, or quantification. Sometimes one step can be eliminated or two steps can be combined, depending on the sample type and quantity of the target. In addition to NA isolation, sometimes it is necessary to prepare samples for amplification or analysis by enzymatic or nonenzymatic means. For example, the enzyme reverse transcriptase may be needed to convert RNA to DNA. MPS usually requires library preparation.

Bisulfite treatment of DNA is often used to analyze the methylation status of cytosine (C) residues in DNA. Sodium bisulfite ($NaHSO_3$) converts C into uracil (U), but does not affect 5-methylcytosine. DNA polymerases used for amplification will

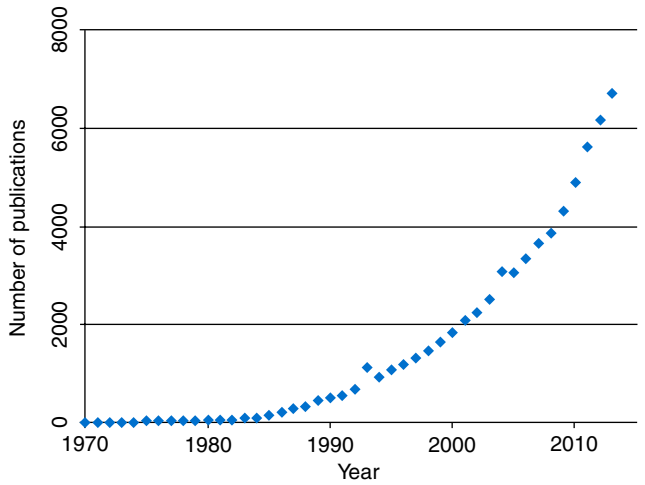


Fig. 48.1 Molecular diagnostics publications, 1970 to 2013. (Reprinted with permission of the publisher from Chiu, R. W. K., Lo, Y. M. D., & Wittwer, C. T. [2015]. Molecular diagnostics: A revolution in progress. *Clinical Chemistry, 61*, 1–3. Copyright 2015 AACC.)

recognize 5-methylcytosine as C (no change), but an unmethylated C is converted to U by the bisulfite and will be recognized as a T (sequence change of C to T). One limitation of bisulfite treatment is significant degradation of DNA. Depending on the protocol, as much as 90% of the DNA can be lost.

AMPLIFICATION TECHNIQUES

Molecular diagnostics requires techniques to detect extremely low concentrations of NAs in a background of complex genomic structure. Achieving sensitive detection is a central concern for clinical applications of NA analysis. Techniques that increase the amount of the NA target or the detection signal are referred to as *amplification methods*. Examples of amplification methods are listed in Table 48.1 and the Points to Remember box. In *target amplification*, a well-defined segment of the NA (the target sequence) is copied many times. Areas outside the target are not amplified. In *signal amplification*, the amount of target stays the same, but the signal is increased. Amplification techniques often can achieve more than a million-fold amplification in less than 1 hour.

Polymerase Chain Reaction—Target Amplification

When the amount of target NA is increased by synthetic in vitro methods, target amplification occurs. **Polymerase chain reaction (PCR)** is the best known and most widely applied of the target amplification methods. Because of the commercial availability of thermostable DNA polymerases, kits, and instrumentation, this method has been widely adopted in research and is routinely used in the clinical laboratory.

POINTS TO REMEMBER

Nucleic Acid Amplification Methods

- May amplify the target, the signal, or the probe
- May be isothermal or cycle through different temperatures
- Cycling speed is limited by instrumentation, not biochemistry
- May be analyzed on gels or in real time
- May be qualitative, quantitative, or digital
- Require positive and negative controls

TABLE 48.1 Common Amplification Techniques

Techniques	Type	Enzymes Required
Polymerase chain reaction (PCR)	Target	DNA polymerase (thermostable)
Transcription-mediated amplification (TMA)	Target	Reverse transcriptase, RNA polymerase, RNase H
Strand displacement amplification (SDA)	Target	*Hinc*II, DNA polymerase I (5′-exo-deficient)
Loop-mediated amplification (LAMP)	Target	DNA polymerase
Helicase-dependent amplification (HDA)	Target	Helicase, DNA polymerase
Recombinase polymerase amplification (RPA)	Target	Recombinase, single-strand binding protein, polymerase
Whole-genome amplification (WGA)	Target	Φ29 DNA polymerase
Antisense RNA amplification (aRNA)	Target	T4 DNA polymerase, Klenow, S1 nuclease, T7 polymerase
Branched DNA (bDNA)	Signal	Alkaline phosphatase
Rolling circle amplification (RCA)	Probe	Φ29 DNA polymerase

Polymerase Chain Reaction Process

PCR requires (1) a thermostable DNA polymerase, (2) deoxynucleotides of each base (collectively referred to as **deoxyribonucleoside triphosphates [dNTPs]**), (3) a DNA strand that includes the sequence to be amplified, and (4) a pair of **oligonucleotides** (referred to as **primers**) that are complementary to opposite strands that define the target sequence. In the first step, target duplexes are denatured into single strands by heat (Fig. 48.2). When the mixture is cooled, primers provided in great excess (usually over 1 million times the concentration of the initial target) specifically anneal to complementary sequences on the target. After the primers are annealed, the action of the polymerase synthesizes two new DNA strands by extending each of the two primers at their 3′ end. The primers are designed to recognize sequences of the target that are close enough together such that the polymerase extends each strand far enough to include the priming site of the other primer. Usually, the optimal temperature for polymerization (or DNA synthesis) is an intermediate temperature between the annealing temperature (at which primer hybridizes to its target) and the denaturation temperature (at which the newly generated DNA strand dissociates from its

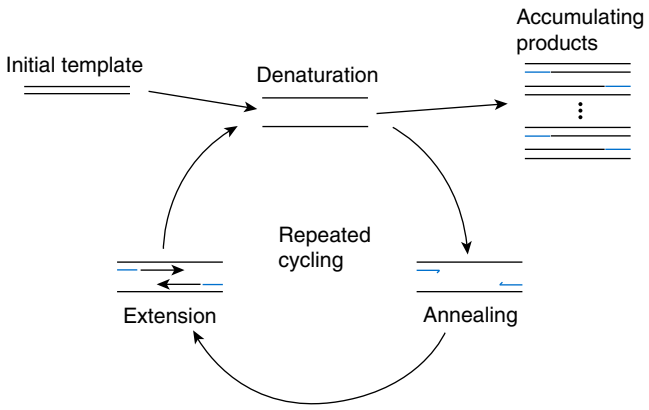
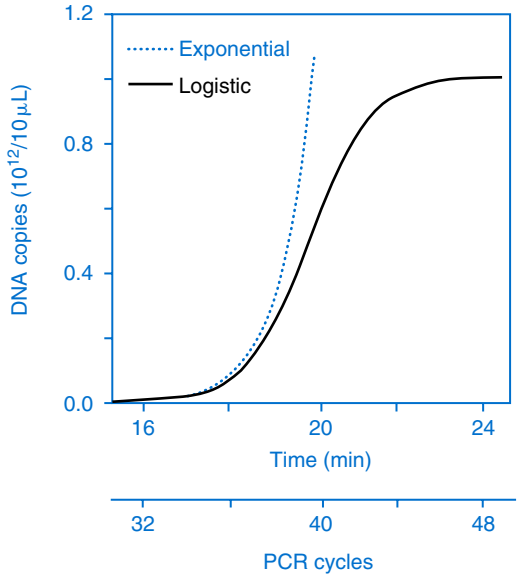


Fig. 48.2 Simple schematic of the polymerase chain reaction. Amplification of the initial template requires denaturation by heat, allowing primers to anneal at a lower temperature, followed by primer extension at an intermediate temperature. In each cycle, a doubling of the DNA product occurs. After 20 to 40 cycles, the product accumulates more than 1 million-fold.



DNA amplification

Fig. 48.3 Exponential and logistic curves for DNA amplified by polymerase chain reaction *(PCR)*. Starting with only one target copy, it takes 46 cycles to amplify the target to saturation when the doubling efficiency is 100% (Modified with permission of the publisher from Wittwer, C. T., Kusukawa, N. [2004]. Real-time PCR. In D. H. Persing, F. C. Tenover, J. Versalovic, [Eds.], *Molecular microbiology: diagnostic principles and practice* [pp. 71–84]. Washington, DC: ASM Press. Copyright 2004 ASM Press.)

template). The second cycle also begins with denaturation, but now twice as many strands (the original genomic DNA and the extension products from the first cycle) are available for primer annealing and subsequent extension. Temperature cycling (typically) uses three temperatures: (1) a high temperature sufficient to denature the target sequence, typically 90°C to 97°C but can be lower depending on the **melting temperature** of the product; (2) a low temperature that allows annealing of the primers to the target, typically 50°C to 65°C; and (3) a third temperature, which is optimal for polymerase extension, typically 65°C to 75°C.

Repetitive thermocycling results in the exponential accumulation of the short product (consisting of primers and all intervening sequences). If the efficiency of each cycle is perfect, the number of target sequences doubles each cycle (the efficiency is 100% or 1.0). PCR efficiency depends on the primers, the temperature-cycling conditions, and any inhibitors that might limit amplification. Amplified products accumulate exponentially in the beginning cycles of PCR. At some point, however, the efficiency of amplification falls and eventually the amount of product plateaus (Fig. 48.3) as the result of exhaustion of the components or competition between the primer and the product annealing (i.e., the single strands of the product are at such high concentrations that they anneal to each other rather than to the primers). The S-curve shape is similar to common models of population growth. In a typical PCR using 0.5 µmol/L of each primer, the maximum DNA concentration typically achieved is about 100 billion copies/µL.

With the addition of reverse transcriptase, RNA targets can be converted into **cDNA** and then successfully amplified. Reverse transcription and DNA amplification are most often catalyzed by two different polymerases. In one-step reverse transcriptase PCR (RT-PCR), both enzymes are present in a common buffer, and typically PCR primers are used for both reverse transcription and DNA amplification. Some thermostable enzymes have both DNA polymerase and reverse transcriptase activities so that both steps can be performed in the same tube with the same enzyme. In two-step RT-PCR, the reverse transcription is performed first, usually with random hexamers or a poly-dT oligonucleotide (to prime the poly-A tail of most mRNAs). After reverse transcription, the second PCR step is performed on cDNA with specific primers.

After amplification, the products can be detected by various methods. Gel **electrophoresis** with ethidium bromide staining is a classical method that separates products by size and may suffice for many applications. For fine size discrimination down to a single base, one of the primers can be fluorescently labeled and the post-PCR fragments separated in capillaries on instruments that are typically used for conventional **dideoxy-termination (Sanger) sequencing**. Alternatively, some form of **hybridization** assay can be used to verify or analyze the amplified product. Automated methods are always attractive, and closed-tube methods, in which amplified products are never exposed to the open environment, are particularly advantageous to avoid contamination of future reactions with the products of a prior reaction. Adding a fluorescent dye or **probe** before amplification allows optical monitoring to follow the reaction as it progresses (**real-time PCR**) or after the reaction is complete (endpoint melting) without the need to process the sample for a separate analysis step.

Polymerase Chain Reaction Kinetics

It is natural to think about PCR in terms of three events—denaturation of double-stranded target, annealing of primers

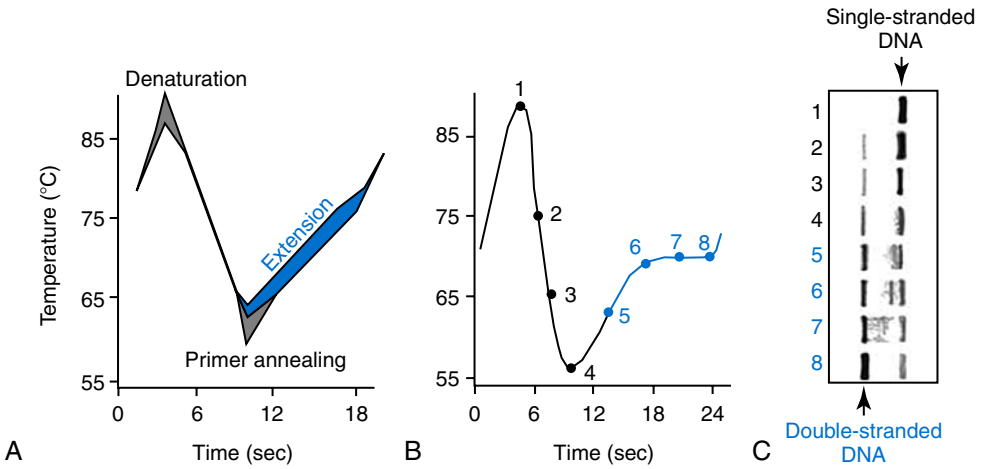


Fig. 48.4 A visual demonstration of polymerase chain reaction (PCR) kinetics. The three phases of PCR (denaturation, annealing, and extension) occur as the temperature continuously changes (A). Toward the end of temperature cycling, the reaction contains single- and double-stranded PCR products. When different points of the cycle (B) are sampled and analyzed, the transition from denatured single-stranded DNA to double-stranded DNA appears as a continuum (C). Progression of the extension reaction can be followed by additional bands appearing between the single- and double-stranded DNA (time points 5 to 7). (Modified with permission from Wittwer, C. T., & Herrmann, M. G. [1999]. Rapid thermal cycling and PCR kinetics. In M. Innis, D. Gelfand, J. Sninsky [Eds.], *PCR applications* [pp. 211–229]. San Diego: Academic Press. Copyright 1999 Academic Press.)

to their targets, and extension of the DNA strand from the primer—that occur at three different temperatures, each requiring a certain amount of time. Indeed, it is common to perform PCR by holding the reaction mixture at three different temperatures (e.g., denaturation at 94°C, annealing at 55°C, and extension at 72°C). Standard thermocyclers that use conical tubes focus on accurate temperature control of the heating block at equilibrium, not on dynamic control of the sample temperature throughout cycling. As a result, sample temperatures are not well defined during transitions, and long cycle times have become standard to ensure that samples reach target temperatures. Reproducibility between instruments and manufacturers is poor, and PCR may require an hour or more to complete 30 cycles of amplification.

The kinetics of denaturation, annealing, and extension suggest that rapid transitions between temperatures with minimal or no pauses (temperature plateaus) provide a better paradigm of PCR amplification (Fig. 48.4). Denaturation, annealing, and extension are very rapid reactions as shown by experiments in capillaries. The use of temperature "spikes" at denaturation and annealing, instead of extended temperature plateaus, allows for rapid cycling with the right instrumentation. The actual time required for PCR depends on the size of the product, but when it is less than 500 base pairs (bp), 30 cycles can be easily completed in 15 minutes.

Initial denaturation of genomic DNA may be required before PCR cycling, depending on how the DNA sample was prepared. Prior boiling of the sample or an initial denaturation step of a few seconds at 95°C to 99°C may be necessary to denature very long strands of genomic DNA. During PCR, however, denaturation of the shorter products occurs very rapidly. Even for long PCR products, denaturation is complete in less than 1 second at temperatures greater than 3°C above the PCR product Tm.

Product specificity is optimal when annealing times are less than 1 second. Longer annealing times may be required if the primer concentrations are low. The required extension time for each cycle depends on the length of the PCR product. Extension is not instantaneous, although it is much faster than common practice would suggest. Extension rates of typical polymerases under typical conditions are 50 to 100 bases per second. Molecules are fast; people and PCR instruments are slow. By increasing the primer and polymerase concentrations, good PCR can be performed in less than 15 seconds.

PCR products can be as small as 40 bp to over 40 kilobases (kb). To amplify long products, mixtures of polymerases that include some 3′-exonuclease activity to edit polymerase mis-incorporations that lead to base mismatches are usually used. Instead of separate annealing and extension temperatures, both processes can be carried out at the same temperature, resulting in two-temperature, instead of three-temperature, cycling.

Polymerase Chain Reaction Optimization

In addition to temperature-cycling conditions, the specificity of PCR depends on the choice of primers and the Mg^{2+} concentration. Mg^{2+} is a polymerase cofactor and stabilizes the DNA double helix. Low concentrations of Mg^{2+} favor specificity, but high concentrations favor sensitivity.

PCR sensitivity and specificity can be compromised by the formation of unintended low molecular weight artifacts. This process is initiated before PCR, when the primers, template, and polymerase are all together at temperatures below the annealing temperature of the PCR primers. This can be

prevented by limiting the activity of polymerase until the temperature is increased (thus the strategy is often collectively called a *hot start*). One method of hot start involves the use of antibody (or an aptamer) to bind and inactivate the polymerase at room temperature. The binding agent is released upon heating, allowing polymerase extension. Another method uses wax or paraffin to create a physical barrier between the essential components in the reaction. Finally, the polymerase, primers, or dNTPs can be chemically blocked so that extension cannot occur until activated by heat, which usually requires an initial denaturation period of 2 to 20 minutes.

Polymerase Chain Reaction Contamination, Inhibition, and Controls

Because PCR typically produces about 1 trillion copies of the target, a small amount of previously amplified product in a new reaction will result in false-positive results. Thus, minute aerosol droplets contain enough PCR product for robust amplification. PCR products can contaminate: (1) reagents, (2) pipettes, and (3) glassware. Simple laboratory precautions can minimize contamination by using (1) physically separated areas for pre- and postamplification steps, (2) positive-displacement or barrier filter pipettes to minimize aerosol contamination, and (3) prior preparation and storage of individual or combined ("master mix") reaction components in small aliquots. The most effective way of preventing contamination is to contain the product within a closed tube as in real-time PCR. Even with contamination precautions, a negative control or blank reaction (including all reactants except target DNA) is one of the most important controls for PCR.

One advantage of PCR is that it does not require highly purified NA to achieve a successful amplification. In practice, however, clinical samples may contain unpredictable amounts of impurities that can inhibit polymerase activity. Consequently, to ensure reliable amplification in clinical analyses, some form of NA purification is typically used. The diverse nature of PCR inhibitors within clinical specimens requires demonstration that the sample (or preparation of NA purified from it) will allow amplification. Although such confirmation is automatic with genotyping assays, detection and quantification reactions typically use a control NA sequence different from the target, which is added to the sample (or to NA extracted from the sample). Failure to amplify this process control indicates that further purification of the sample is required to remove inhibitors.

Multiplex Polymerase Chain Reaction and Nested Polymerase Chain Reaction

Multiplex PCR refers to amplifying more than one target simultaneously in the same solution. This can be done by using consensus primers that bind to and amplify more than one bacterial species, for example, ribosomal sequences with high variation internal to the primers, but flanking identical sequence that can be used for consensus primer binding. Usually, however, multiple primer sets are designed for multiple targets and are amplified in the same solution at the same time. Although potential primer interactions increase exponentially with multiplex PCR, it works surprising well most of the time as long as the number of PCR cycles is kept low. If lower primer concentrations are used, the annealing time may need to be increased to maintain efficient annealing. Multiplex PCR can be called "preamplification" if it is followed by an additional amplification reaction.

If PCR is performed with a pair of outer primers and then that PCR product is amplified yet again with a set of inner primers, it is referred to as *nested PCR*. Typically, the first round PCR product is diluted 1000 to 1 million times before the inner (nested) primers are added. The advantage of nested PCR is that both sensitivity and specificity increase. The disadvantage is increased potential for contamination, particularly if the first round of PCR products is handled manually for dilution and transfer. Multiplex PCR followed by nested PCR can be used in a closed-tube system for multiplex detection of infectious agents.

Asymmetric Polymerase Chain Reaction

Conventional PCR uses primers that are present in equal amounts, thereby ensuring that most of the DNA products are double stranded. **Asymmetric PCR** uses different concentrations of the two primers to generate more of one strand than of the other. For instance, the use of one primer at 0.5 μmol/L and the other at 0.005 μmol/L produces mostly ssDNA. However, yield of the product may be low with this technique. With less extreme ratios (e.g., one primer at 0.5 μmol/L and the other at 0.2 μmol/L), the yield is mostly preserved, with one strand produced in enough excess to make it readily available for probe hybridization.

Selective Amplification of Sequence Variants by Polymerase Chain Reaction

Allele-specific PCR enables preferential amplification of one genetic allele over another by placing the 3′-end of one primer at the polymorphic site. The variant that is better matched to the primer extends more readily than the other allele. This strategy can be used to distinguish a gene from its pseudogenes and for genotyping **single nucleotide variants (SNVs)**. Allele-specific PCR also can be used to determine haplotypes. In another approach, the minor Tm differences created when single base changes occur on one allele, anywhere along the sequence, can be exploited for preferential amplification of a mutant allele versus a wild-type allele. This strategy is combined with a lower denaturation temperature during PCR and can be used to identify low levels of mutant alleles.

Detection Limits of Polymerase Chain Reaction

When PCR is performed under optimal conditions, a single copy of the target can be detected. In practice, however, the statistical probability of distributing at least a single copy from a dilute template solution into the PCR must be considered. The Poisson distribution indicates that if, on average, one target copy is present per tube, 37% of the tubes will have no target, 37% will have one target, and the remainder will have more than one target. If there is an average of two

copies per tube, approximately 14% of the tubes will have no template and will be false negatives. About three copies on average are necessary for 95% of the tubes to include at least one copy. Therefore, the limit of detection (95% probability) of any single PCR cannot be lower than three copies per reaction. About five copies on average are necessary for 99% of the tubes to include at least one copy. This limitation of low copy analysis holds true for any amplification technique. However, digital PCR analyzes many reactions in parallel and can quantify less than one copy (on average) per reaction.

Digital Polymerase Chain Reaction

Conventional PCR averages the amplification results of many individual template molecules. Digital or single-molecule PCR is a technique that uses a dilute solution of template distributed across many reaction compartments or "partitions." Each partition either has or does not have PCR template molecules to amplify. After PCR, the partitions are scored as either positive (one or more initial templates) or negative (no initial template) for a digital readout. Thousands or even millions of partitions are typically scored. The partitions may be aqueous PCR droplets in oil or formed on chips by microfluidics.

Digital PCR can identify and quantify rare sequence variants, precisely determine copy number changes, establish the concentration of PCR standards, and determine the haplotype of variants that are on the same PCR product. When properly performed, digital PCR does not require the standard curves routinely used in quantitative PCR. Digital PCR is less prone to background DNA competition because competing DNA is divided between positive and negative partitions. Furthermore, common PCR inhibitors may not be as apparent in digital PCR because a positive threshold is reached even under conditions of moderate inhibition that would affect bulk quantitative PCR. Single-molecule amplification and digital analysis also has been reported for other amplification methods **including loop-mediated isothermal amplification (LAMP)** and recombinase polymerase amplification (RPA).

Additional Target Amplification Methods

In addition to PCR, many other methods of target amplification have been developed; some have found clinical use and are described in the next sections.

Transcription-Based Amplification Methods

Transcription-mediated amplification (TMA) is modeled after the replication of retroviruses (Fig. 48.5). As in PCR, amplification is exponential with completion in less than 1 hour. Unlike PCR, temperature cycling is not required (except for an initial heat denaturation if a DNA template is used). TMA is particularly advantageous when the target is RNA (e.g., human immunodeficiency virus [HIV] and hepatitis C virus [HCV] in blood bank NA testing).

Loop-Mediated Amplification Methods

Instead of producing products of a defined length, LAMP produces a wide range of different DNA structures with branches and loops. Starting with 6 primers, the chain reaction includes both extension of free 3′-ends and additional priming from inner primers to the exposed single strands in the loops. The amplification results in a mixture of products with ever more loops and branching structures of increasing complexity.

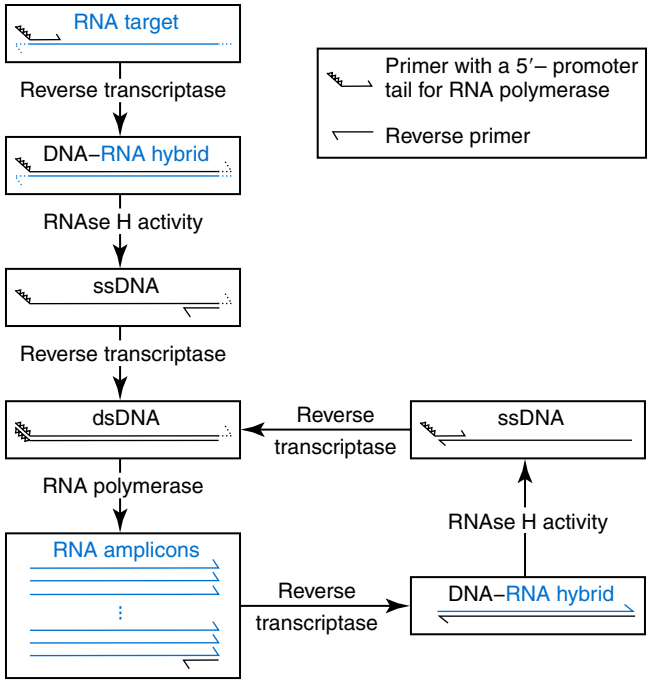


Fig. 48.5 Schematic diagram of transcription-mediated amplification. Starting with a single-stranded RNA target, a primer with an RNA polymerase promoter on its 5′-end is extended by reverse transcriptase to form a DNA–RNA hybrid. The reverse transcriptase also has RNAse H activity that subsequently degrades the RNA strand to leave single-stranded DNA (ssDNA). A second primer then binds to the ssDNA, and extension forms double-stranded DNA (dsDNA) with the attached RNA polymerase promoter. RNA polymerase then makes 100 to 1000 copies of RNA, some of which are again primed by the second primer. Repeated cycles of reverse transcription, DNA–RNA hybrid degradation by RNAse H activity, dsDNA formation by reverse transcriptase, and further transcription by RNA polymerase exponentially produce ssRNA amplicons. Single-stranded targets are amplified isothermally, and double-stranded targets are first denatured to single strands.

Strand Displacement Amplification

Strand displacement amplification (SDA) requires four primers: two outer displacement primers and two inner primers. Exponential amplification occurs at 37°C by repetitive strand synthesis and displacement catalyzed by a restriction enzyme.

Variants of Polymerase Chain Reaction That Do Not Require Heat Denaturation

Variants of PCR have been developed that replace the need for heat denaturation with enzymatic separation of the double helix. These methods do not require thermal cycling and better reflect the normal DNA replication process, although the end products are the same as in the PCR. For example, helicase-dependent amplification (HDA) uses the unwinding

enzyme, helicase, to separate the double helix into single stands. Another technique, RPA, uses a recombinase to scan dsDNA for priming sites, causing strand exchange to anneal the primers and single-stranded binding proteins to stabilize the loop structure long enough for strand displacement primer extension. Two opposing primers exponentially replicate a short DNA fragment as in PCR, but the reaction is performed at 37°C without temperature cycling.

Whole-Genome and Whole-Transcriptome Amplification

Instead of specific amplification of one target to improve sensitivity, methods that amplify all genomic DNA or mRNAs are useful when the target is in short supply. For example, *multiple-displacement amplification* uses exonuclease-resistant random hexamers and a highly processive polymerase to amplify DNA nonspecifically. Initial DNA denaturation is not necessary, and the reaction proceeds isothermally. Similarly, messenger RNA can be generically amplified with a poly(T) primer modified with an RNA polymerase promoter. After reverse transcription, second-strand DNA synthesis and transcription, antisense RNA is produced. Both whole genome and antisense RNA amplification are also useful as NA purification methods before amplification or detection.

Signal Amplification: Branched-Chain DNA

Instead of increasing the concentration of target, signal amplification techniques use NAs to magnify the detection signal. The branched-chain DNA (bDNA) method is one of these techniques in common use. The bDNA approach hybridizes the target NA to multiple capture probes affixed to a microtiter well. This is followed by hybridization to a series of "extender," "preamplifier," and "amplifier" probes. The final, highly branched amplifier probe includes multiple copies of signal-generating enzymes that act on a chemiluminescent substrate to produce light.

Rolling Circle Amplification

If a primer is annealed to a closed circle of DNA in the presence of a processive, displacing polymerase, the complement of the circle will be synthesized over and over again with displacement of the tandem repeats. If two primers are used in opposite orientation, progressively more complex branches will be formed in an exponential reaction. The rolling circle can be formed by ligation of the two ends of a linear probe on template DNA. Ligation may happen directly, after polymerization through a gap, or after annealing of an additional, allele-specific oligonucleotide.

Endpoint Quantification in Amplification Assays

Molecular diagnostic assays may be qualitative (detect the presence or absence of a target or identify a genotype) or quantitative (quantify the original concentration of a target sequence in the sample). When amplification is part of the assay, many variables need to be controlled carefully for accurate and precise quantification. Variation in extraction efficiency, the presence of enzyme inhibitors, lot-to-lot variation in enzyme and reagent performance, and day-to-day variation in reaction and detection conditions need to be addressed by quantitative methods. With reverse transcriptase assays, even more variables (reverse transcription efficiency and choice of reference genes) need to be considered.

DETECTION TECHNIQUES

Molecular diagnostics uses both generic and specific methods of NA detection. NAs can be quantified in bulk by optical techniques of absorbance and **fluorescence**. Specific methods of detection and quantification usually use sequence-specific primers or probes with fluorescent or electronic detection. Specificity in NA assays almost always comes from the hybridization of two complementary NA strands.

Fluorescent Labels

Advances in oligonucleotide synthesis and fluorescence detection have made fluorescently labeled probes the preferred reporter for NA analysis. Many fluorescent labels are now available, allowing color multiplexing for applications such as DNA sequencing, fragment length analysis, DNA arrays, and real-time PCR as reviewed later in this chapter. Techniques such as fluorescence polarization, fluorescence resonance energy transfer (FRET), and fluorescence quenching can provide additional detection specificity (see Chapter 9). Molecular rotation primarily depends on the size of the molecule, so binding of a small probe onto a large target results in a polarization increase that can be measured. FRET techniques depend on the distance between two spectrally distinct fluorescent labels. Two labeled probes may be brought closer together by adjacent hybridization. Alternatively, two labels on the same probe may end up farther apart by hydrolysis or hybridization. Fluorescence quenching or augmentation does not always require FRET. For example, fluorescence may change merely by hybridization of a fluorescent oligonucleotide to its target, or quenching moieties can be directly incorporated into probes.

Electrochemical Detection

Electrochemical detection of NAs recognizes the DNA duplex by conductivity or capacitance. Usually, PCR amplification is performed before detection, so that many molecules are available and a bulk signal is generated to increase sensitivity. Electronic detection is also used in one type of MPS, in which single nucleotide extension (SNE) from a clonally covered DNA bead produces a change in pH that is detected by electronic sensors. Direct sequencing of single molecules is also possible by detecting current changes that occur when a single strand of DNA passes through a nanopore.

DISCRIMINATION TECHNIQUES

NA discrimination techniques are divided into three categories: (1) electrophoretic methods that physically separate NAs based on molecular size or shape; (2) methods that determine the size, base content, or sequence of NAs without electrophoresis, including high-performance liquid chromatography (HPLC), mass spectrometry, and **pyrosequencing**; and (3)

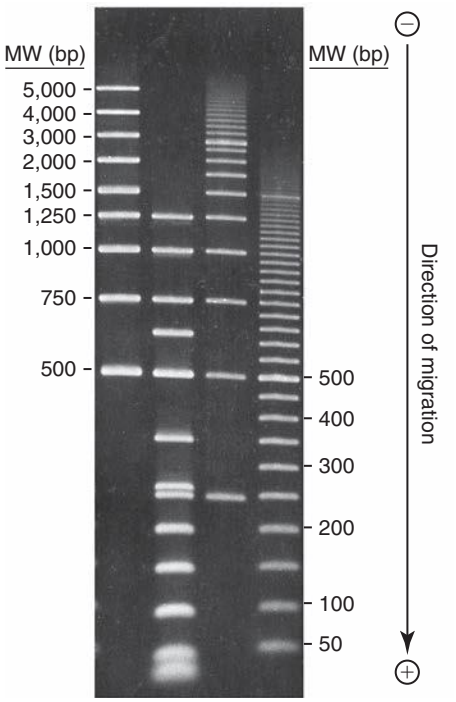


Fig. 48.6 A photograph of multiple DNA fragments after agarose gel electrophoresis (1% w/v, SeaKem LE agarose gel) showing the separation of double-stranded DNA molecules by size. *MW,* Molecular weight. (Photograph courtesy Lonza Rockland Inc., Rockland, ME.)

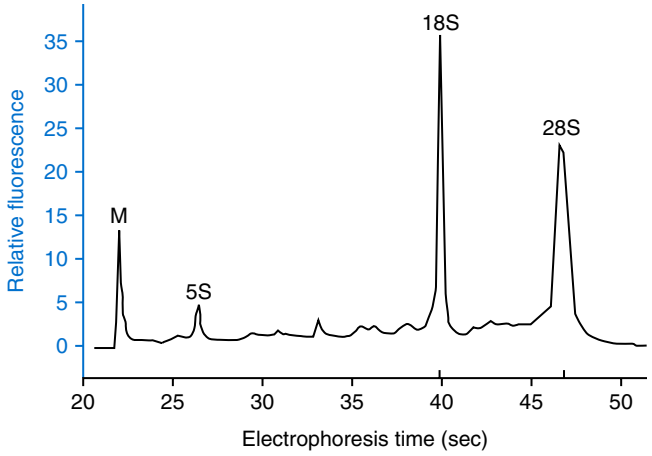


Fig. 48.7 Microelectrophoresis of human white blood cell (WBC) RNA. After isolation of WBCs and extraction of total RNA, samples were denatured, stained with a fluorescent dye, and applied to a commercial microelectrophoresis platform for assessment of RNA quality. Prominent 18S and 28S bands of ribosomal RNA suggest the RNA is largely intact. Also indicated are a reference marker (M) and the 5S ribosomal band. Note that electrophoresis was performed in less than 1 minute.

hybridization assays that identify specific NAs by the annealing or melting of complementary NAs. Some techniques use both electrophoresis and hybridization.

Electrophoresis

Both DNA and RNA are negatively charged and migrate toward a positively charged electrode when an electrical field is present within an appropriately buffered solution. Separation of different NAs occurs when mixtures are allowed to travel through a gel under an electrical field. Separation is based primarily on molecular size, with smaller molecules traveling faster through the gel than larger ones (Fig. 48.6). The shape or physical conformation of the molecule also affects migration, so that mismatches and hairpins can be detected. Separation based on shape can provide useful information, but it can also confuse size-based analysis.

RNA electrophoresis is commonly performed as a quality control check before transcript quantification or microarray expression analysis. Microfluidic chips with integrated microelectrophoresis channels are commercially available to rapidly assess RNA integrity by inspection of ribosomal RNA peaks (Fig. 48.7). Although specific transcripts are not detected by this method, only small amounts of starting RNA are needed.

Agarose and *polyacrylamide* are the two types of gel polymers commonly used in electrophoresis. Agarose gels can separate NA fragments as small as 20 bp to more than 10 mb (10,000 kb), including chromosomes of yeast, fungi, and parasites. However, the resolution of agarose is limited to about 2% to 5%. Polyacrylamide gels are suited for high-resolution separation (down to about 0.1% size differences) of short molecules (up to about 2 kb) and are the primary polymers for single-stranded NA separation, such as in dideoxy-termination sequencing. The optical clarity of polyacrylamide gels makes them ideal for visualizing fluorescent emission using laser-induced fluorescence detection. Table 48.2 lists common electrophoresis-based techniques.

TABLE 48.2 Commonly Used Electrophoresis-Based Techniques

Techniques Using Electrophoresis	Abbreviation	Primary Application
Polymerase chain reaction (or reverse transcriptase polymerase chain reaction) length	PCR RT-PCR	Detection
Polymerase chain reaction/ restriction fragment length polymorphism	PCR/RFLP	Detection
Dideoxy-termination sequencing (Sanger)		Sequencing
Single-nucleotide extension assay	SNE	Genotyping
Oligonucleotide ligation assay	OLA	Genotyping
Multiplex ligation-dependent probe amplification	MLPA	Quantification

Polymerase Chain Reaction Product Length

PCR product analysis by electrophoresis is frequently used to query the product size and specificity of PCR. PCR products are visualized by staining the gel with a fluorescent DNA-binding dye, such as ethidium bromide. In some cases, the presence of an amplification product is directly diagnostic (e.g., detection of sequences found only in a bacterium, virus, or fungus in a human sample). The specificity of the amplification reaction is verified by the known size of the fragment

and the lack of extraneous bands. Internal negative and positive controls are used to control for potential contamination and PCR inhibitors.

Small insertions, deletions, rearrangements, and changes in the number of repeated sequences also can be detected by monitoring PCR product length on gels. Length differences may be large and easily visualized with agarose gel electrophoresis, or they may be small enough to require a denaturing polyacrylamide matrix. Fluorescent primers may be incorporated into the product during PCR to simplify detection of fragment lengths. These techniques are commonly used in the diagnosis of inherited diseases and in identity assessment.

Restriction Fragment Length Polymorphism

DNA extracted from a cell is extremely long and is usually cut into shorter fragments before electrophoresis to enhance mobility. Restriction endonucleases cut dsDNA into fragments of reproducible size; the same enzyme produces the same fragments in different specimens if they contain the same DNA sequence. If an alteration in the DNA abolishes or creates a cleavage site recognized by the enzyme (or changes the spacing between two cleavage sites), different sized fragments will be produced. These changes in fragment lengths (or polymorphisms) that result from restriction digestion are called **restriction fragment length polymorphisms (RFLPs)**. However, restriction digestion of genomic DNA produces thousands of fragments. To be useful, specific fragments need to be visualized with probes.

Polymerase Chain Reaction/Restriction Fragment Length Polymorphism

Many sequence alterations (e.g., single base changes) do not affect the length of DNA. However, they can be amplified easily by PCR, and many can be detected as RFLPs after treatment with restriction enzymes. After PCR, the products are digested with one or more enzymes and analyzed by electrophoresis. For example, if a sample has a variant that disrupts an enzyme recognition site, it can be distinguished from a sample that does not have the variant. Such an assay will produce one uncut PCR fragment when the variant is present and two shorter fragments with normal DNA (Fig. 48.8). If the variant is present as a heterozygote (one normal and one variant copy of DNA), then one long and two shorter fragments will be observed. Usually it is possible to design the assay so that the fragments can be easily resolved by agarose gel electrophoresis. One variant of this method uses reverse-transcribed mRNA, which lacks the introns that would be present in the DNA. In this way, multiple exons can be analyzed in a single PCR.

Dideoxy-Termination Sequencing (Sanger Sequencing)

Dideoxy-termination sequencing of DNA is routinely performed in the clinical laboratory. RNA can also be sequenced by first converting RNA to DNA by reverse transcriptase. NA sequence is determined and compared with a reference sequence with an error rate of approximately 0.1% (1 misidentified base in 1000). Often the sequence is analyzed on both strands (sense and antisense) for even greater accuracy. Any deviation from the reference sequence is identified by using computer software. Base changes, including synonymous changes that do not alter the protein

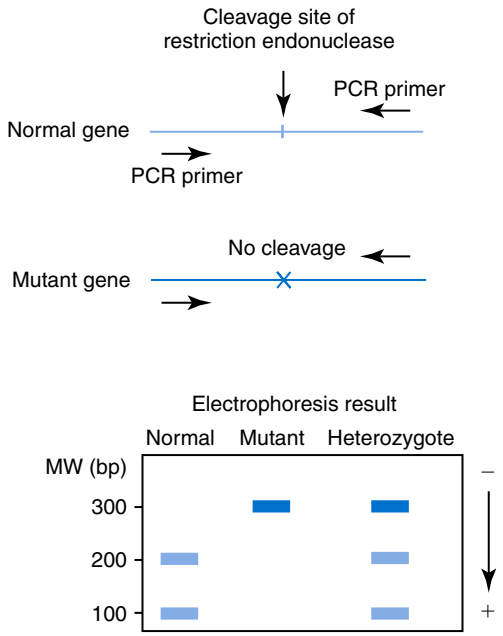


Fig. 48.8 An example of polymerase chain reaction *(PCR)* restriction fragment length polymorphism (PCR-RFLP). A DNA fragment amplified by PCR carries a site (a unique sequence of generally four or more bases) that is recognized and cleaved by a restriction endonuclease. If a mutation is present, this site is altered and is no longer recognized by the enzyme. Electrophoresis reveals that the fragment from a normal specimen was indeed cut by the enzyme, generating two fragments shorter than the original length, but the fragment from a homozygous mutant was not cut, and the original length of the PCR product is preserved. In a heterozygous mutant, both the original fragment and the shorter fragments are visible. *MW,* Molecular weight.

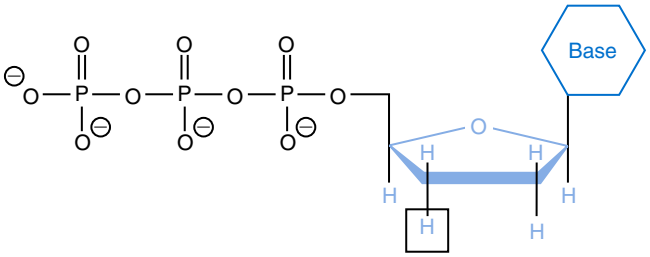


Fig. 48.9 A dideoxynucleotide. Notice the absence of the 3′-OH that is usually present in standard deoxynucleotides. This lack of a 3′-OH prevents polymerase extension because no phosphodiester bond can form. Incorporation of a dideoxynucleotide forces termination of polymerase extension.

sequence, as well as nonsynonymous changes resulting in altered amino acids, stop codons, and small insertions and deletions, are identified. However, larger deletions and rearrangements spanning the sequencing primer binding sites are not detected.

The most common sequencing strategy uses PCR in the first step to amplify the region of interest followed by a chain-termination reaction developed in the late 1970s. This reaction generates fragments that are terminated at various lengths by incorporation of one of the four dideoxynucleotide base analogs during extension from a sequencing primer. Dideoxynucleotides lack both 3′- and 2′-hydroxyl groups on the pentose ring (Fig. 48.9). Incorporation of dideoxynucleotides terminates chain extension. The most common method for generating these terminated fragments is *cycle*

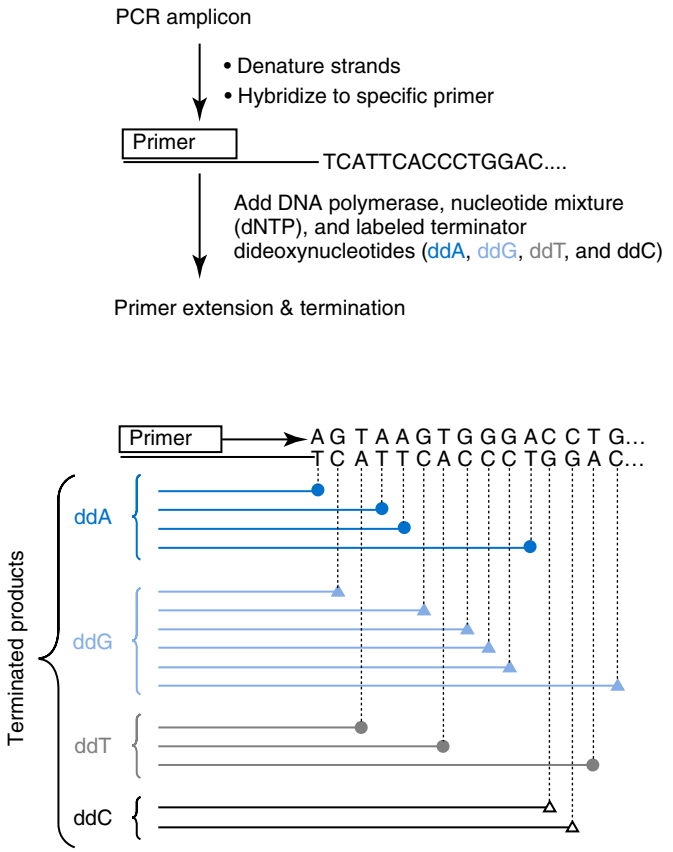


Fig. 48.10 The dideoxy-termination reaction for sequencing. A polymerase chain reaction *(PCR)* product is denatured and hybridized to a specific oligonucleotide primer. As the *DNA* polymerase extends the primer by incorporating bases (deoxynucleotides *[dNTPs]*) complementary to the template, it occasionally incorporates a terminator dideoxynucleotide base analog *(ddA, ddG, ddT, or ddC)* that stops further extension. The result is a mixture of extended products of varying lengths. Each terminator base may be labeled with one of four different fluorescent tags (shown as different symbols in the diagram).

sequencing, repeating the steps of annealing, chain extension-termination, and denaturation by temperature cycling, similar to PCR (Fig. 48.10). The fragments generated are tagged and terminated with a fluorescent dideoxynucleotide, separated by electrophoresis, and detected by fluorescence as the fragments travel past a detector (Fig. 48.11). About 600 to 800 bases can be resolved in less than 2 hours by capillary electrophoresis, and 96 or 384 samples can be run in parallel.

Single Nucleotide Extension

Also known as *single-base primer extension* or *minisequencing,* SNE assays involve the annealing of an oligonucleotide primer to a single-stranded PCR product at a location that is immediately adjacent to the site of the single base variant, followed by enzymatic extension of the primer in the presence of polymerase and dideoxynucleotide terminators. For example, each of the four terminators can be labeled with a unique fluorescent label so that it is possible to detect which base was incorporated. SNE assays can be multiplexed on capillary electrophoresis instruments by varying the lengths of the primers so that each SNV is resolved by size in one electrophoresis run. Many SNE detection methods are available other than electrophoresis, including (1) photometric detection on microtiter plates, (2) product capture detection systems on **DNA microarrays**, (3) bead hybridization assays detected by **flow cytometry**, (4) solution-based fluorescence polarization detection systems, and (5) mass spectrometry. SNE assays are particularly useful when the target of interest contains 5 to 50 disease-causing single base variants. SNE assays do not work well if variants are present in the primer-binding site. Nor are they usually designed to detect variants at a position other than immediately adjacent to the 3′-end of a primer.

Oligonucleotide Ligation

Another assay format frequently used for variant detection is the **oligonucleotide ligation assay (OLA)**. Two oligonucleotide probes are hybridized to adjacent sequences of amplified target DNA, with the variation site positioned at the end of one probe (Fig. 48.12). DNA ligase covalently joins the two probes only if both probes are perfectly hybridized to the target, including the polymorphic base. A probe matching the normal base and another probe matching the variant base are usually prepared. These two can be discriminated through differential electrophoretic mobility by varying the length of their 5′-tails. The probe hybridizing to both alleles (the *common probe*) provides the reporter molecule, usually a fluorescent label. Multiplexing is achieved by attaching different fluorescent labels to the common probes and by varying the length of the tails on the allele-specific probes. After ligation, probes for multiple variants are separated by denaturing capillary electrophoresis.

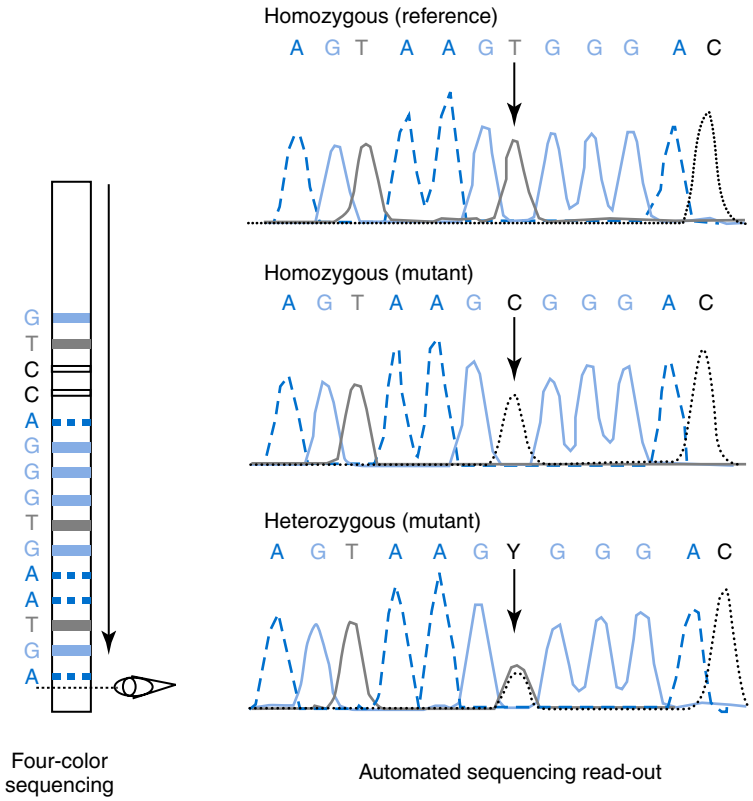


Fig. 48.11 Schematic of dideoxy-termination DNA sequencing. Extension products (see Fig. 48.10) are labeled with different color dyes for each of the four terminator reactions and are separated by electrophoresis. The four-color strategy allows automated endpoint fluorescence detection *(eye icon)*. The direction of fragment migration is from top to bottom. The sequence is read from left to right in the electropherogram generated by the sequencer. Examples of a reference sample (homozygous T at the polymorphic site), a mutant sample (homozygous C), and a heterozygous mutant sample (T and C, reported as Y) are shown.

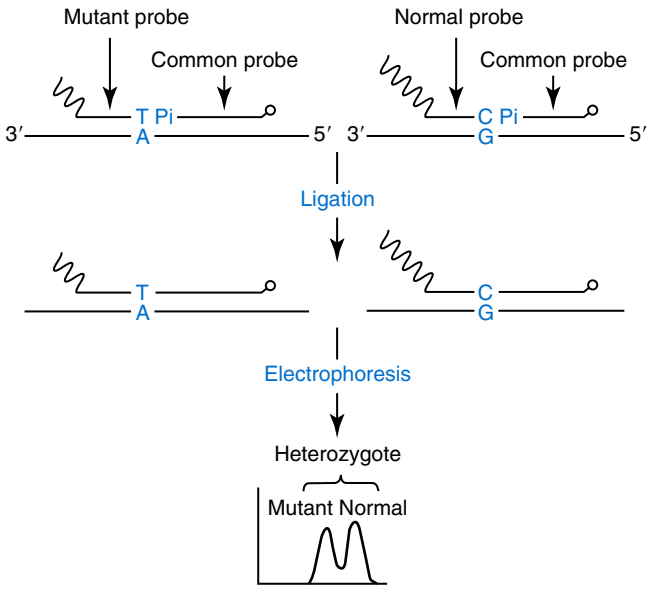


Fig. 48.12 Oligonucleotide ligation assay of a heterozygous single nucleotide variant. A mutant probe with a 3′-T *(upper left)* hybridizes to the mutant DNA with an opposing *A*, and a normal probe with a 3′-C *(upper right)* hybridizes to the normal DNA with an opposing *G*. The probes are also attached to mobility modifying tails of different lengths *(wavy lines)*. Hybridized next to these probes is the common probe with a 5′-phosphate (Pi) and a 3′-fluorescent tag. In the presence of ligase, adjacent probes are covalently joined to generate longer probes, each with a fluorescent tag and a mobility-modifying tail. Probes that are mismatched to the target at their 3′-ends may hybridize, but they are not ligated. Electrophoresis and endpoint, laser-induced detection of ligated probes reveal the different alleles by their different mobility.

Multiplex Ligation-Dependent Probe Amplification

Multiplex ligation-dependent probe amplification (MLPA) is a convenient method for relative quantification of up to 10 to 50 targets. This method is particularly useful in screening for deletions or duplications of exons, multiple exons, or whole genes. For each target, two probes are designed that hybridize adjacent to each other so that they can be ligated. The two probes have unique tails that do not hybridize to the target and that are the same between targets. After hybridization and ligation, the probes are amplified by PCR with a common primer pair (complementary to the tails). One of the primers is fluorescently labeled at its 5′-end. Because probes of different lengths are used, multiple PCR products of different sizes are produced and separated by capillary electrophoresis. Relative peak heights or areas are compared for relative quantification.

Alternatives to Electrophoresis

Electrophoresis is not the only method to determine the size, base content, and/or sequence of NAs. Some of these alternatives to electrophoresis are attractive in the clinical laboratory because of compatibility with automation and capacity for high throughput. Some examples include pyrosequencing, mass spectrometry, and HPLC.

Pyrosequencing is a sequencing-by-synthesis method that does not require dideoxy-termination or electrophoresis. One of the four dNTPs is added to the reaction, and if the base is complementary to the template strand, DNA polymerase catalyzes its incorporation. Each incorporation event is accompanied by release of a pyrophosphate (PPi), which is coupled through enzymatic reactions to the generation of visible light. The light produced is proportional to the number of nucleotides incorporated. As the process is repeated by adding one dNTP at a time, the complementary DNA strand is built, and the nucleotide sequence is determined.

Matrix-assisted laser-desorption ionization time-of-flight (MALDI-TOF) mass spectrometry can be used to genotype sequence variants. With mass spectrometry, no label is necessary because the alleles differ in mass. After isolation of genomic DNA, a specific DNA fragment, including the variant site, is amplified by PCR. After purification of the PCR product, an extension primer is hybridized directly or closely adjacent to the polymorphic site. Unlabeled deoxynucleotides are incorporated and terminated with a dideoxynucleotide, generating allele-specific diagnostic products of different mass. MALDI-TOF then measures the mass of the extension products. After the mass is measured, the genotype is determined (Fig. 48.13). Despite the complexity, automated systems processing 384 to 1536 samples at once are available.

HPLC is commonly used to separate and purify oligonucleotides. A variant of this technology is denaturing HPLC (dHPLC). dHPLC is run at a single elevated temperature to partially denature dsDNA. dHPLC reveals the presence of **heteroduplexes** as additional peaks that are shifted in retention compared with **homoduplexes**. DNA detection is most often performed by ultraviolet absorbance, although fluorescence or mass spectroscopy can also be used. Analysis by dHPLC is sequential (one at a time).

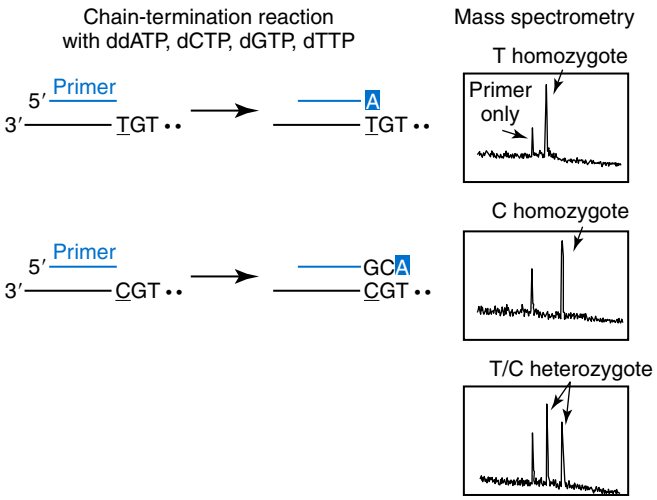


Fig. 48.13 Sequence polymorphism analysis by mass spectrometry. The underlined base is the polymorphic site in the template (T or C). The single-stranded template is primed and extended in the presence of three deoxynucleotides (dNTPs) and one dideoxy nucleotide triphosphate (ddNTP), producing fragments of different mass depending on the sequence. The boxed "A" in this example indicates the incorporated terminator adenine base. The mass of terminated products is precisely measured by matrix-assisted laser-desorption ionization time-of-flight mass spectrometry.

Hybridization Assays

All hybridization assays are based on the ability of single-stranded NAs to form specific double-stranded hybrids. The process requires (1) that probe and target NAs are mixed under conditions that allow specific complementary base pairing, and (2) that a method is available to detect any resulting double-stranded NAs. A *probe* indicates a NA whose identity is known, and the *target* or *sample* is a NA whose identity or abundance is revealed by hybridization. In some of the methods discussed here, hybridization occurs between a target in solution and a probe that is tethered to a solid surface. In *homogeneous* or *real-time* techniques, both the probes and the targets are in solution, and hybridization and detection occur without the washing steps. Some of the homogeneous methods also monitor the dissociation of hybridized duplexes under controlled heating, revealing the identities of the hybridized duplexes by *melting curve* signatures.

The noncovalent binding between two DNA strands is sequence dependent and reversible. Denaturing agents (e.g., high temperature, formamide, or extremes of pH) favor dissociation of the double-stranded molecule into two separate random coils (Fig. 48.14). On removal of the denaturant, single strands reform duplexes. Because temperature is the denaturant most easily manipulated, double- to single-strand transformation is referred to as *melting*, and the temperature at which half of the DNA is melted is the **melting temperature (Tm)** of the duplex. Duplexes with mismatched base pairs are less stable than those with a perfect sequence match and thus melt at lower temperatures. The reverse process, in which two complementary strands recombine to form a stable duplex molecule, is referred to as *annealing* or *hybridization*.

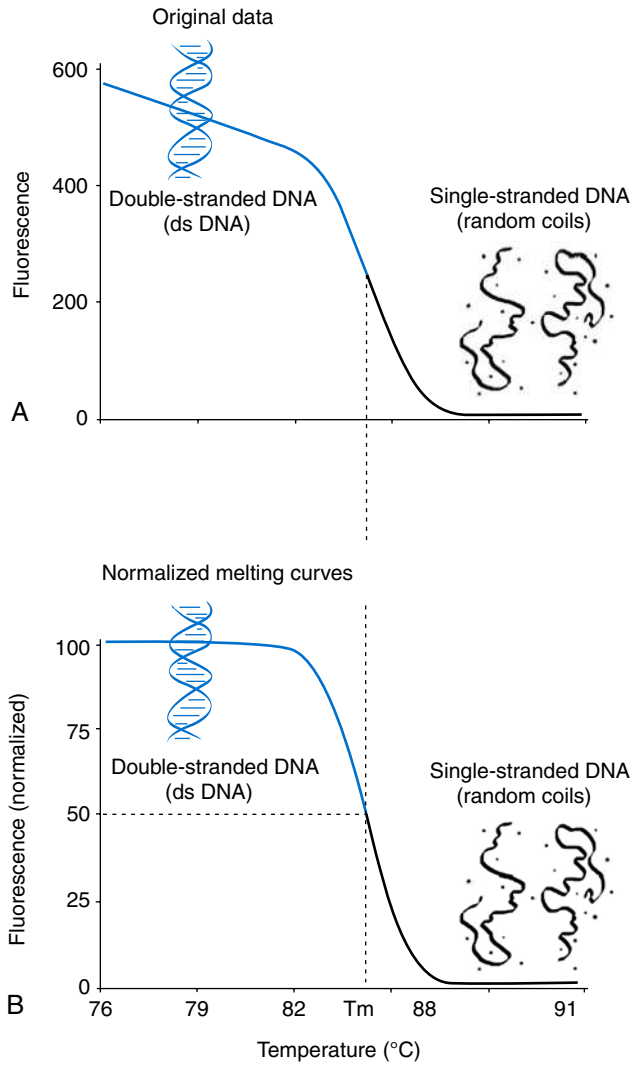


Fig. 48.14 Fluorescence melting curve of a polymerase chain reaction (PCR) product. PCR-amplified *DNA* was melted in the presence of a dye that fluoresces in the presence of double-stranded DNA, but not single-stranded DNA. (A) Fluorescence gradually decreases as the temperature is increased until a more rapid decrease occurs near the DNA melting temperature (Tm) of the PCR product. (B) The data are normalized between 0% and 100% after background removal to obtain constant fluorescence outside of the transition. (Modified with permission of the publisher from Wittwer, C. T., & Kusukawa, N. [2011]. Real-time PCR and melting analysis. In D. H. Persing, F. C. Tenover, Y.-W. Tang, J. Versalovic, T. J. White, & E. Unger [Eds.], *Molecular microbiology: diagnostic principles and practice* [2nd ed., pp. 63–82]. Washington, DC: ASM Press. Copyright 2011 ASM Press.)

Probes

Similar to antibodies in immunoassays, probes in NA hybridization assays can be labeled in many ways to detect hybridization. Probes may be cloned (recombinant), generated by PCR, or synthesized (oligonucleotides). They may be DNA, RNA, or NA analogs and single or double stranded. Oligonucleotide probes are the most common probes used today. These probes are usually 20 to 80 bases of single-stranded NA that are chemically synthesized. Automated, efficient, and accurate methods of synthesis continue to lower the cost of production. Sequence information is now routinely available in public databases. Probe sequences must be carefully chosen to minimize cross-hybridization with pseudogenes (eukaryotes) or related species (bacteria and viruses). The hybridization temperature of the probe should allow both favorable hybrid stability and discrimination against related sequences. Oligonucleotide probes are often prepared with covalent attachment of a reporter molecule (e.g., a fluorescent dye) or a reactive linker that allow them to be attached to solid supports. Probes used in homogeneous (real-time) PCR are usually oligonucleotides with a fluorescent label.

BOX 48.1 Hybridization Assays

Solid-Phase Hybridization
Dot blot and line probe assays
Arrays (microarrays and medium-density arrays)
Micro-bead assays
In situ hybridization
Massively parallel sequencing on beads or planar flow cells

Solution-Phase Hybridization
Real-time (or homogeneous) polymerase chain reaction
Melting analysis
Single-molecule sequencing
Many classical techniques

Examples of Hybridization Assays

Hybridization reactions can be divided into two broad categories: (1) *solid phase,* in which either probe or target is tethered to a solid support while the other is in solution, and (2) *solution phase,* in which both are in solution (Box 48.1). Several classical methods first hybridize in solution and then separate the bound from the unbound labeled probe. Exclusion chromatography and binding by hydroxyapatite, magnetic particles, or other capture methods allow selective measurement of the labeled probe–target hybrid.

Solid-Phase Hybridization

Solid-phase assays are useful because multiple samples can be processed together, which facilitates (1) control, (2) washing, and (3) separation procedures. However, hybridization on a solid support is less efficient than solution-hybridization, and the kinetics are slower and more difficult to predict. Both solid-phase and liquid-phase assays are used routinely in the clinical laboratory. Solid-phase assays include (1) dot blots, (2) line probes, (3) microspheres, and (4) microarrays.

Dot blot and line probe assays. Conventional hybridization assays on membranes are known as dot blot or line probe assays, depending on the geometry of the individual spots. The NAs are applied with suction, forming a shape that is either round (dot) or elongated (line or slot). After immobilization, the membrane is incubated with complementary NA at a constant temperature followed by one or more washes to discriminate matched from mismatched NA. The method allows multiple simultaneous hybridizations under identical conditions.

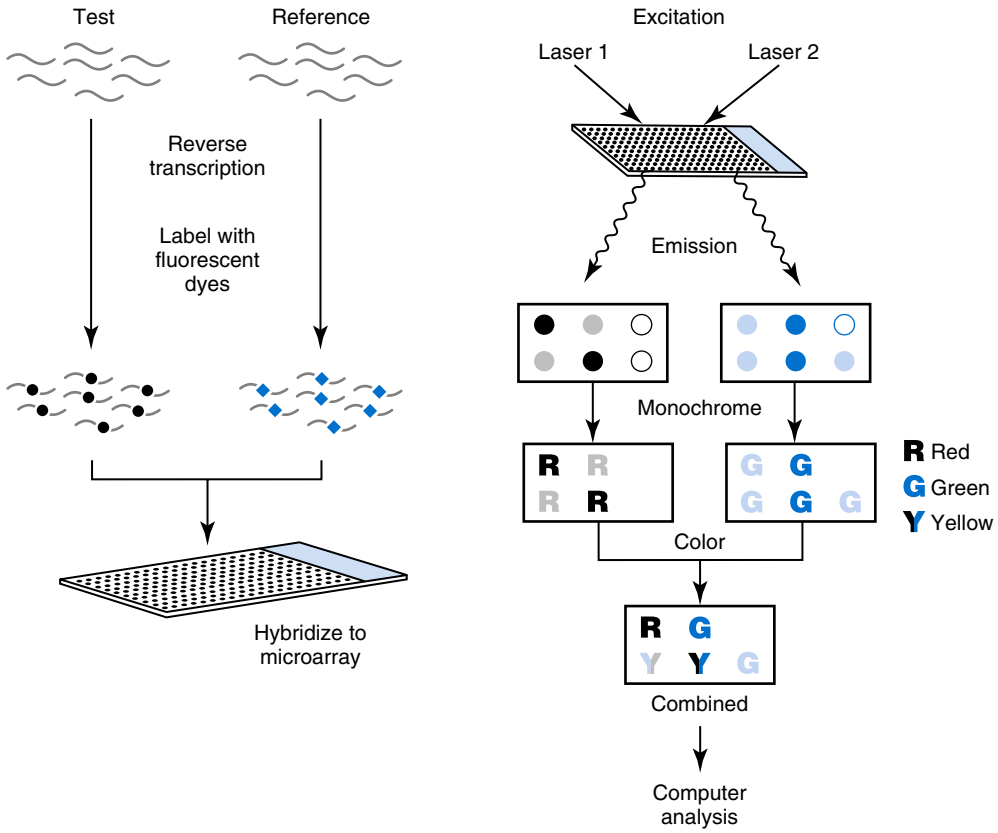


Fig. 48.15 A two-color microarray experiment. An array of DNA oligonucleotides with sequences of messenger RNA are affixed to a glass slide. Messenger RNAs in the test and reference specimen are converted into differentially labeled cDNA by reverse transcription and incorporation of two different fluorescent dyes. The two samples are hybridized together onto the array. The array is washed, and the image is captured twice, each time with a laser of a wavelength that excites one of the dyes but not the other. The monochromatic images are then converted to two colors (*green [G]* for the test sample and *red [R]* for the reference), and the images are combined. If the abundance of cDNA is the same in each of the two samples, then the composite spot will be shown as *yellow (Y)*. If one is in greater abundance, then that color will be preserved.

Medium-density arrays. Dot blot and line probe assays have largely been replaced by medium-density arrays that typically analyze 20 to 500 spots. The bound NAs are often oligonucleotides 20 to 80 bases in length that may be conventionally synthesized and spotted onto the chip or directly synthesized on site. Medium-density arrays are used for testing multiple mutations simultaneously in genetic diseases, oncology, and pharmacogenetics. These arrays do not need to be attached to a two-dimensional surface as long as their "address" can be decoded. For example, microspheres can be coded by fluorescence intensity in two different channels, and fluorescence in a third channel monitors hybridization. All channels can be read simultaneously using a flow cytometer.

Microarrays. Microarrays (also called DNA arrays, DNA chips, or Biochips) were introduced in the mid-1990s. Compared with medium-density arrays, spot sizes in microarrays are decreased (typically to <200 microns in diameter) such that one array contains thousands to millions of spots. This dimensional change requires (1) specialized detection equipment, (2) software, and (3) informatics to analyze the data. Because of their high density, microarrays have attracted intense interest among researchers who wish to monitor the whole genome for (1) SNVs, (2) gene expression, or (3) copy number variation.

Because SNVs represent the most common genetic difference among individuals, much effort has focused on correlating SNV genotypes to phenotype and disease association. SNV microarrays have been used in many genome-wide association studies (GWAS). Microarrays that analyze human SNVs ("SNV chips") provide the technology to genotype most known human SNVs in one experiment. Nearby SNV alleles tend to cluster together as haplotypes, so disease association by haplotype simplifies the analysis.

Gene expression microarrays quantify the relative amounts of different messenger RNAs in test and reference samples. An example of a two-color microarray for gene expression is shown in Fig. 48.15. Probes that hybridize to mRNA are usually directly synthesized on microarrays. Modern gene expression arrays have been used to measure the mRNA transcribed from all human genes in one experiment. Most arrays are used in marker discovery projects for selection of a smaller panel of expression targets that are then analyzed by other quantitative methods, such as real-time PCR, that provide greater precision and dynamic range.

Another important clinical application of microarrays is the genome-wide analysis of deletions and duplications, referred to as copy number variants (CNVs). CNV analysis using microarrays is replacing traditional cytogenetic

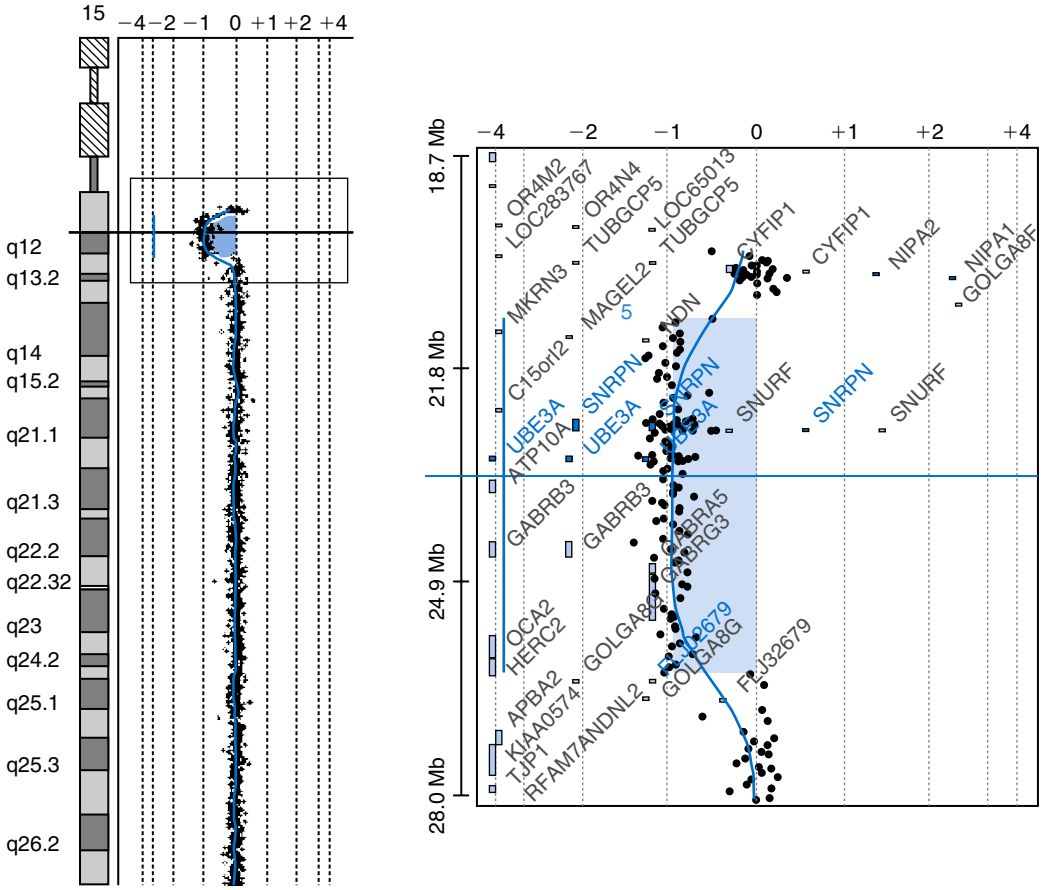


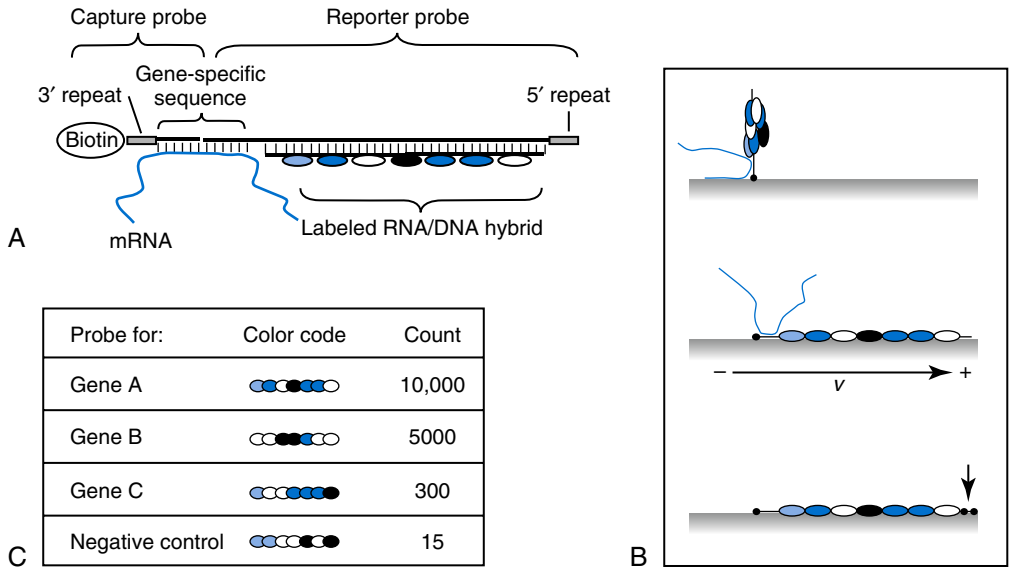
Fig. 48.16 Copy number variation identified with a comparative genomic hybridization array made from oligonucleotides. DNA from a subject is fragmented, labeled with Cy5, and hybridized onto a microarray, together with Cy3-labeled reference DNA. On the array are nearly 44,000 oligonucleotide probes, each about 60 bases long and tiled across the whole genome at an average spacing of 75 kb. Shown on the *left panel* are results of probes on chromosome 15 (all other chromosomes are analyzed in this assay but are not shown). Each *dot* represents a specific probe to which the subject's DNA hybridizes. Their positions (0, −1, +1, and so on) reflect the dosage of the subject's DNA relative to the reference DNA. A majority of the probes line up on "0," indicating no quantitative difference compared with the reference DNA. Probes in the 15q11 to 15q13 region, however, are on the "−1" line, indicating that the subject has a deletion of that region in one of the chromosomes. *kb*, Kilobases. (Courtesy Sarah South, PhD, ARUP Laboratories.)

chromosome analysis (**karyotyping**) and **fluorescence in situ hybridization (FISH)** analysis for the detection of genome-wide copy number alterations. Similar to gene expression arrays, many of the CNV arrays use two-color comparative hybridization to determine the gene dosage in a specimen compared with a normal reference genome (array **comparative genomic hybridization** [aCGH]). Arrays for CGH use oligonucleotide probes for very high resolution and data density. An example of CNV analysis using aCGH is shown in Fig. 48.16. Unlike aCGH, SNV arrays have the advantage of analyzing the specimen without the need to mix in a reference genome. SNV arrays also are able to detect copy number neutral changes caused by inversions or uniparental disomy that are not detected by aCGH methods.

Single-copy visualization. If an NA probe is labeled with many fluorescent molecules, it is possible to see a single copy of the NA target by fluorescent microscopy. One technique uses reporter probes that are labeled with a long string of multicolored fluorescent labels. Several tandem color segments are placed on the reporter probe with each segment consisting of about 100 fluorophores. The combination of different color segments uniquely identifies the target. The target NA is hybridized to capture probes that are (1) washed, (2) immobilized, (3) stretched, and (4) oriented on the surface of an optical slide. Each captured target is then identified by the color code of the reporter and is counted (Fig. 48.17). Although the sensitivity of this technique is not as high as that of real-time PCR, up to 150 reporter probes have been multiplexed in one reaction.

Solution-Phase Hybridization

Real-time PCR and melting analysis are considered "dynamic" hybridization assays in which the formation or dissociation of the probe–target duplex (or product duplex) is monitored in solution in real time. Data are collected throughout NA amplification rather than just at the endpoint. The technique uses fluorescent reporters and instruments to record hybridization during thermal cycling. The data obtained provide information on the (1) identity, (2) quantity, and (3) melting characteristics of the NA sample. Fluorescent dyes or probes are added to the PCR mixture



Probe for:	Color code	Count
Gene A		10,000
Gene B		5000
Gene C		300
Negative control		15

Fig. 48.17 Single copy visualization. A pair of probes (capture and reporter probes) hybridize to the messenger RNA (mRNA) target in solution through gene-specific probe sequences (A). The reporter probe has seven color segments, each segment is made of ~900 RNA bases that are labeled with about a hundred fluorophores of one color. The labeled portion of the probe is a DNA/RNA hybrid that can be observed as a ~3 nm fluorescent spot. Excess probes and unbound DNA are removed and the target complex is immobilized on a slide (B, *top*). An electrical current is applied, and the complex is stretched (B, *middle*) and immobilized in extended form (B, *bottom*). The color code of each probe is read through a microscope and counted (C).

before amplification. By measuring the fluorescence at each cycle of amplification, the amount of initial template present before PCR can be calculated.

Real-time PCR monitors the accumulation of double-stranded PCR products with the fluorescence signal recorded once each cycle (Fig. 48.18). If target DNA is present, fluorescence increases. How early during PCR one begins to see a signal depends on the initial amount of target DNA, and this provides a systematic method of quantification. Furthermore, when fluorescence is continuously monitored as the temperature is raised, a melting curve can be generated. Melting analysis can be used to verify the identity of the amplified product and to detect sequence variants.

Dyes and probes. Many different fluorescence generating systems are used in real-time PCR; some of the more common ones are shown in Fig. 48.19. Many methods use probes complementary to the target, and others rely on dsDNA binding dyes and depend on the specificity of the PCR primers. Some have the additional option of melting analysis to measure the melting temperature of the probe or product. Points to remember about fluorescent indicators are summarized in the box.

POINTS TO REMEMBER

Fluorescent Indicators

- May be dyes or probes
- Dyes are less expensive and less specific
- Melting analysis increases specificity
- May function by hybridization, hydrolysis, ligation, or displacement
- Single base variants can be detected by probes or high-resolution melting
- Enable real-time PCR in a closed system

Some dyes increase their fluorescence in the presence of dsDNA, which is produced during PCR (see Fig. 48.19A). dsDNA binding dyes are commonly used for real-time quantification, particularly in the research setting when the specificity of a probe and its added cost is not needed. They also allow melting analysis of the product at the end of PCR.

Fluorescent probes are also commonly used in real-time PCR. They hybridize to PCR products during amplification and change fluorescence by two possible mechanisms: (1) a covalent bond between two dyes is broken by hydrolysis, or (2) the fluorescence change follows reversible hybridization of the probe to the target. Following this distinction, when an irreversible covalent bond is involved, the probes are called *hydrolysis probes.* When probes reversibly change fluorescence on duplex formation, they are called *hybridization probes.* One difference between the two probe types is that melting analysis is characteristic of hybridization probes, but hydrolysis probes usually have a weak, if any, melting signature.

Hydrolysis probes (exonuclease probes) are synthesized with an absorbing molecule positioned to quench the fluorescence of another label. If the probe is hydrolyzed between the fluorophore and the quencher during PCR, fluorescence will increase. The most common implementation uses the 5′-exonuclease activity of some DNA polymerases to hydrolyze the probe and to dissociate the labels (see Fig. 48.19B). Hydrolysis probes generate fluorescence through changes in covalent bonds. The change in fluorescence signal is irreversible, and melting analysis of the hydrolyzed probe is seldom useful.

Hybridization probes change fluorescence upon hybridization, usually by FRET. Two interacting fluorophores are typically placed on adjacent probes (see Fig. 48.19C). Only one probe with one fluorophore may be necessary if fluorescence is quenched by deoxyguanosine residues. The

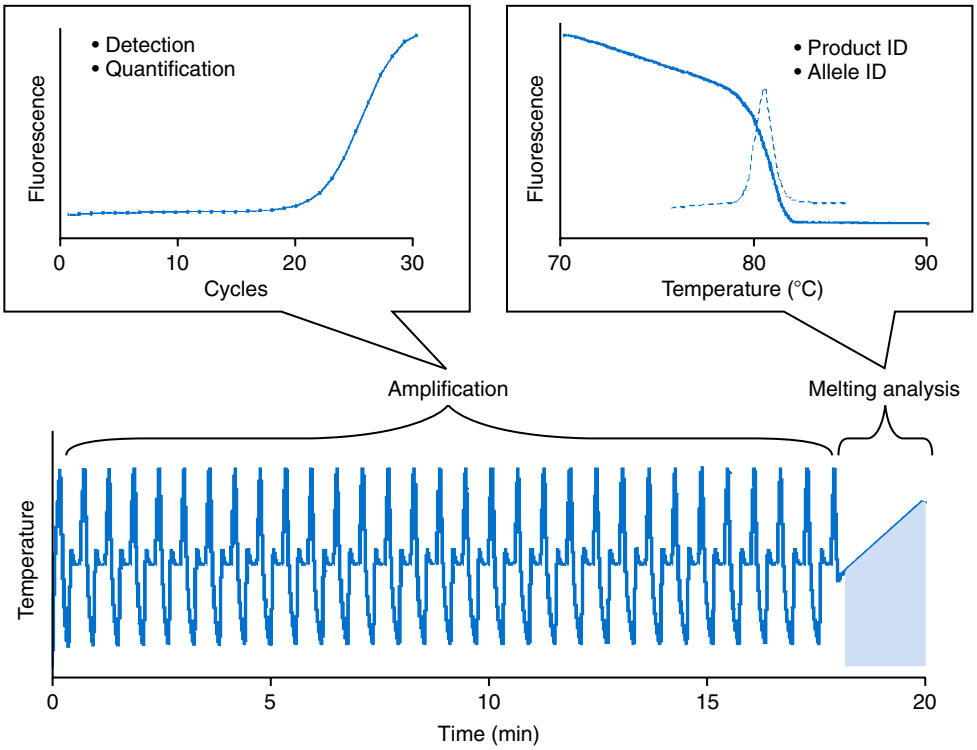


Fig. 48.18 Real-time monitoring during amplification and melting analysis. The *bottom panel* shows a typical rapid-cycle temperature profile that is followed by a temperature ramp for melting analysis. When fluorescence is monitored once each cycle during amplification, the presence or absence and the quantification of the specific target are obtained. When fluorescence is monitored continuously through the melting phase *(shaded area),* melting analysis can verify target identification and establishes genotype. (Modified with permission of the publisher from Wittwer, C. T., & Kusukawa, N. [2004]. Real-time PCR. In D. H. Persing, F. C. Tenover, J. Versalovic, Y.-W. Tang, E. R. Unger, D. A. Relman, et al [Eds.], *Molecular microbiology: diagnostic principles and practice* [pp. 71–84]. Washington, DC: ASM Press.)

fluorescence change that occurs with hybridization is reversible upon melting.

Hairpin probes (molecular beacons) typically have a fluorophore and a quencher at the 3′ and 5′-ends of a hairpin stem (see Fig. 48.19D). Fluorescence increases when the distance between the quencher and the reporter increases upon target hybridization. Hairpin probes of different colors can be combined with melting curve analysis for highly multiplexed assays.

Self-probing primers are modified at their 5′-end to include a hairpin probe with a fluorophore and quencher on opposite ends of the stem (see Fig. 48.19E). The hairpin loop is complementary to the primer extension product of the same strand that is generated during PCR. Through intramolecular hybridization, the quencher is separated from the fluorophore, and fluorescence increases. A blocker prevents copying the hairpin during PCR.

Highly quenched probes have a very efficient quencher at the 3′-end and a minor groove binder and fluorophore at the 5′-end (see Fig. 48.23F). Background fluorescence is very low because of the combined effect of the quencher and minor groove binders. Upon hybridization, fluorescence increases because the fluorophore and quencher are separated, a process that can be reversed by melting. The minor groove binders increase probe stability, allowing the use of smaller probes when sequence variation of the target is high.

Competitive displacement probes have a fluorophore and quencher on opposite strands, usually at 3′- and 5′- ends. Generation of target strands during PCR displaces one labeled strand, separating fluorophore and quencher, and generating fluorescence. In one design, partially double-stranded probes with one strand shorter than the other favor displacement (see Fig. 48.19G).

Detection and Quantification in Real-Time Polymerase Chain Reaction

When fluorescence is monitored once each cycle in the presence of a dye, the data closely follow the expected logistic shape discussed earlier (see Figs. 48.3 and 48.20, *top left*). However, with hydrolysis probes, fluorescence is cumulative and continues to increase even after the amount of product reaches a plateau (see Fig. 48.20, *top middle*). In contrast, reactions monitored with hybridization probes may show a decrease in fluorescence at a high cycle number (see Fig. 48.20, *top right*). Despite differences in the curve shape, all real-time systems follow the amount of product being produced during PCR, and this information is used for detection and quantification.

Detection. A fluorescent signal that increases during PCR and follows one of the expected curve shapes suggests that the specific target is present and was amplified. In contrast, a signal that stays at background even after 40 PCR cycles suggests

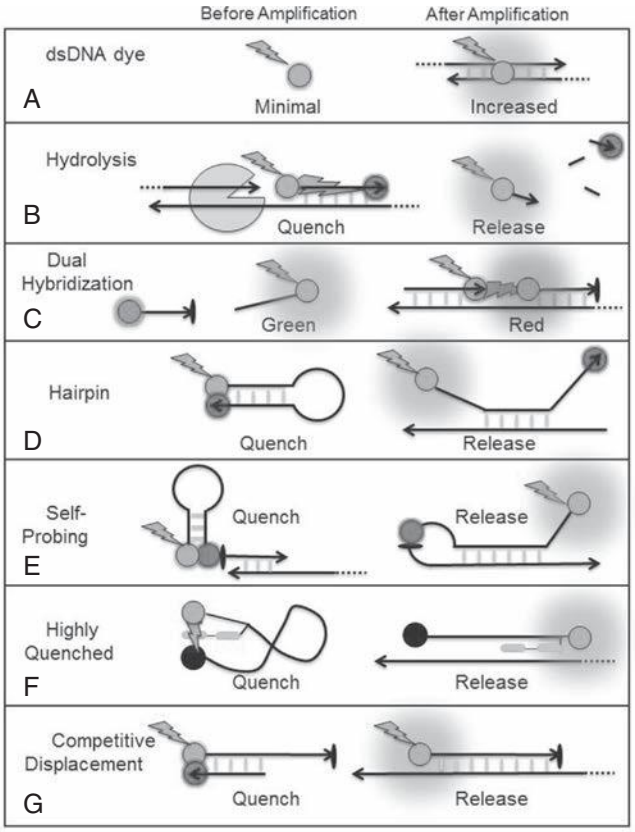


Fig. 48.19 Common probes and dyes for real-time polymerase chain reaction. (A) Double-stranded DNA dyes show a significant increase in fluorescence after DNA amplification. (B) Hydrolysis probes are cleaved between a fluorescent reporter and a quencher, resulting in increased fluorescence. (C) Dual hybridization probes. Fluorescence resonance energy transfer is illustrated between a donor and acceptor fluorophore. The *thin vertical oval* indicates termination of the 3′-end of the probe to prevent polymerase extension. (D) Hairpin probes are quenched in the native conformation but increase in fluorescence when hybridized. (E) Self-probing primers retain their native, quenched conformation until they are incorporated into a double-stranded product. (F) Highly quenched probes use a combination of minor groove binders and a very efficient quencher to limit background fluorescence. Upon hybridization, fluorescence increases. (G) Competitive displacement probes have reporters and quenchers on opposite strands. The quenching probe is displaced by the accumulating polymerase chain reaction product and fluorescence increases. There are several forms of this type of probe. (Modified with permission of the publisher from Nolte, F. S., & Wittwer, C. T. [2015]. Nucleic acid amplification methods overview. In D. H. Persing, F. C. Tenover, R. T. Hayden, M. Leven, M. B. Miller, F. S. Nolte, et al [Eds.], *Molecular microbiology: diagnostic principles and practice* [3rd ed.]. Washington, DC: ASM Press.)

that the target is absent and that no amplification occurred. Positive controls (to rule out inhibitory factors) and negative controls (to rule out product contamination and nonspecific signal generation) are necessary. If the fluorescent signal is reversible with hybridization, melting analysis can verify the expected Tm of the probe or product. Multiplex detection is possible with probes that are labeled with different colors or different melting temperatures. Examples in the clinical laboratory include probe multiplexing to detect the presence of more than one infectious organism, or to discriminate an internal control template from the target.

Quantification. Real-time PCR offers a convenient approach to quantification by monitoring the amount of product at each cycle. Digital PCR has some advantages over real-time PCR, including precision and rare-allele analysis. However, one of the advantages of real-time PCR is its large dynamic range. Fig. 48.21 shows an extended range of quantification standards in a typical real-time PCR. As the initial template concentration increases, the curves shift to earlier cycles. The extent of this shift depends on the PCR efficiency (Table 48.3). The cycle at which fluorescence rises correlates inversely with the log of the initial template concentration and is the quantification cycle or Cq. This "cycle" is actually a *virtual* cycle that includes a fractional component determined by interpolation, which can be calculated by several methods.

Accuracy and precision. The accuracy of real-time PCR quantification depends on the quality of the quantification standards used. Synthesized oligonucleotides, purified PCR products, plasmids, and genomic DNA can be used as calibrators. As previously mentioned, digital PCR is a great method to quantify PCR standards. Sometimes absolute quantification is not needed, and quantification relative to one or more reference genes is performed. In this case, selection of the reference genes and PCR efficiency are critical. The precision of quantitative real-time PCR depends on the initial number of template copies in the reaction. When the initial target concentration is low, imprecision is high. When this is the case, digital PCR may be a better choice.

Melting Analysis

When fluorescence is monitored continuously within each cycle of PCR, the hybridization characteristics of PCR products and probes can be observed. Continuous spirals of fluorescence versus temperature are produced (see Fig. 48.20, *bottom panels*). With dyes, the melting characteristics of the amplified DNA identify the product. With probes, melting occurs at a characteristic temperature that can be exploited to confirm target identity and to analyze sequence variants under the probe.

For routine testing in the clinical laboratory, a single melting curve is usually performed at the end of PCR instead of monitoring hybridization throughout the entire PCR process (see Fig. 48.18). Genotyping is best performed in the same tube by monitoring the melting of hybridized duplexes during controlled heating, producing a melting curve signature for the duplex. Such a signature monitors melting over a range of temperatures in contrast to the single-temperature analysis of conventional hybridization techniques, such as dot blots or microarrays. Real-time amplification and melting analysis make up a powerful combination that requires only temperature control and sampling of fluorescence.

High-Resolution Melting Analysis

An example of SNV genotyping by high resolution melting analysis is shown in Fig. 48.22. The PCR product is long enough that it melts in two domains because of unequal guanosine/cytosine content throughout the PCR product. Applications of **high-resolution melting analysis** include genotyping, mutation scanning, methylation analysis, sequence matching, and relative copy number assessment.

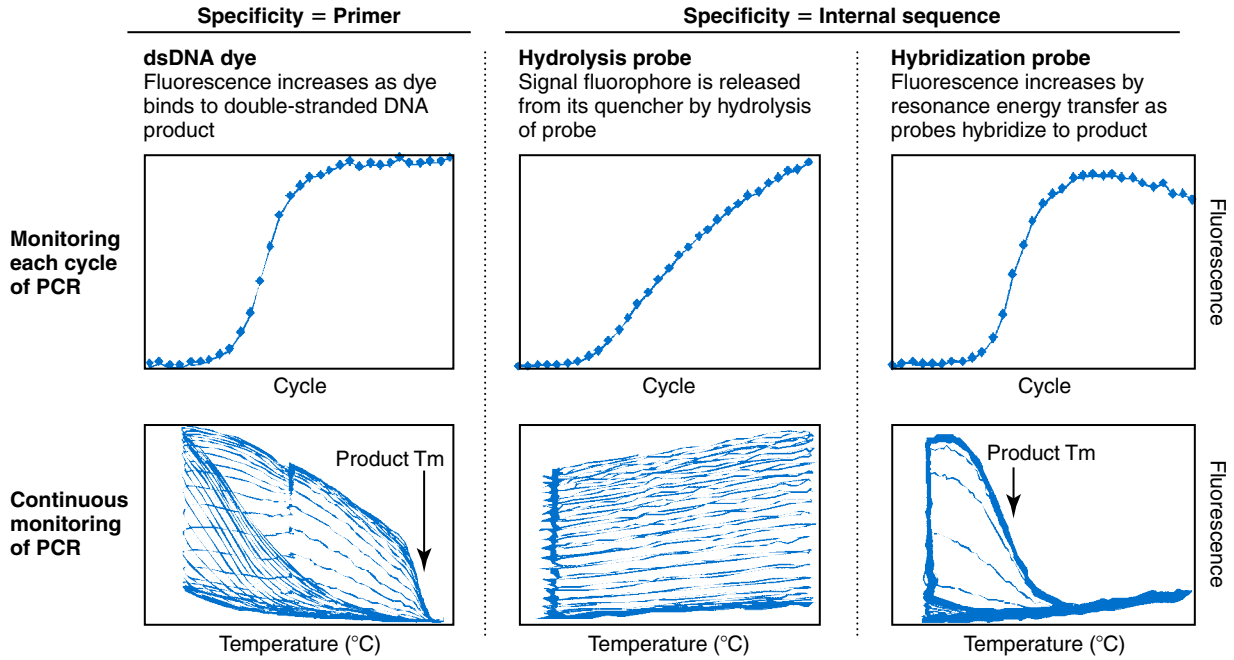


Fig. 48.20 Monitoring polymerase chain reaction *(PCR)* in real time. The *top row* shows data collected once each PCR cycle, and the *bottom row* shows data collected continuously (five times per second) during all PCR cycles. Three different reporter systems are shown. *dsDNA,* Double-stranded DNA; *Tm,* melting temperature. (Modified with permission of the publisher from Wittwer, C. T., & Kusukawa, N. [2004]. Real-time PCR. In D. H. Persing, F. C. Tenover, J. Versalovic, Y.-W. Tang, E. R. Unger, D. A. Relman, et al [Eds.], *Molecular microbiology: diagnostic principles and practice* [pp. 71–84]. Washington, DC: ASM Press.)

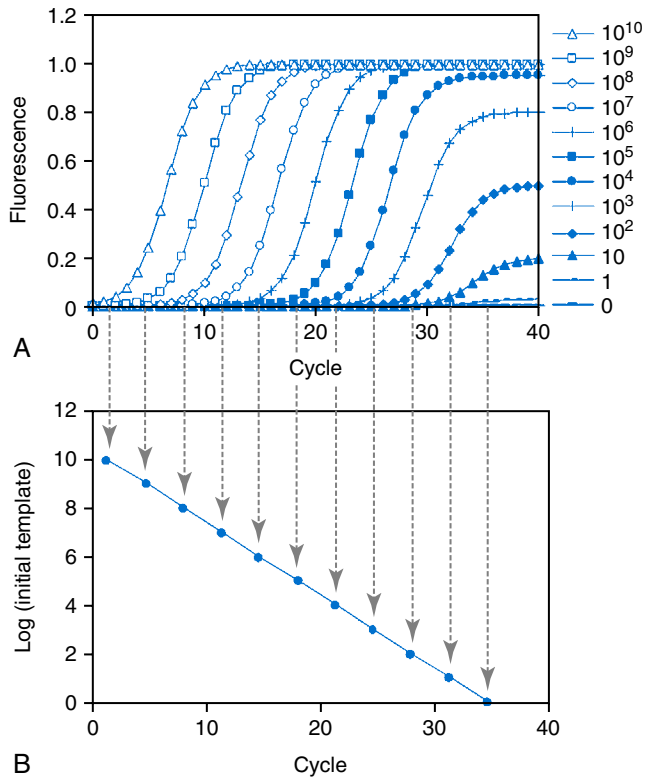


Fig. 48.21 Quantification by real-time polymerase chain reaction. Shown are typical real-time curves for amplification reactions of varying initial target concentrations (A) and the log of the initial concentration plotted against the quantification cycle (B). (Modified with permission of the publisher from Wittwer, C. T., & Kusukawa, N. [2004]. Real-time PCR. In D. H. Persing, F. C. Tenover, J. Versalovic, Y.-W. Tang, E. R. Unger, D. A. Relman, et al [Eds.], *Molecular microbiology: diagnostic principles and practice* [pp. 71–84]. Washington, DC: ASM Press.)

TABLE 48.3 Correlation Between Polymerase Chain Reaction Efficiency, the Slope of the Standard Curve, and the Percentage of Polymerase Chain Reaction Product After 30 Cycles

Slope of Standard Curve[a]	PCR Efficiency (%)	PCR Product After 30 Cycles (% Expected)
−3.32	100	100
−3.35	99	86
−3.38	97.5	69
−3.45	95	47
−3.59	90	22
−3.74	85	10
−3.92	80	4
−4.34	70	1
−4.90	60	0.1
−5.68	50	0.02

[a]Assuming the log (initial template) is plotted on the *x*-axis as the independent variable and the quantification cycle is plotted on the *y*-axis as the dependent variable, the slope of the calibration curve is as follows: Slope = ΔCycle/ΔLog [initial template]. Percent PCR efficiency is calculated as $(10^{-1/\text{Slope}} - 1) \times 100$.
PCR, Polymerase chain reaction.

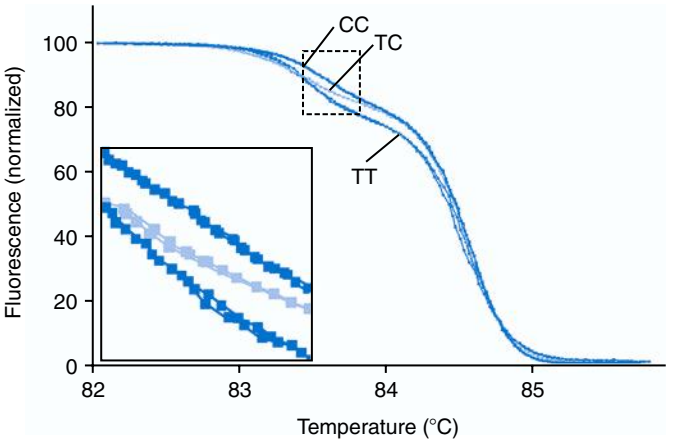


Fig. 48.22 A single base change in a 544-bp fragment detected by melting analysis. Shown are high-resolution melting curves of polymerase chain reaction products from the gene *HTR2A* carrying a single nucleotide variant. Results are shown for six individuals, two different individuals for each of the three genotypes: wild-type homozygote *(TT)*, mutant homozygote *(CC)*, and heterozygote *(TC)*. Two melting domains are present because of differing GC content. The single nucleotide variant was present in the lower melting domain. The *inset* magnifies a portion of the data, showing that all three genotypes can be discriminated. (Modified by permission of the publisher from Wittwer, C. T., Reed, G. H., Gundry, C. N., Vandersteen, J. G., Pryor, R. J. [2003]. High-resolution genotyping by amplicon melting analysis using LC Green. *Clinical Chemistry*, *49*, 853–860. Copyright AACC.)

TABLE 48.4 Characteristics of Massively Parallel Sequencing Methods

Method	Principle	Detection	Clonal	Run Time	Output Per Run	Read Length (bp)
Synthesis	Pyrophosphate release	Chemiluminescence	Emulsion PCR	10–23 h	40–700 mb	400–700
Synthesis	pH change	Electronic CMOS	Emulsion PCR	3–4 h	1.5–10 gb	125–400
Synthesis	Reversible terminator	Fluorescence	Bridge amplification	2.7–12 days	15–600 gb	200–600[a]
Ligation	Multiple ligation events	Fluorescence	Emulsion PCR	10 days	300 gb	110
Single molecule	Zero-mode waveguide	Fluorescence	No	2 days	5 gb	10 kb
Single molecule	Conductivity	Electronic	No	Minutes to days	Depends	5 kb

[a]Includes both paired end reads (sequencing from both ends).
bp, Base pairs; *CMOS,* complementary metal oxide semiconductor; *kb*, kilobases; *PCR,* polymerase chain reaction.

Massively Parallel Sequencing

Compared with dideoxy-termination sequencing, MPS generates up to 1 billion more bases of sequence data in one operation at 10,000 to 100,000 times lower cost per base. Several methods are available that continue to evolve. As the cost, turnaround time, and convenience of these methods continue to decrease, they will be used increasingly in the clinical laboratory. Characteristics of MPS methods are summarized in Table 48.4.

Sequencing From Clones

Clonal sequencing methods start with producing a shotgun (random) library of fragments that are typically 70 to 1000 bases in length, although some methods require 6- to 20-kb fragments. Fragmentation is usually physical or enzymatic. Physical methods include sonication, acoustic shearing, and hydrodynamic shearing. Enzymatic fragmentation can be from restriction endonucleases or nonspecific DNAses.

Adapter sequences are typically added to each end of the random fragments. The primary role of these **adapters** is to provide common priming sites for each fragment to initiate MPS reactions. One primer set amplifies a massive array (beads or planer flow cell) of library **inserts**. Adapters also facilitate initial capture of DNA fragments onto solid surfaces and spatially restrict clonal amplification products generated from the fragments onto beads or spots on an array surface. If multiplexing of different DNA samples is desired, a sequence "barcode" is added to identify which DNA sample the clone arose from. A typical library insert with adapters and a barcode is shown in Fig. 48.23A. The libraries are then partitioned according to size to select a band optimal for the downstream sequencing technology. Clonal amplification is usually performed by either emulsion PCR or bridge amplification.

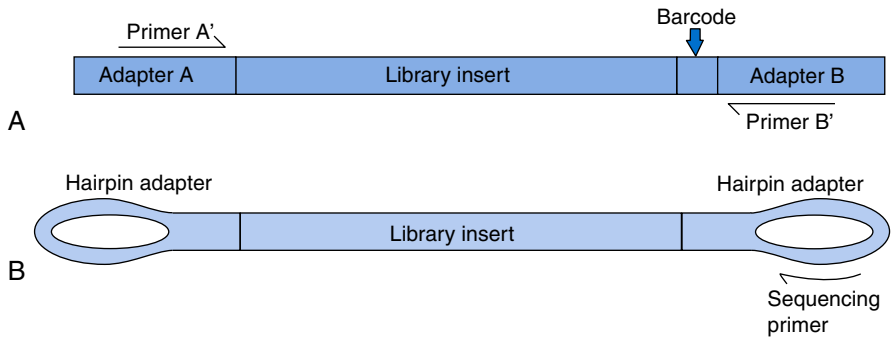


Fig. 48.23 Diagram of different library formats used in massively parallel sequencing. (A) Two adapters that include consensus polymerase chain reaction priming sites are ligated onto each end of the library inserts produced by fragmentation. If multiplexing different samples is desired, a barcode is added so that each read can be assigned to a specific sample. (B) A library insert is bounded by hairpin adapters that allow primer binding to the single-stranded loops on each end for rolling circle amplification.

Emulsion polymerase chain reaction. In emulsion PCR, one strand of a library element is captured on one bead and is clonally amplified inside a water-in-oil droplet, generating a bead covered with single-stranded PCR products (Fig. 48.24). The emulsion is formed by mixing beads (each covered with one primer), aqueous PCR components (including the other primer, polymerase, and dNTPs), and a mixture of oils under agitation to ideally form droplets that each contain only one bead and one library insert. The two primers are complementary to the adapters; one coats the bead surface, and one is free in solution. During emulsion PCR, all of the beads are amplified together in aqueous microdroplets dispersed in oil. After PCR and denaturation, millions of copies of identical single-stranded PCR products are on each bead with each bead carrying distinct, oriented inserts flanked by the adapters. The emulsion is then broken and, after the elimination of empty beads, is ready for sequencing.

Bridge amplification. Bridge amplification generates clusters of single-stranded PCR products tethered to the surface of a planar flow cell (Fig. 48.25). In contrast to the clonal bead generation of emulsion PCR, the amplification occurs on a flat surface. The primers, complementary to the adapters, are both attached to the surface, either randomly or in a fixed pattern. The library DNA is then denatured to form single strands that hybridize to the primers on the surface. After extension of the surface-bound primers, the original template strands are washed away under denaturing conditions. What follows is called bridge amplification, which is very similar to PCR except that both primers are bound to the surface, so that single strands bound to the surface must bend over to find their opposite primer, resulting in a double-stranded bridge after extension. Instead of denaturing with heat as in PCR, the flow cell is kept at 60°C, and a chemical denaturant is introduced to dissociate the two bridge strands, except now both strands are attached to the planer surface. When the flow cell is flushed with a polymerase and dNTPs under favorable extension conditions, both can find new primers to form additional bridges. The process repeats until about 1000 copies are formed. One of the bound primers can be designed to include a cleavable site (either chemical or enzymatic) so that one strand can be removed after denaturation, providing an amplified template for sequencing.

Sequencing by Synthesis

Sequencing by synthesis can be detected by (1) pyrophosphate release, (2) a pH decrease, or (3) fluorescence. Clonal amplification methods allow parallel observation of thousands to millions of strand extensions, greatly increasing the signal strength. However, the extensions must be controlled at each step because continuous strand extension would not remain synchronous between strands. This is achieved by immobilizing the clones into arrays so that reagents can be applied sequentially (no pun intended).

Pyrosequencing. Clonal beads are fitted into picoliter reaction chambers formed in etched individual fibers of a fiberoptic cable. Solutions of nucleotides are passed over chambers one at a time under conditions that favor extension. If there is a base match, the nucleotide will be incorporated, and pyrophosphate will be released, producing light. The light is captured by the individual light fibers and is detected on a sensitive camera. If more than one base occurs in a row (a homopolymer stretch), multiple bases of that nucleotide will be incorporated, and the signal will be proportionately higher.

Semiconductor sequencing. Similar to pyrosequencing detection, clonal beads are used as templates. However, the beads are arrayed on semiconductor sensors modified to detect pH changes. The chip does not detect light but a slight change in pH as a consequence of dNTP incorporation. By leveraging semiconductor technology, these pH chips have rapidly increased in performance by decreasing the size of the beads and sensor wells, and increasing the size of the chip. Run times are as short as 3 to 4 hours.

Reversible fluorescent terminators. After bridge amplification on a planar flow cell, four fluorescent nucleotides are passed through the flow cell under conditions that favor extension. Each nucleotide has a different fluorescent label so that each can be distinguished by color. Furthermore, the fluorescent terminators are reversible, which means that the blocked 3′-end can be regenerated by simple chemical means

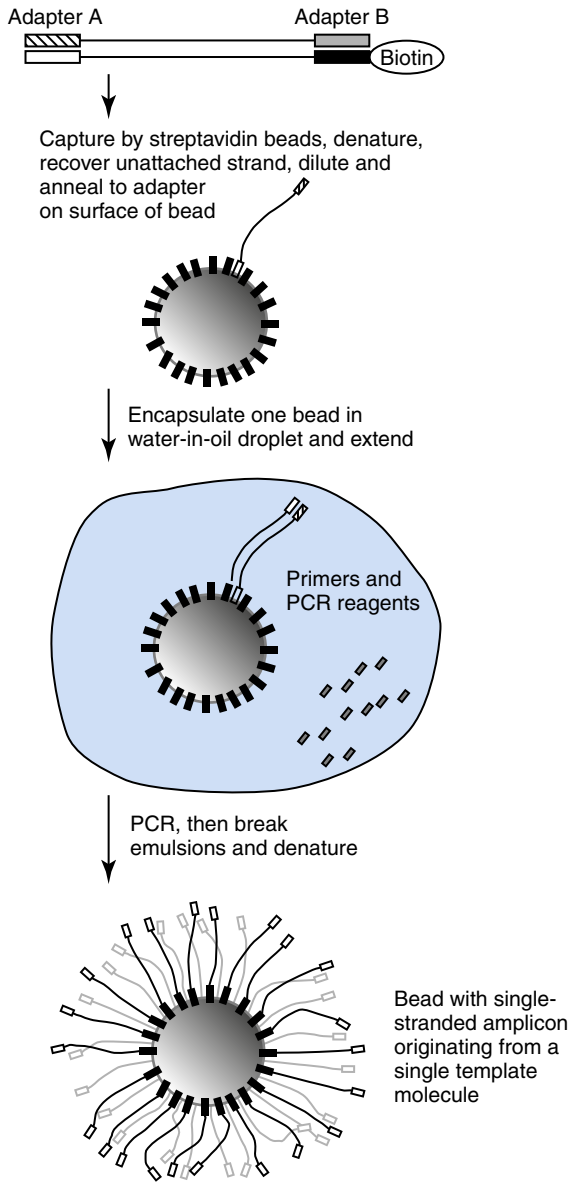


Fig. 48.24 Emulsion polymerase chain reaction *(PCR)*. Two adapters (*adapters A* and *B*) are randomly ligated to DNA fragments. Adapter B has a biotin on its 5′ end. Fragments with adapter B on one or both ends are captured by streptavidin beads, but fragments with only adapter A are washed away (not shown). Then the fragments are denatured, and the free strand with adapter A and adapter B at each end is collected (fragments with adapter B on both ends will not be released from the streptavidin bead). One molecule of the single-stranded template is then captured on a bead coated with adapter and is encapsulated inside a water-in-oil droplet that contains PCR reagents and primers. After PCR, the emulsion is broken, and the DNA is denatured. This generates a bead with a large number of clonal single strands tethered to it. The bead is then deposited into one of many wells on a fiber optic or onto a glass slide for sequence analysis (not shown).

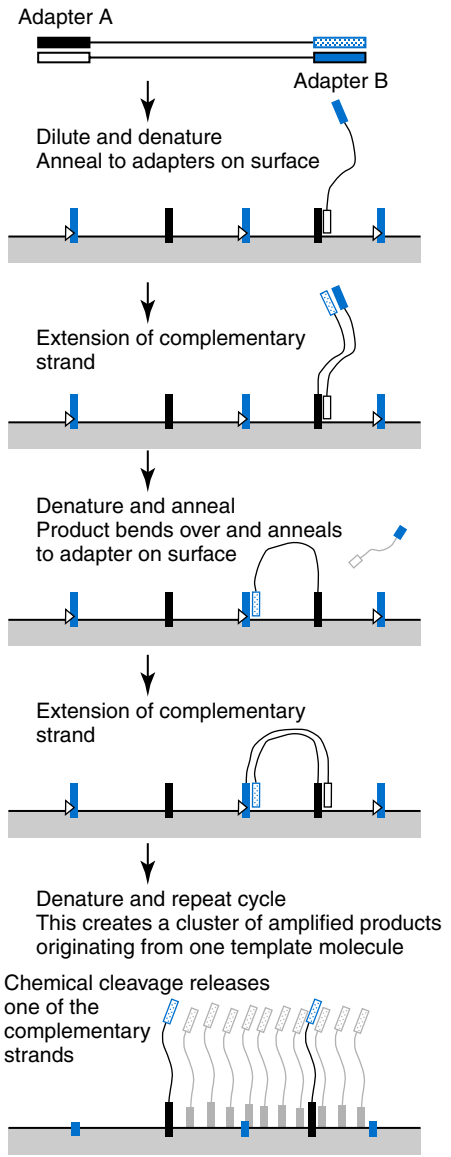


Fig. 48.25 Bridge amplification. Two adapters (*adapters A* and *B*) are ligated to a DNA template. After they are diluted and denatured into single strands, the template is captured onto a flow cell surface by annealing to one of the two surface-bound primers that share sequences with adapter A or B. The polymerase reagent introduced into the flow cell extends the primer and generates the complementary strand of the template. The denaturant (usually sodium hydroxide) is introduced to the flow cell to release the original template strand. The free end of the newly synthesized strand anneals to a nearby primer by bending over, and a second round of reagent addition catalyzes the synthesis of another complementary strand. By repeating many of these cycles, a clonal cluster that consists of about 1000 copies of single-stranded template tethered to the surface is generated. This cluster is still a mixture of both complementary strands. One of the strands is selectively eliminated by treatment with periodic acid that cleaves a diol linkage present in one of the surface-bound primers *(open triangle on red primer)*. The cluster now contains only one of the template strands and is ready for sequence analysis.

provided in the flow cell. Each cycle involves (1) adding polymerase and dNTPs as fluorescently labeled terminators under conditions that favor extension, (2) washing the flow cell, (3) imaging the fluorescence, (4) cleaving the fluorescent terminator, and (5) washing the flow cell. Output per run is up to 600 gb in 1 day.

Single-Molecule Sequencing

Single-molecule sequencing methods do not require template amplification. Base **reads** do not require synching with other clonal strands because there are none. Sensitive optical or electronic methods are required to detect base sequence in a single molecule. If long reads and high accuracy are achieved,

advantages include efficient sequence assembly, analyses of repetitive sequences, de novo sequencing, mapping of chromosomal rearrangements, and fusions. In contrast, MPS methods usually generate short sequence reads (30 to 700 bases long) that have to be aligned and analyzed to derive a consensus, then stitched together by software, and compared with reference genome sequences.

Real-time single molecule sequencing with fluorescent nucleotides. Library preparation is unique for this method because fragmentation is tuned to provide 10-kb inserts, and the adapters are designed as hairpins. The result is a double-stranded insert bounded by identical 27 base single-stranded loops on each end (see Fig. 48.23B). Sequencing primers are annealed to the loop region and bound to a single polymerase molecule at the bottom of an optical waveguide. The waveguide allows single molecule detection of transient fluorescent labels that are covalently attached to the terminal phosphate of each dNTP. The four dNTPs are added to the wells with different labels that can be distinguished optically. When a fluorescent dNTP pairs with its complement near to the polymerase active site, it is in a perfect position for fluorescence detection. After being incorporated into the growing strand, the terminal fluorescent label (attached to pyrophosphate) diffuses away from the polymerase. The fluorescent signals are acquired continuously at high speed as the primer proceeds around the loop by **rolling circle amplification**. The process can be stopped after one loop or continued for multiple reads.

Nanopore sequencing. Another single-molecule sequencing approach that does not require amplification and uses electronic, rather than optical, detection is nanopore sequencing. Single DNA strands are channeled through nanopores formed by immobilized proteins. Passage of DNA bases through the protein nanopore generates characteristic electrical signals that can reveal the identity of the bases traversing the nanopore. In essence, this is similar to a nanosized Coulter counter that quantifies base differences on a single strand of DNA. Kilobases of NAs can be sequenced in a single read. The method is nondestructive and can discriminate methylated DNA bases as well as normal bases.

REVIEW QUESTIONS

1. PCR requires:
   - **a.** Product annealing
   - **b.** Primer denaturation
   - **c.** Polymerase denaturation
   - **d.** Primer annealing
2. Restriction fragment length polymorphisms:
   - **a.** Restrict the length of polymorphisms
   - **b.** Require treatment with a restriction enzyme for visualization
   - **c.** Are polymorphisms in restriction enzymes
   - **d.** Are commonly visualized by massively parallel sequencing
3. Which method does not use electrophoresis?
   - **a.** Single nucleotide extension
   - **b.** PCR/RFLP
   - **c.** Dideoxy-termination sequencing
   - **d.** Pyrosequencing
4. Hybridization probes:
   - **a.** Are the same as hydrolysis probes
   - **b.** Can consist of one or two probes
   - **c.** Irreversibly lose fluorescence once hybridized
   - **d.** Never function by resonance energy transfer
5. Real time PCR:
   - **a.** Is a qualitative assay
   - **b.** Can be observed during amplification
   - **c.** Never uses fluorescence
   - **d.** Requires melting analysis
6. How fast can PCR be performed?
   - **a.** In about an hour
   - **b.** In 15 minutes
   - **c.** In 5 minutes
   - **d.** In less than 1 minute
7. Massively parallel sequencing:
   - **a.** Costs more per base than dideoxy sequencing
   - **b.** Is simpler than dideoxy sequencing
   - **c.** Is more descriptive than *next generation sequencing*
   - **d.** Has a fast turnaround time
8. Single molecule sequencing:
   - **a.** Can produce longer read lengths than typical massively parallel sequencing
   - **b.** Is less informative than massively parallel sequencing
   - **c.** Makes it difficult to haplotype 2 variants
   - **d.** Is impossible
9. DNA melting analysis:
   - **a.** Is expensive and time consuming
   - **b.** Is performed before PCR
   - **c.** Plots fluorescence and temperature to produce melting curves
   - **d.** Is destructive
10. Hydrolysis probes:
    - **a.** Are reversible
    - **b.** Increase fluorescence after the probe is hydrolyzed
    - **c.** Are very good for melting analysis
    - **d.** Are less specific than DNA dyes

SUGGESTED READINGS

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*, 53–59.

Dingle, T. C., Sedlak, R. H., Cook, L., et al. (2013). Tolerance of droplet-digital PCR vs real-time quantitative PCR to inhibitory substances. *Clinical Chemistry*, *59*, 1670–1672.

Farrar, J. S., & Wittwer, C. T. (2015). Extreme PCR: Efficient and specific DNA amplification in 15-60 seconds. *Clinical Chemistry*, *61*, 145–153.

Higuchi, R., Dollinger, G., Walsh, P. S., et al. (1992). Simultaneous amplification and detection of specific DNA sequences. *Biotechnology (N Y)*, *10*, 413–417.

Ma, Z., Lee, R. W., Li, B., et al. (2013). Isothermal amplification method for next-generation sequencing. *Proceedings of National Academy Sciences United States of America*, *110*, 14320–14323.

Saiki, R. K., Gelfand, D. H., Stoffel, S., et al. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, *239*, 487–491.

Sakurai, T., & Husimi, Y. (1992). Real-time monitoring of DNA polymerase reactions by a micro ISFET pH sensor. *Analytical Chemistry*, *64*, 1996–1997.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of National Academy Sciences United States of America*, *74*, 5463–5467.

Schena, M., Shalon, D., Davis, R. W., et al. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, *270*, 467–470.

Steinbock, L. J., & Radenovic, A. (2015). The emergence of nanopores in next-generation sequencing. *Nanotechnology*, *26*, 074003.

Vossen, R. H., Aten, E., Roos, A., et al. (2009). High-resolution melting analysis (HRMA): more than just sequence variant screening. *Human Mutation*, *30*, 860–866.

49

Molecular Applications

Frederick S. Nolte, D. Hunter Best, Todd William Kelley, Jay L. Patel, Evi Lianidou, Dave S.B. Hoon, Rossa W.K. Chiu, Y.M. Dennis Lo

OBJECTIVES

1. Describe the role of HIV-1 viral load testing in monitoring antiviral resistance testing.
2. List common organisms that cause upper respiratory infections and how they can be detected.
3. Describe the clinical specimens and procedures used for collecting specimens of *Chlamydia trachomatis* and *Neisseria gonorrhoeae.*
4. Discuss the role of the human microbiome in health and disease.
5. Describe the difference between autosomal dominant, autosomal recessive and X-linked diseases.
6. List common genetic disorders and their mode of inheritance.
7. Explain the advantages that massively parallel sequencing has over traditional testing methods.
8. List the types of genetic abnormalities that are recurrent in hematopoietic malignancies.
9. Describe the testing strategies that are necessary to reliably detect the spectrum of genetic abnormalities in hematopoietic malignancies.
10. Discuss the role of the *TP53* gene in cancer.
11. What are the different forms of ctDNA detectable in cancer patients?
12. Detail why viral ctDNAs are useful for cancer monitoring.
13. What are the differences in ctDNA and CTC in terms of assay merits?
14. Describe the biological properties of cell-free fetal DNA in maternal plasma.
15. Explain the methodological principles of fetal chromosomal aneuploidy assessment based on circulating cell-free fetal DNA analysis.
16. Detail the analytical issues pertinent to successful circulating cell-free fetal DNA analysis.

KEY WORDS AND DEFINITIONS

Antiretroviral therapy (ART) A regimen and schedule of drugs aimed at slowing or eliminating the replication of an RNA virus. Most effective therapies include multiple drugs.

Autosomal dominant A disorder where the dysfunction of a single allele is sufficient to result in a phenotype.

Autosomal recessive A disorder where the dysfunction of both copies of a gene (one inherited from each parent) is necessary for a phenotype to manifest.

Cell-free fetal DNA (cffDNA) DNA fragments derived from the fetus, mainly of the placenta, found in the circulation of pregnant women. The access of cffDNA allows for noninvasive DNA-based prenatal assessment of the fetus.

Circulating tumor cell free DNA(ctDNA) Circulating tumor cell-free DNA detected in body fluids such as blood. Can be used to detect actionable genomic aberrations such as mutations, deletions, amplifications, and fusions in cancer patients.

Circulating tumor cells (CTC) Circulating tumor cells detected in the blood of cancer patients. More prevalent in advance stages of disease; useful for the evaluation of disease recurrence, prognosis, and therapy response.

Fetal fraction The proportion of fetal DNA in relation to the total amount of DNA detectable in plasma of pregnant women.

Fluorescence in situ hybridization (FISH) A targeted technique for assessment of large scale chromosomal abnormalities such as translocations or large copy number variants.

Fusion transcript Abnormal mRNA transcript that contains a portion of one gene fused to a portion of another, which results from the presence of a chromosomal translocation.

Karyotype Low resolution pan-genomic technique for identifying large scale structural chromosomal abnormalities such as monosomies, trisomies, deletions, duplications, and translocations.

Liquid biopsy Real-time monitoring of cancer evolution in the peripheral blood of cancer patients, based mainly on CTC and ctDNA analysis.

Massive parallel sequencing (MPS) A high-throughput DNA sequencing technology capable of generating data on a genomic scale over a short period of time.

Massively parallel sequencing panel A targeted technique for identifying variants (mutations) by massively parallel sequencing in a series of genes that are clinically associated with the same or related conditions; for example, myeloid malignancies.

Metagenomics The study of genetic material recovered from the environment.

Methylated ctDNA Methylation status, typically of gene promoter regions that may be hypo- or hyper-methylated in ctDNA.

Microbiome The spectrum of microorganisms in a particular environment.

Microdeletion A reduction in copy number in a subchromosomal region.

Microduplication An increase in copy number in a subchromosomal region.

Nucleic acid amplification test (NAAT) A molecular assay that identifies or quantifies a pathogen by amplification of its RNA or DNA, most commonly by the polymerase chain reaction or transcription-mediated amplification.

Quantitative polymerase chain reaction (qPCR) Amplification of a specific targeted short sequence, often to detect a genomic aberration, such as mutation, including single nucleotide variants, indels, deletions, or fusions.

Size-selection Methods to analyze DNA fragments of specified size ranges.

Syndromic testing Testing for a panel of microbes, any of which may cause a set of medical signs and symptoms.

Translocations An event that leads to the juxtaposition of a chromosomal segment to a region of a nonhomologous chromosome.

Viral load The number of viral genome copies per milliliter of blood.

Virologic failure A sustained viral load during antiviral therapy of more than a certain amount; for example, greater than 200 copies/mL for HIV-1.

Whole exome sequencing Targeted sequencing of all of the coding regions (and directly adjacent intronic regions) of all genes in the human genome.

X-linked A disorder caused by a gene located on the X chromosome.

There are many applications of nucleic acid molecular diagnostics in the clinical laboratory. Sometimes, the mere presence of DNA or RNA of a particular sequence provides a diagnosis. For example, HIV or *Escherichia coli* sequences are not expected in blood, allowing the diagnosis of infectious disease when their respective sequences are detected in blood. If sequence alterations in human genes affect function and can be passed to future generations, genetic diseases can be identified. Finally, sequence changes within an individual that develop during their lifetime that lead to malignant proliferations are useful to diagnose and monitor cancer. Currently, molecular diagnostics are most readily applied to infectious diseases, genetics, and cancer.

Here we describe specific applications of molecular diagnostics currently used to characterize and help treat patients with a variety of ailments, including infectious diseases, genetic diseases, and cancer. Included are hematologic malignancies and the study of circulating tumor cells (CTC) and circulating tumor cell free DNA (ctDNA). The chapter concludes with a description of using cell-free circulating nucleic acids for prenatal diagnostics.

INFECTIOUS DISEASE

Nucleic acid (NA) amplification techniques are now commonly used to diagnose and manage patients with infectious diseases. The growth in the number of US Food and Drug Administration (FDA)-approved test kits and analyte-specific reagents has facilitated the use of this technology in clinical laboratories. Technological advances in NA amplification techniques, automation, NA sequencing, and multiplex analysis have invigorated the field and created new opportunities for growth. Simple, sample-in, answer-out, molecular test systems are now widely available that can be deployed in a variety of laboratory and clinical settings. Molecular microbiology remains the leading area in molecular diagnostics in terms of both the number of tests performed and the clinical relevance. NA-based tests have reduced the dependency of the clinical microbiology laboratory on more traditional antigen detection and culture methods and created new opportunities for the laboratory to impact patient care. A complete and current list of all FDA-cleared and FDA-approved microbial NA-based tests can be found at http://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/InVitroDiagnostics/ucm330711.htm. As examples of important applications, we focus on human immunodeficiency virus 1 (HIV-1) infection, sexually transmitted infections due to *Chlamydia trachomatis* and *Neisseria gonorrhoeae*, respiratory tract infections, and consideration of the human microbiome.

Human Immunodeficiency Virus 1

HIV-1, the causative agent of the acquired immunodeficiency syndrome (AIDS), is an RNA virus. Replication of the virus is complex and involves reverse transcription of the RNA genome into a double-stranded DNA molecule or provirus, which is integrated into the host genome. The HIV-1 reverse transcriptase (RT) enzyme does not have proofreading capabilities, leading to the significant genetic diversity of HIV-1.

The management of patients with HIV-1 infection was transformed by tests that measure the concentrations of HIV-1 RNA in blood (viral load testing) and tests for resistance to antiviral drugs. With these tools, it is possible to maximize the effectiveness of antiretroviral therapy (ART) for an individual, one of the earliest examples of "personalized medicine."

The first HIV-1 viral load test was approved by the FDA in 1996 and rapidly became the standard of care for monitoring response to ART. Most clinical laboratories are currently

using real-time reverse transcription-polymerase chain reaction (RT-PCR) assays to monitor HIV-1 viral load. Early studies found that patients who have higher viral loads progressed more rapidly to AIDS and death than those with low viral loads. The intra-assay imprecision of HIV-1 viral load assays is approximately 0.12 to 0.2 $\log_{10}$ copies/mL. The biologic variation of viral load in patients not receiving antiretroviral therapy is approximately 0.3 $\log_{10}$ copies/mL. Consequently, changes in viral load greater than or equal to 0.5 $\log_{10}$ copies/mL suggest a significant change in viral replication.

Viral load testing is now widely accepted as a marker of response to ART. After treatment is initiated, viral load testing is crucial for monitoring the response to therapy. The standard of care is to treat with a combination of antiretroviral drugs, which are classified based on their viral targets. After the initiation of appropriate therapy, there is typically a 2 $\log_{10}$ or greater decrease in viral load within 2 to 3 months. Data have shown that the lower the absolute viral load, the better the clinical and virologic outcomes. Guidelines recommend quantifying plasma HIV-1 RNA immediately before initiating therapy and 2 to 8 weeks later, with the goal of achieving an undetectable viral load level (<20 to 50 copies/mL) within 16 to 24 weeks of initiating therapy. It is important to determine early in the treatment course if there is suboptimal viral load suppression, so that factors affecting adherence can be assessed and, if needed, the regimen altered. After the initial response has been characterized, the viral load should be monitored every 3 to 4 months to ensure that the response to therapy is sustained.

Of note, viral blips (defined as a detectable viral load usually <400 copies/mL) occur in successfully treated individuals and are not predictive of virologic failure. For this reason, *virologic failure* is defined as a sustained viral load of more than 200 copies/mL.

Viral load testing also aids in the diagnosis of acute HIV-1 infection (the window period after infection that occurs before detectable antibody production), although the currently available viral load assays are not FDA approved for diagnostic purposes. During this period of early infection, patients typically have very high viral loads ranging from 10^5 to 10^7 copies/mL.

A proviral, PCR DNA test, although not FDA cleared, is helpful in neonates born to HIV-1-infected mothers receiving ART. Both antiretroviral agents and maternal antibody cross the placenta. Antiretroviral agents can suppress the replication of the virus in neonates, and RNA tests can be falsely negative early after birth. Maternal HIV-1 antibody can persist in neonates for up to 2 years after birth and is therefore not a reliable marker of neonatal infection.

Viral load testing is routinely performed on plasma specimens, and ethylenediaminetetraacetic acid (EDTA) is the anticoagulant of choice. Acid–citrate–dextrose (ACD) is also an acceptable anticoagulant, but blood anticoagulated in heparin is unacceptable for most tests. It is critical to minimize the risk of RNA degradation during specimen collection and transport. Plasma should be separated within 6 hours of collection and ideally stored at −20°C, although plasma viral RNA is stable at 4°C for several days. Special blood collection containers, or tubes, are available that contain a gel that provides a physical barrier between the plasma and cells after centrifugation. The tubes can be shipped without the need to transfer the plasma into a separate tube. Tubes should not be frozen before pouring off of the plasma because this may lead to falsely elevated viral load assays.

The clinical utility of antiviral resistance testing in HIV-1-infected individuals is well established and should be performed in the following situations: (1) before the initiation of ART in treatment-naïve patients, (2) to guide the selection of active drugs when changing antiretroviral regimens, (3) for the management of suboptimal viral load reduction, and (4) in all pregnant women before the initiation of therapy.

HIV-1 resistance testing can be done by genotypic or phenotypic methods. Genotypic assays identify specific mutations or nucleotide changes that are associated with decreased susceptibility to an antiviral drug. Phenotypic assays are performed by the creation of a pseudoviral vector, and by measuring its replicative capacity in varying concentrations of a drug and comparing it with replication of wild-type virus. Both genotypic and phenotypic methods are used clinically for assessing antiviral resistance in patients, with phenotypic testing usually reserved for drug-experienced patients with complicated resistance profiles.

The first step in genotypic assays is the isolation of HIV-1 RNA from plasma followed by RT-PCR amplification and sequencing of RT protease and integrase genes. Analysis of the results involves sequence alignment and editing, mutation identification by comparison with wild-type sequence, and interpretation of the clinical significance of the mutations identified.

A limitation of currently used genotypic and phenotypic assays is that they can detect only those mutants that make up at least 20% of the total viral population. Regimens that are chosen based on resistance testing may not always be effective because the minority populations will quickly predominate in the presence of the drug. Drug selection pressure is also needed for some resistance mutations to persist at detectable concentrations in the viral population; when drug therapy is discontinued, the wild-type virus may quickly predominate. For this reason, it is recommended that specimens for resistance testing be obtained while the patient is on ART. The minimum viral load required for reliable resistance testing is approximately 1000 copies/mL.

POINTS TO REMEMBER

- HIV exhibits genetic diversity globally and within individual patients
- HIV RNA is the earliest marker of infection
- Viral load testing is a widely accepted marker of response to therapy with the goal of suppression to undetectable amounts
- Special care should be taken in sample processing to avoid spurious increases in HIV viral load tests
- HIV resistance genotyping should be performed on all therapy naïve patients prior to the initiation of antiretroviral therapy and in the selection of active drugs when changing therapy

SEXUALLY TRANSMITTED INFECTIONS

Chlamydia trachomatis and *Neisseria gonorrhoeae*

Testing for *C. trachomatis* (CT) and *N. gonorrhoeae* (NG) is discussed together because several of the available NA amplification tests (NAATs) for these pathogens are duplex assays. Although both CT and NG can cause a variety of clinical symptoms, the focus here is on genital infections.

The detection of CT is a challenging and important public health issue. CT is a major cause of sexually transmitted infections (STIs), with an estimated 1 million cases occurring annually among sexually active adolescents and young adults in the United States. More than half of the infections are asymptomatic. Even when symptomatic, the diagnosis can be missed as the manifestations are protean. In men, CT infection may present as urethritis, epididymitis, prostatitis, or proctitis; in women it may present as cervicitis, endometritis, and urethritis, with 10% to 40% of infections in women progressing to pelvic inflammatory disease (PID) if untreated. Related complications include chronic pelvic pain, ectopic pregnancy, and infertility. In the United States, CT infection is the likely cause of most secondary infertility in women. In pregnant women, there is the additional risk of transmitting the infection to the newborn during labor and delivery, leading to pneumonia or conjunctivitis in the newborn.

N. gonorrhoeae infection, too, may present in various ways, and the clinical presentations overlap with those of CT. Men may have acute urethritis with discharge, epididymitis, prostatitis, and urethral strictures. In women, NG infection can produce cervicitis, which, if left untreated, can lead to PID, abscesses, or salpingitis.

Traditional methods for the diagnosis of CT infection include cell culture, antigen detection by immunofluorescence, enzyme immunoassay (EIA), and nonamplified NA probes. These traditional methods have been replaced in most laboratories by amplified NA tests, which provide greater sensitivity in detecting CT from genital specimens. For NG, which was traditionally diagnosed based on culture methods that relied on selective culture media, NA testing does not offer significant improvement in sensitivity compared with culture. However, NA testing for NG offers a sensitive and reliable alternative to culture since it is easier to maintain the integrity of the target NA than it is to maintain the organism's viability.

In addition to high diagnostic and analytical sensitivity and specificity, NA testing offers several advantages over conventional culture and antigen detection methods for the diagnosis of CT and NG. Testing for both pathogens can be done on a single specimen, and for some duplex assays, testing is performed in a single reaction. The stability of NA avoids the necessity of immediate transport to the laboratory, and specimens may be stored refrigerated or at room temperature before transport. Transport and storage requirements vary among tests, so it is important to refer to the package insert for specific details. An additional advantage of NA testing is the use of urine specimens, which for women allows testing to be done without the need for a pelvic examination. In men, urine offers a convenient and diagnostically sensitive alternative to collection with a urethral swab and increases the likelihood that asymptomatic men will agree to be tested.

NAATs for the detection of CT and NG from clinical specimens use a variety of specimens, including cervical and vaginal swabs, urethral swabs, and urine from both asymptomatic and symptomatic individuals. The diagnostic sensitivity of the tests varies according to the specimen type and whether the patient is asymptomatic or symptomatic. Interpretation of the results of NA testing for CT can be challenging because these assays are more sensitive than culture, which was previously used as the gold standard for clinical trials. For men, the diagnostic sensitivity of testing urine specimens is nearly equivalent to that of testing urethral swabs. A limited volume (20 to 50 mL) of first-passed urine is preferred because larger volumes will lead to a decreased concentration of the organism in the sample and thus reduce the diagnostic sensitivity. With proper specimen collection, male urethral swabs and urine specimens have a sensitivity of nearly 100% for the detection of NG or CT infection. For women, vaginal and cervical swab specimens provide the highest sensitivity for the detection of NG and CT infection, with many studies showing a sensitivity of 90% to 95%. Vaginal swabs are preferred because they are easier to collect. Urine specimens can be used, but they generally result in a lower diagnostic sensitivity than cervical swabs (75% to 85%). An alternative to urine testing in women is the use of self-collected vaginal swabs, which have been shown in some studies to have a diagnostic sensitivity that is equal to that obtained with cervical swabs. Because DNA can persist in urine samples for up to 3 weeks after completion of therapy, test of cure using NA testing is discouraged. If this must be done, then testing should be delayed for at least 3 weeks after therapy is completed to allow time for clearance of the pathogen DNA.

Other important sexually transmitted infections for which NAATs are now the standard of care include the protozoan *Trichomonas vaginalis*, herpes simplex virus, and human papilloma virus.

RESPIRATORY TRACT INFECTIONS

The viruses that infect the respiratory tract consist of large and diverse groups that cause disease in humans, and new ones continue to be discovered. The more common viruses that infect humans include influenza A and B, parainfluenza virus (PIV) types 1 to 4, respiratory syncytial virus (RSV), metapneumovirus, adenoviruses (>50 different types), rhinoviruses (>100 different types), and coronaviruses (5 types). The disease spectrum ranges from the common cold to severe life-threatening pneumonia. It is difficult to differentiate the viral origin based on signs and symptoms alone, and treatment options vary depending on the viral etiology. Infection with these viruses has demonstrated the potential for global public health threats of epidemic and pandemic proportions. The 1918 influenza A pandemic, the avian influenza A H5N1 in 1997, the severe acute respiratory syndrome (SARS)

coronavirus outbreak in 2003, the 2009 pandemic of influenza A H1N1, and the emergence of Middle East respiratory syndrome (MERS) coronavirus in 2012 are all reminders of the potential threats to human health posed by novel respiratory viruses. Detection of emerging respiratory viruses will require multiple modalities, but molecular methods have been crucial to their discovery and characterization, and in the development of diagnostic tools.

Acute respiratory viral infection is (1) a leading cause of hospitalization and death in infants and young children; (2) contributes to problems of asthma exacerbation, otitis media, and lower respiratory tract infection; and (3) exacerbates acute disease in immunocompromised and elderly patients. Rapid diagnosis aids in effective treatment and management.

Rapid antigen-based EIAs provide short turnaround times (TATs; minutes) but are hampered by poor diagnostic sensitivity and low positive predictive values compared with culture methods or molecular assays. Direct fluorescent antibody (DFA) detection assays for viral antigens on centrifuged cellular material from nasopharyngeal swabs, aspirates, or wash specimens demonstrate greater rates of detection than the rapid antigen assays and provide results in a relatively short time frame of 2 to 4 hours. However, detection rates are lower for antigen detection methods than for NAATs.

Cell culture methods, although slower than antigen detection methods, have been considered the gold standard for detection of a wide range of viral pathogens. In recent years, culture methods have been optimized for detection by combining multiple cell lines and improving the TAT from weeks to days through the use of shell-vial spin amplification cultures. Here, the patient's specimen is concentrated onto cells grown on a coverslip, and fluorescent antibody detection is performed after 16 to 24 hours of incubation instead of waiting for the development of a cytopathic effect. Although this has hastened the time to detection, 1 to 2 days is still required along with significant technologist labor, and it is not quite as sensitive as molecular methods of detection.

Molecular detection of respiratory viruses offers several advantages over traditional virologic culture or antigen detection. Most important, analytical sensitivity of molecular assays, primarily using PCR or real-time PCR, is consistently better than traditional methods. Results from molecular testing are more accurate, and thus the patient benefits from the most appropriate treatment decision; also, infection control practitioners can more effectively implement strategies to prevent or reduce nosocomial transmission. Molecular assays can be designed to detect a wide range of viral pathogens, including viruses that are difficult to culture.

Despite the advantages of NAATs for respiratory virus detection, their adoption by clinical laboratories was initially slow because of the limited capacity of real-time PCR assays for multiplexing. To provide comprehensive coverage for respiratory viruses, a panel of assays needed to be deployed, which is an approach not practical for many laboratories.

Currently, there are several FDA-cleared multiplexed respiratory virus panels capable of detecting up to 20 different viral targets, thus providing simplified approaches to comprehensive diagnostics for respiratory viruses. See Table 49.1 for an overview of the important parameters of these respiratory virus panels. These tests are truly transformative for the laboratory in that they can replace the combination of limited multiplex NAATs, antigen detection tests, and culture-based methods that were traditionally used in clinical laboratories to detect respiratory viruses and thus dramatically increase diagnostic yield.

One approach is multiplexed RT-PCR with fluorescently color-coded microsphere (bead) hybridization for simultaneous detection and identification of 17 respiratory viruses and subtypes (see Table 49.1, column 1). The multiplexed RT-PCR primers amplify conserved regions of the viruses, and the products are labeled with biotin-containing deoxynucleotides in a second target-specific primer extension reaction. The extension product has a proprietary tag sequence incorporated for hybridization to the virus-specific probe on the color-coded bead. After hybridization, fluorescently-labeled streptavidin is bound by the biotin-labeled primer extension products, and the fluorescent signal is quantified on a flow cytometer. The instrument contains at least two lasers to identify the microbe by a color-coded bead and to detect the primer extension product. The data are recorded as mean fluorescent intensities. Because there are a number of post-PCR processing steps, care must be taken to avoid amplicon cross-contamination and false-positive results.

Another system is automated and provides sample-to-answer results with freeze-dried reagents to perform NA purification, reverse transcription, and nested, multiplex PCR followed by high-resolution melting analysis (see Table 49.1, column 2). The respiratory panel on this instrument simultaneous detects 17 viral and 3 bacterial respiratory pathogens. The test is initiated by loading water and an unprocessed patient nasopharyngeal swab specimen mixed with lysis buffer into a single use disposable that is placed into the instrument. The software has a simple interface that requires only identification of the specimen and pouch barcode to initiate a run. Multiplexed two-stage RT-PCR followed by high-resolution melting analysis of the target amplicons is used to detect each of the panel analytes. Results are reported in an hour; currently, the instrument is designed to test a single sample per run, though multiple instruments can be linked. A second version of the panel with 45-minute analysis time, and a scalable, higher throughput instrument have recently been cleared by the FDA. Because it is a completely closed system, false-positive results caused by amplicon cross-contamination are not an issue. A version of this system has been Clinical Laboratory Improvement Amendments (CLIA) waived.

Another respiratory multiplex system uses electrochemical-detection-based DNA microarrays to detect PCR products from a multiplex PCR reaction (see Table 49.1, column 3). These microarrays are composed of a printed circuit board consisting of an array of gold-plated electrodes. Each electrode is modified with a multicomponent, self-assembled monolayer that includes presynthesized oligonucleotide capture probes. NA detection is based on a sandwich assay principle.

TABLE 49.1 Parameters of Different Food and Drug Administration-Cleared Respiratory Virus Panels

Detection Method	Hybridization to Labeled Bead Array[a]	Melting Curve Analysis[b]	Hybridization and Electronic[c]	Hybridization on Gold Nanoparticles[d]
Amplification	PCR, primer extension	Nested PCR	PCR	PCR
On-board sample processing	No	Yes	Yes	Yes
Post-PCR manipulation	Yes	No	No	No
Hands-on-time (min)	45	3	2	5
Throughput	High	Low	Moderate	Low
Analysis time (h)	4	1	1.5	2
TAT with NA extraction (h)	6	1.1	1.5	2.1
Complexity	High	Low	Low	Moderate
Pathogens detected	ADV	ADV	ADV	ADV
	INF A (H1, H3,09H1)	INF A (H1, H3, 09H1)	INF A (H1, H3, 09H1)	INF A (H1, H3)
	INF B	INF B	INF B	INF B
	MPV	MPV	MPV	MPV
	RSV	RSV	RSV (A, B)	RSV (A, B)
	RV	RV/EV	RV/EV	RV
	PIV (1, 2, 3, 4)	PIV (1, 2, 3, 4)	PIV (1, 2, 3,4)	PIV (1, 2, 3, 4)
	COV (HKU1, NL63, 229E, OC43)	COV (HKU1, NL63, 229E, OC43)	COV (HKU1, NL63, 229E, OC43)	*B. pertussis*
	Bocavirus	*Bordetella pertussis*	*B. pertussis*	*B. parapertussis/B. bronchiseptica*
		Chlamydophila pneumoniae	*C. pneumoniae*	*B. holmesii*
		Mycoplasma pneumoniae	*M. pneumoniae*	

[a]Luminex xTAG RVP Fast v2.
[b]BioFire FilmArray TORCH RP.
[c]GenMark ePlex.
[d]Luminex Verigene RP Flex.
ADV, Adenovirus; *COV,* coronavirus; *EV,* enterovirus; *INFA,* influenza A virus; *INFB,* influenza B virus; *MPV,* metapneumovirus; *PCR,* polymerase chain reaction; PIV, parainfluenza virus; *RSV,* respiratory syncytial virus; *RV,* rhinovirus.

Signal and capture probes are designed with sequences complementary to immediately adjacent regions on the corresponding target DNA sequence. A three-member complex is formed between the capture probe, target sequence, and signal probe based on sequence-specific hybridization. This process brings the 5′ end of the signal probe containing electrochemically active ferrocene labels into close proximity to the electrode surface. The ferrous ion in each ferrocene group undergoes cyclic oxidation and reduction, leading to loss or gain of an electron, which is measured as current at the electrode surface using alternating-current voltammetry.

A simple sample-to-answer, self-contained cartridge has been developed that automates and fully integrates nucleic acid extraction, amplification, and electrochemical detection of 19 respiratory viruses and subtypes, and 3 bacterial respiratory pathogens. The analysis time is approximately 90 minutes and it is performed on a moderate throughput instrument. Each instrument is modular and scalable with one to four towers with each tower capable of processing 192 tests in 8 hours.

Yet another respiratory multiplex system uses PCR amplification and gold nanoparticle-labeled probes to detect target NA hybridized to capture oligonucleotides arrayed on a glass slide (see Table 49.1, column 4). Silver signal amplification is then performed on the gold nanoparticle probes that are hybridized to the captured DNA targets of interest. The reader optically scans the slide for silver signal, processes the data, and produces a qualitative result for 13 respiratory viruses and subtypes and 4 *Bordetella* spp.

Molecular testing for respiratory viruses will likely continue to include tests designed to detect a limited number of viruses of particular importance (e.g., influenza A and B viruses and RSV), as well as tests that detect a broad array of viruses because there are clinical needs for both types of tests. The use of comprehensive respiratory virus panels greatly increases diagnostic yield and the ability to detect mixed viral infections. However, the clinical significance of mixed infections is not well documented or understood. In addition, there are test options that range from simple, "sample in answer out" systems to complex tests that require multiple manual steps, meeting different niches for various clinical laboratory settings. In fact, point-of-care molecular testing for respiratory viruses is now possible with the recent development of a CLIA-waived tests for influenza A and B viruses and RSV. Using isothermal amplification or real-time PCR these systems deliver results within 20 minutes and can be performed by nonlaboratory personnel with performance characteristics similar to those of NAATs performed in laboratories.

HUMAN MICROBIOME AND METAGENOMICS

Microbial inhabitants outnumber our own body's cells by about 10 to 1 and at a genomic level have 100-fold greater gene content than the human genome. Interest in elucidating the role of resident organisms in human health and disease has flourished over the past decade with the advent of new technologies for interrogating complex microbial communities. The microbiome is the totality of microbes, their genetic information, and the milieu in which they interact. It includes bacteria, fungi, viruses and phages, and parasites, but most of the emphasis to date has been on the bacterial component of the microbiome. However, progress is being made toward defining the human virome and the role that it plays in complex microbial communities.

Metagenomics refers to the concept that a collection of genes sequenced from the environment could be analyzed in a way analogous to the study of a single genome. Facilitated by advances in NA sequencing technology, this technology has permitted the study of microbial communities directly in their natural environment, thus bypassing the need for isolation and cultivation of individual species. We now know that the majority of microorganisms from the human body cannot be cultured in vitro. Most taxonomic metagenomic studies have used targeted sequencing of the phylogenetically informative regions of the 16s rRNA gene from bacteria because it has long been the gold standard method for bacterial identification, and there are large sequence databases and sophisticated analysis tools available. However, 16s rRNA gene sequencing does not provide enough information for comprehensive microbiome studies. To overcome the limitations of single gene-based amplicon sequencing, researchers have used whole genome shotgun sequencing on massively parallel sequencing platforms (MSP, also known as NGS—next-generation sequencing). Whole-genome approaches permit identification and annotation of diverse sets of microbial genes that encode many different biochemical or metabolic functions, thus providing functional metagenomic information.

Alterations of the microbiome in many different disease states have been described (Table 49.2). Establishing a causal link between microbiome changes and a specific disease often is challenging because most studies have been observational, the disease entities themselves may not be well defined, and the pathogenesis may be multifactorial.

However, it seems clear that future approaches in medical microbiology will be shaped in part by developments in metagenomics and human microbiome research. The identification of single agents of infection will be supplemented by techniques that will determine the relative composition of microbiomes in the context of different infections and other disease states. Recent evidence that recurrent *C. difficile* infections can be treated by reconstituting the normal colon microbiota in the patient by transferring feces from a normal donor is a good example of how a better understanding of changes in the microbiome composition can lead to effective treatments options. Differences in the composition of the microbiome that may cause or contribute to noninfectious diseases may offer new opportunities for clinical microbiology laboratories to impact other areas of medicine. Finally, metagenomic techniques will facilitate the discovery of previously unrecognized pathogens and increase our understanding of how changes in the microbiome may contribute to infectious diseases.

TABLE 49.2 Association of Human Disease With Changes in the Microbiome

Disease	Relevant Change
Psoriasis	Increased ratio of *Firmicutes* to *Actinobacteria*
Reflux esophagitis	Esophageal microbiota dominated by gram-negative anaerobes; gastric microbiota with low or absent *Helicobacter pylori*
Obesity	*Reduced ratio of Bacteroidetes* to *Firmicutes*
Childhood-onset asthma	Absent gastric *H. pylori* (especially cytotoxin-associated gene genotype)
Inflammatory bowel disease (colitis)	Increased Enterobacteriaceae
Functional bowel diseases	Increased *Veillonella* spp. and *Lactobacillus* spp.
Colorectal carcinoma	Increased *Fusobacterium* spp.
Cardiovascular disease	Gut-dependent metabolism of phosphatidylcholine

Modified from Cho, I., & Blaser, M. J. (2012). The human microbiome: at the interface of health and disease. *Nat Rev Genet, 13,* 260–270.

Molecular microbiology will continue to be one of the leading growth areas in laboratory medicine as it becomes less technically complex and more accessible. However, now more than ever, clinical and financial outcomes data will be needed to justify the use of this often expensive technology in an era of declining reimbursement and increased cost consciousness.

GENETICS

Advances in technology and bioinformatics have revolutionized the field of human genetics. Examples of common inherited autosomal recessive, autosomal dominant, and X-linked diseases are presented here as examples.

Autosomal Recessive Disorders

An individual with an autosomal recessive disease has inherited two abnormal alleles at a given locus by receiving one mutant allele from each carrier parent; the disease-causing gene is on one of the autosomes (1 to 22) and not on a sex chromosome (X or Y). Typically, the carrier parents with one abnormal allele have no clinical features of the disease yet possesses a 50% risk of donating the mutant allele to his or her offspring. Matings in which both partners are carriers of an abnormal allele have a 25% chance of producing a child with both normal alleles, a 50% chance of having a child that has received only one abnormal allele, and a 25% chance of having an affected child. The affected patient may

be homozygous for a specific mutation by receiving the same mutation from each parent or may be a compound heterozygote having received a different mutation from each parent. The specific mutations present influence the clinical severity of the disease and account for variability in the expression of the disease between patients, referred to as genotype–phenotype correlation. Modifier genes and environmental factors also play a role in determining the patient's phenotype. Among pedigrees illustrating autosomal recessive disorders, males and females are equally affected, and for rare diseases, consanguinity is likely to be observed.

Cystic Fibrosis

Cystic fibrosis (CF; Online Mendelian Inheritance in Man [OMIM] #219700) is one of the most common autosomal recessive diseases in people of Northern European ancestry with an estimated incidence in the United States of about 1 in 2500 and a carrier frequency of about 1 in 25. Within other ethnic populations, the disease has an estimated incidence of 1 in 2300 Ashkenazi Jews, 1 in 8500 Hispanics, 1 in 17,000 African Americans, and 1 in 35,000 Asian Americans. CF is a multisystem disorder characterized by progressive pulmonary disease, pancreatic insufficiency, elevated sweat electrolytes, male infertility, and a predisposition to sinonasal disease. An abnormal sweat chloride concentration is considered the gold standard for the diagnosis of CF, and a result of 60 mmol/L or greater is considered diagnostic of CF. Interestingly, the phenotypic expression of the disease is heterogeneous, ranging from meconium ileus and severe respiratory disease in infants to mild pulmonary symptoms and no evidence of gastrointestinal problems in adulthood. Atypical CF patients with a nonclassic presentation may have involvement of only one organ, as in congenital bilateral absence of the vas deferens (CBAVD), pancreatitis, rhinosinusitis, or nasal polyps. Variability in expression is explained by both allelic heterogeneity at the CF gene locus and genetic variation in modifier genes. Loss or decreased amounts of the disease-associated protein cause mucous accumulation and airway obstruction, recurrent infection with pathogens and excessive inflammation with progressive lung damage, and ultimately respiratory failure. Because morbidity and mortality are most related to severe lung disease, daily management of these patients is to prevent bacterial lung infections through physical therapy of the chest and early or aggressive treatments to eradicate bacterial colonization. Although originally considered a fatal childhood disease, the US Cystic Fibrosis Foundation reported that in 2008, more than 45% of patients were older than 18 years of age. Furthermore, the median survival age among approximately 23,000 patients with CF who received care through one of the nationwide CF care centers was 37.4 years with the life expectancy of patients born now in their 50s. The increase in survival age is due to organ transplantation, improved nutrition, and new therapies.

The severity and frequency of the disease led to an intensive search for the gene, which was eventually cloned in 1989. The CF gene maps to chromosome 7q31.2 with 27 exons encoding a transcript of approximately 6.5 kb. The CF transmembrane conductance regulator protein (CFTR) has 1480 amino acids and is a member of the ATP-binding cassette (ABC) transporter superfamily of membrane transport proteins. The wide clinical diversity of CF is based in part on the varying effects conferred on this protein from the more than 2000 mutations reported within this gene.

Because CF is an autosomal recessive disorder, patients with CF must have two mutant *CFTR* alleles to develop the disease. Some mutations are "private" and unique to a family; others may be common among CF patients. More than half of all mutations are missense or frameshift mutations, with exon 14 containing the largest number of different disease-causing mutations. The types of mutations and frequencies of each differ significantly among populations.

The most common mutation, p.Phe508del (c.1521_1523delCTT) is detected in about 70% of CF alleles in whites of Northern European descent. This CFTR protein is misfolded and not properly processed by the endoplasmic reticulum with the majority of the protein rapidly degraded. Understanding the effect of various mutations on the CFTR protein is important for choosing the corrective drug therapy required for each patient.

The *CFTR* genotype and clinical phenotype correlations are most closely related for pancreatic involvement rather than for pulmonary manifestations of the disease. Most patients with two "severe" mutations, which cause a "severe" CF phenotype have pancreatic insufficiency, but patients with one or two "mild" mutations, which cause a "mild" CF phenotype have pancreatic sufficiency but have an increased risk of developing pancreatitis.

DNA testing for the identification of *CFTR* mutations is performed for a variety of reasons (Box 49.1). It is performed to confirm the diagnosis of disease in patients with equivocal sweat chloride results or when an insufficient amount of sweat is collected and no results are obtained from the test. The presence of two known pathogenic *CFTR* mutations is diagnostic for CF. A diagnosis of CF can be considered in the patient with a *CFTR*-related disorder such as chronic pancreatitis, CBAVD, or sinusitis. In known CF patients, mutation

BOX 49.1 Referrals for Cystic Fibrosis Transmembrane Conductance Regulator Mutation Analysis

- Confirm diagnosis of cystic fibrosis
- Determine prognosis
- Screen patient with pancreatitis
- Family member testing
- Newborn screening
- Preconception couples
- Expectant couples
- Prenatal testing—at-risk fetus
- Prenatal testing—hyperechogenic bowel
- Preimplantation genetic diagnosis
- Infertile male with congenital bilateral absence of the vas deferens
- Semen and oocyte donors

analysis can be requested to help predict the prognosis based on genotype–phenotype correlations and initiate mutation-specific treatment options. At the same time, identifying the *CFTR* gene mutations that are present in the proband and how they segregate in the family enables preimplantation diagnosis or prenatal testing for subsequent pregnancies and enables carrier or diagnostic testing for other at-risk family members. Similarly, state-sponsored newborn screening (NBS) programs have detected infants with CF and at the same time have identified at-risk family members, and have enabled carrier or diagnostic genetic testing.

Cystic fibrosis carrier screening may be offered as a single gene-specific test or may be included within high-throughput large carrier screening panels that detect both common *CFTR* gene mutations as well as targeted mutations in other genes. Although these more extensive panels provide additional information to patients, they contain genes for inherited conditions for which population-based screening is not recommended in current practice guidelines. A core panel of the most common mutations includes disease alleles with an estimated prevalence of at least 0.1% of CF mutations and a carrier detection rate of about 88% for non-Hispanic whites. Although CF is more common in the white and Ashkenazi Jewish populations, the standard of care is to make CF testing available to all preconception or expectant couples, especially because it is becoming more difficult to assign a single ethnicity to a patient to best determine the carrier risk. Counseling for CF carrier testing is complex and should include information about CF and CFTR-related disorders and the a priori likelihood of having a child with CF based on personal and family history and ethnicity. Furthermore, it is important for the patient to understand the inability of the screening panel to detect all CF mutations and the residual risk of being a carrier despite a negative test result. This is especially important for patients in ethnic groups for which the *CFTR* gene mutation detection level is reduced. The genetics professional should discuss the possibility of stigmatization or anxiety associated with being a carrier of a genetic disease and how knowing this information may affect her pregnancy. After DNA testing, it is very important for the genetics professional to know relevant family history when interpreting and reporting the test results to the patient and reproductive partner to accurately assess the risk to the couple of having a child with CF.

In families at risk for a child with CF, prenatal testing of the fetus may be requested. In these prenatal cases and for all DNA testing of fetal DNA, it is the standard of care that the DNA obtained from the fetus be tested for the presence of maternal contamination because the presence of maternal DNA can interfere with test result interpretation. This is best performed by using PCR amplification for highly polymorphic short tandem repeat loci coupled with capillary electrophoresis.

CFTR gene mutation testing is performed using a laboratory developed test or one of a variety of commercially available platforms with kits cleared by the FDA. The number of mutations detected by each assay is variable, and the median turnaround time is 14 days. Furthermore, because the detection rate of the 23-mutation gene panel is lower in some ethnicities, laboratories serving such populations should consider supplementing this screening panel with additional mutation analysis. Although full gene sequencing by Sanger or massively parallel sequencing can be done to detect *CFTR* gene mutations, current practice guidelines do not recommend this for CF carrier screening. Rather, sequencing of the *CFTR* gene should be performed to confirm the diagnosis of suspected CF when no or only one *CFTR* gene mutation was detected with the 23 common mutation panel. If *CFTR* gene mutations are identified by sequencing in the proband, targeted analysis to specifically screen for that mutation in at-risk family members may be performed.

Autosomal Dominant Diseases

In autosomal dominant disorders, a single abnormal allele is sufficient to cause disease despite the presence of a normal allele. An individual with an autosomal dominant disease may have inherited an abnormal allele from an affected parent, or the mutant allele may have arisen de novo as a new mutation during gametogenesis in an unaffected parent. The disease-causing gene is on one of the autosomes (1 to 22) and is not on a sex chromosome (X or Y). An affected individual has a 50% risk of donating the mutant allele to each offspring. Different mutations within the gene may have varying effects on the protein, causing variability in clinical expression between patients who have pathogenic mutations. In some instances, known mutant gene carriers have no clinical symptoms of the disease, a phenomenon referred to as *reduced penetrance*; however, they still possess a 50% risk of donating the mutant allele to each offspring. Differences in phenotypic expression of the disease between patients who share identical gene mutations are commonly explained by the effects of modifier genes or environmental influences (or both). Among pedigrees illustrating autosomal dominant inheritance, both males and females are affected, and male-to-male transmission is observed (unlike X-linked inheritance).

Huntington Disease

Huntington disease (HD; OMIM#143100) is an autosomal dominant, late-onset neurodegenerative disorder with an incidence of about 3 to 10 per 100,000 in most populations but may be as high as 10 to 15 in some populations of Western European origin. First described by George Huntington in 1872, this progressive disease is characterized by choreic movement, cognitive decline, and ultimately dementia and psychiatric disturbances. The mean age of onset is between 35 and 44 years, but subtle changes in personality may be evident before diagnosis. Approximately 25% of patients first display symptoms after the age of 50 years, and about 5% to 10% of patients have juvenile HD with the age of onset before 20 years. The median survival time is 15 to 20 years after the onset of symptoms.

The molecular basis for HD is an expansion of a glutamine-encoding CAG trinucleotide repeat in exon 1 of the HD gene, *HTT*. Normal CAG repeat lengths range from 10 to 26,

repeats of 27 to 35 are considered intermediate or "mutable," repeats of 36 to 39 are considered HD alleles associated with reduced penetrance, and repeats of 40 or greater are diagnostic of HD (Fig. 49.1).

HD is one of many trinucleotide repeat expansion diseases associated with neurologic and neuromuscular dysfunction. In HD, the number of CAG repeats is inversely correlated with the age at onset of the disease. Patients with onset as early as 2 years of life have a repeat number approaching 100 or greater, and late-onset-disease patients have repeat numbers of 36 to 39. Although the CAG-repeat number accounts for the majority of variance in age of onset for HD (≈70%), the remainder of the variance comes from modifier genes and environmental factors.

The *HTT* gene contains 67 exons and encodes a novel protein, huntingtin (htt), with 3144 amino acids and a molecular mass of approximately 350 kDa. Htt is ubiquitously expressed in both neural and nonneural tissue with highest levels in the brain. Htt is required for neuronal development with the complete absence of htt being lethal in mice. Htt is predominantly localized in the cytoplasm, but lesser amounts are also found in the nucleus. Mutant htt with expanded polyglutamine tracks are effectively transcribed and translated, but as a result of the increase in glutamine residues, the protein is misfolded. This abnormal folding conveys a gain of function and an increased amyloid-like ability to form protein aggregates, sequestering other cellular proteins preventing their normal function, thereby disrupting normal protein homeostasis and conferring neurotoxicity. Interestingly, patients with two expanded CAG-repeat alleles do not have more severe disease than heterozygotes.

DNA testing for HD is performed using PCR to determine the CAG-repeat number. The most common method includes PCR with fluorescently labeled primers coupled with capillary electrophoresis to size the PCR products and determine the number of repeats.

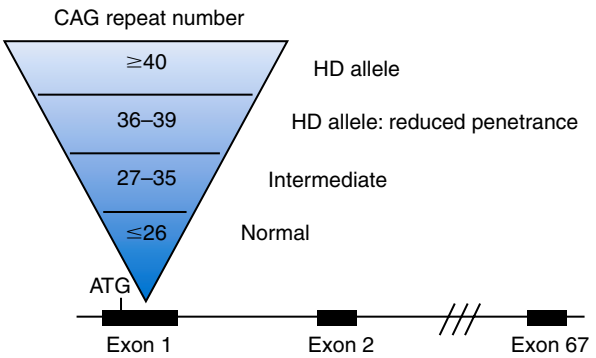


Fig. 49.1 Schematic representation of the polyglutamine-encoding CAG repeat in exon 1 of the Huntington disease *(HD)* gene and associated alleles. A CAG-repeat number of 26 or less is considered normal. CAG-repeat numbers of 27 to 35 are intermediate, and although they are not associated with an abnormal phenotype, these alleles are susceptible to meiotic expansion to an HD allele. CAG repeats of 36 to 39 are considered HD alleles with reduced penetrance, indicating that both unaffected and affected patients have been reported with alleles of this size. CAG repeats of 40 or more are associated with HD with complete penetrance.

Many ethical issues are associated with HD testing, primarily as they relate to presymptomatic testing (Box 49.2). The approach used for presymptomatic HD testing has become the model for late-onset single-gene disorders, and similar formats have been applied to other late-onset inherited diseases. This model includes a multidisciplinary team involving neurology, psychiatry, and genetics specialists; pretest evaluation and counseling; individual support from family members or close friends; and posttest follow-up sessions. Predictive testing should be performed for patients 18 years of age or older and only with informed consent. Informed consent implies that the patient has been thoroughly counseled and clearly understands both the advantages and the disadvantages of knowing the results. An advantage of having this test is the removal of uncertainty regarding whether patients have or have not inherited the mutant allele and thus a feeling of relief for those who have not inherited a mutant *HTT* allele. This knowledge can help patients plan their career goals and personal affairs involving marriage, children, and long-term care insurance. Disadvantages of knowing this information include, but are not limited to, (1) the feeling of "survivor's guilt" in those who learn that they have not inherited a mutant allele while other family members have, (2) fear from learning that they have inherited a mutant *HTT* allele and will develop this incurable disease, (3) potential risk of discrimination in employment, (4) concern for passing this gene on to their offspring, and (5) uncertainty of developing the disease if a mutant *HTT* allele with 36 to 39 CAG repeats is identified.

Because HD is a delayed-onset disease, as the asymptomatic at-risk patient ages, the risk of testing positive with an expanded CAG repeat decreases. Thus if the patient elects not to have predictive testing, the genetic counselor can provide information regarding the probability that an HD mutation exists, which, based on the individual's age, may provide some comfort to the patient.

Prenatal testing for HD, another complicated issue associated with this disease, may not be provided in all laboratories

BOX 49.2 Ethical Issues Associated With Presymptomatic DNA Testing for Huntington Disease

- Patients must be 18 years of age or older.
- The decision to proceed with testing must be voluntary and informed.
- Genetic counseling regarding the benefits and pitfalls of testing is required.
- A support partner is needed for the patient for counseling and the testing process.
- Diagnosis of Huntington disease in the family should be confirmed by DNA testing before presymptomatic testing.
- Psychiatric assessment of patient is necessary before testing.
- Follow-up genetic counseling is recommended after delivery of results.
- Prenatal testing of fetuses is controversial; preimplantation diagnosis is available.

that perform routine HD testing for several reasons. Ethical issues include possible pregnancy termination for a late-onset disorder, presymptomatic testing of the child if the parents choose not to terminate, and technical issues that may compromise testing of chorionic villus and amniocentesis samples. As an alternative for prenatal testing, preimplantation genetic diagnosis (PGD) can be performed. In PGD, in vitro fertilization is used to produce embryos that then undergo a single cell biopsy for genetic analysis. After PCR testing has been used to determine the *HTT* CAG-repeat numbers, embryos with normal HD alleles are implanted. This method, which combines direct mutation analysis and PGD, eliminates the necessity for subsequent prenatal testing to determine the HD status of the fetus. Exclusion testing whereby only embryos who have not inherited an affected allele are implanted without disclosing information regarding the *HTT* CAG repeat numbers in all tested embryos is an option for some families that do not wish to know the genotype of the asymptomatic at-risk parent.

X-Linked Diseases

In X-linked diseases, the mutant allele resides on the X chromosome. In X-linked recessive diseases, females are heterozygous carriers of the disease with one normal and one mutant allele and are typically not affected. Males receiving the mutant allele from their mothers are considered hemizygous with one mutant allele and no normal allele. All daughters of affected males are carriers of a mutant allele. A carrier female has a 25% chance of transmitting her normal allele to a son; a 25% chance of having an affected son; a 25% chance of having a daughter who carries the mutant allele; and a 25% chance of having a daughter who receives her normal allele. In the absence of a family history, an affected male can have a mutant allele that arose de novo as a new mutation during gametogenesis in the formation of the egg. Roughly one-third of all cases of X-linked disorders represent de novo new mutations with the absence of a family history. In these cases, the mother is not a carrier of a mutant allele and is not at risk for having subsequent affected children. In pedigrees associated with X-linked recessive conditions, typically only males are affected, and male-to-male transmission of the disease is not seen. In less frequent, X-linked dominant diseases, one copy of the mutant allele is sufficient to cause disease despite the presence of a normal allele. In these disease processes, females are affected, and in males with only a single mutant allele, these diseases are often lethal.

Duchenne Muscular Dystrophy

Duchenne muscular dystrophy (DMD; OMIM #310200) is a fatal X-linked recessive disorder characterized by progressive skeletal muscle wasting. The incidence of DMD is about 1 in 3500 male births, making it the most common severe neuromuscular disease in humans. Classic DMD presents in early childhood, with delayed motor skills or an abnormal gait. This is followed by progressive muscle weakness, calf hypertrophy, and grossly elevated serum creatine kinase (>10× normal) caused by degenerating muscle fibers. Progressive weakness initially affects the lower extremities, causing most DMD patients to require wheelchairs between 10 and 15 years of age. Continual degeneration and regeneration and inflammation of muscle eventually lead to the replacement of muscle tissue by adipose and connective tissue. Scoliosis is common and affects respiratory function. Chronic respiratory insufficiency develops in all patients. Cardiac disease is most commonly dilated cardiomyopathy (DCM) caused by cardiac fibrosis or rhythm and conduction abnormalities. Cardiorespiratory failure is the primary cause of death. Clinical management of patients with DMD is complex, requiring a multidisciplinary team approach. In addition to healthcare professionals, team members in the areas of psychosocial, gastrointestinal, pain, and speech and language are required. Although elevated serum creatine kinase (CK) levels can be measured from blood spots, NBS for DMD, enabling early identification of affected children and potential better outcomes from early intervention has not been implemented in the United States.

Because DMD is an X-linked recessive disorder, most carrier females are asymptomatic. Similar to other X-linked diseases, the varying degree of clinical manifestations among carrier females depends on the degree of inactivation of the X chromosome harboring the mutant *DMD* gene in various tissues where the DMD protein is expressed. Up to as many as 20% of carriers can display some symptoms of DMD. Most frequently observed is muscle weakness with elevated serum CK levels or cardiac involvement.

The *DMD* gene (Xp21.2) is complex and is one of the largest genes in the human genome, spanning 2.4 megabases. DMD contains 79 exons, representing less than 1% of the gene, and encodes a 14 kb mRNA. The gene has multiple tissue-specific promoters that transcribe various full-length dystrophin isoforms differing in their amino terminal sequences. The full-length protein product, dystrophin, contains 3685 amino acids; has a molecular weight of 427 kDa; and contains four distinct domains, including an actin-binding domain, a central rod domain with spectrin-like repeats, a cysteine-rich domain, and a unique COOH-terminal domain. Dystrophin is predominantly expressed in skeletal, cardiac, and smooth muscle.

DMD gene mutations are heterogeneous. Intragenic deletions encompassing one or more exons represent 70% of mutations and affect the translational reading frame of the protein leading to a truncated and nonfunctional protein. Duplications of one or more exons are observed in about 5% of patients. Point mutations account for the majority of remaining mutations, but small insertions, deletions, or splice site mutations are also detected.

Because of the tremendous size and diversity of mutations within the *DMD* gene, DNA testing for DMD presents a challenge for clinical laboratories. DNA testing confirms the diagnosis, identifies the pathogenic mutations segregating in the family, and enables targeted analysis for carrier testing of at-risk females as well as prenatal or preimplantation testing as requested. Deletion and duplication testing for all 79 exons is most frequently performed by multiplex

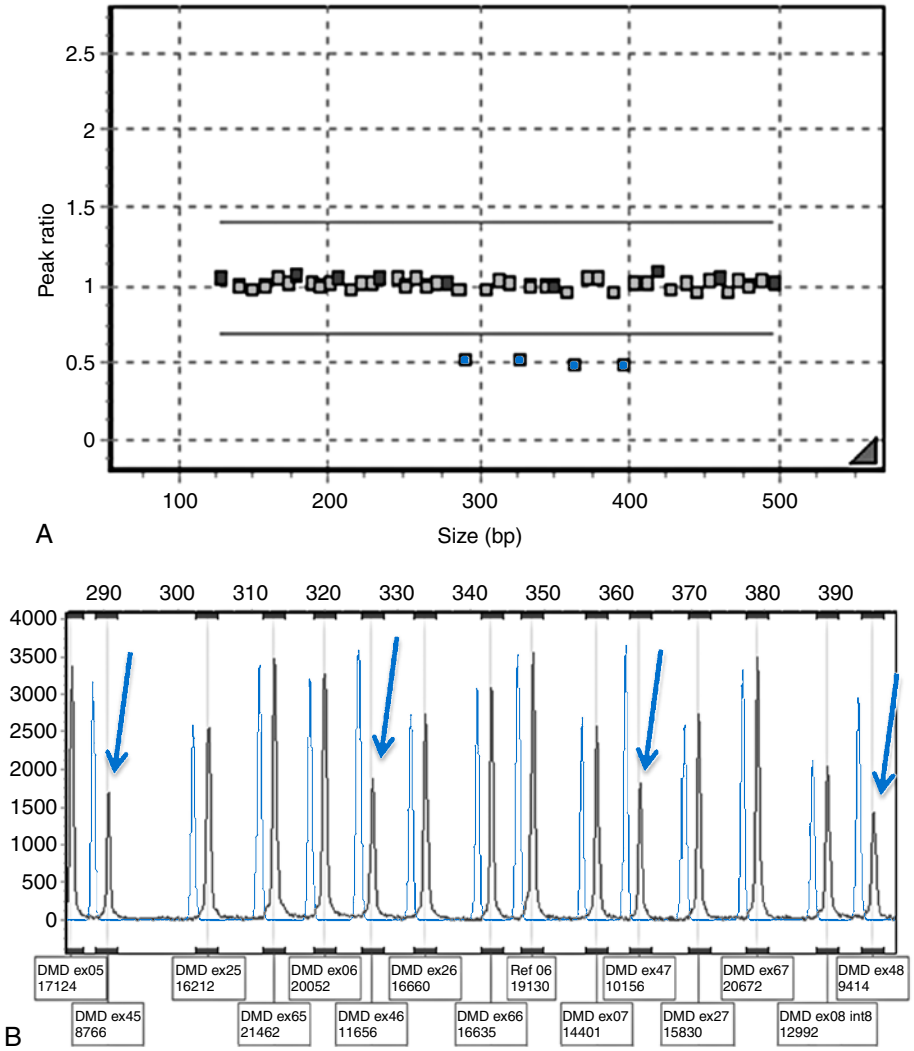


Fig. 49.2 Deletion/duplication analysis of the Duchenne muscular dystrophy gene in a female patient by multiplex ligation-dependent probe amplification (MLPA). (A) Wild-type (normal) alleles demonstrate a peak ratio of about 1. In this example, the patient is positive for a deletion of exons 45 to 48 *(blue squares)*. Note that the deleted exons show a peak ratio of approximately 0.5. (B) Electropherogram data showing the reduction in peak height (*y*-axis, relative fluorescence) for exons 45 to 48 *(arrows)* in one haplotype *(gray peaks)* relative to the other control haplotype *(blue)*.

ligation-dependent probe amplification assay (Fig. 49.2). This method determines whether each exon is present, deleted, or doubled compared with control DNA. Alternatively, microarray-based comparative genomic hybridization can be used for deletion and duplication screening. Sanger sequence analysis is typically performed for the remaining 30% of mutations. However more recently, targeted massive parallel sequencing (MPS) has been used for DMD analysis.

Although no curative treatment for DMD is available, innovative therapies are emerging. Recombinant adeno-associated virus vectors have been used to restore 90% of strength in *mdx* mice by delivery of shortened dystrophin gene constructs. Suppression of nonsense mutations causing premature truncation of the dystrophin protein generated by point mutations in about 10% to 15% of patients can be accomplished with some drugs. Alternatively, for the majority of patients with gene deletions or duplications, exon skipping using antisense oligonucleotide therapy targeting specific exons has been successful for exon 51. Clinical trials for these and other drugs are ongoing. Alternative therapeutic approaches and new targets linked to the pathogenesis of the disease are being evaluated including use of the patient's stem cells, gene editing by homologous recombination, nonhomologous end joining, and increased expression of dystrophin similar proteins.

MASSIVELY PARALLEL SEQUENCING

Massively parallel sequencing is a high-throughput DNA sequencing technology capable of generating data on a genomic scale in a short period of days. One of the most significant advantages of MPS for diagnostic testing is the ability to target all genes known to be associated with a specific diagnosis (or phenotype) in a single test that is comparable

TABLE 49.3 Commonly Ordered Massively Parallel Sequencing–Based Gene Panels

Panel Name	Included Disorders[a]	Targeted Patient Group
Aortopathy disorders	Ehlers-Danlos syndrome (I, II, and IV), Loeys-Deitz syndrome, Marfan syndrome	Individuals with disease affecting any aortic section
Breast and ovarian hereditary cancer	Hereditary breast and ovarian cancer syndrome	Individuals with a strong family history of breast and ovarian cancer; individuals with early onset of breast or ovarian cancer
Cardiomyopathy disorders	Dilated cardiomyopathy, hypertrophic cardiomyopathy, long QT syndrome	Individuals with a suspected diagnosis of a hereditary cardiomyopathy disorder
Expanded carrier screening	Numerous (>100)	Individuals planning pregnancy or prenatal reproductive partners
Hearing loss	Keratitis–ichthyosis–deafness syndrome, nonsyndromic hearing loss, Usher syndrome	Individuals with a suspected diagnosis of either syndromic or nonsyndromic hearing loss
Hereditary endocrine cancer	Multiple endocrine neoplasia type 1, multiple endocrine neoplasia type 2, Von Hippel-Lindau disease	Individuals with a family history of endocrine cancer; individuals with a personal history of endocrine cancer
Hereditary gastrointestinal cancer	Familial adenomatous polyposis, juvenile polyposis syndrome, Lynch syndrome	Individuals with a personal or family history of gastrointestinal cancer
Noonan spectrum disorders	Cardiofaciocutaneous syndrome, Costello syndrome, Noonan syndrome	Individuals with a suspected diagnosis of Noonan syndrome or a related disorder
Periodic fever syndromes	Familial Mediterranean fever, Majeed syndrome, Muckle–Wells syndrome	Individuals with a suspected diagnosis of a periodic fever syndrome
X-linked intellectual disability	Rett syndrome, Duchenne muscular dystrophy, ornithine transcarbamylase deficiency	Individuals with intellectual disability inherited in an X-linked manner

[a]This is a selected list of commonly included disorders and is not comprehensive.

in price and turnaround time to that of Sanger sequencing for a single-gene analysis. These targeted gene panels are the most commonly ordered clinical tests using MPS technology. Before the development of MPS, testing patients for causative mutations in multiple genes that cause a single syndrome was a very expensive and time-consuming process. Table 49.3 lists some commonly ordered MPS-based gene panels. MPS also enables the sequencing of very large genes at a reasonable cost (e.g., the *DMD* gene). The flexibility of MPS selection and library preparation can provide a comprehensive test of more than 100 genes (e.g., an X-linked intellectual disabilities panel) or a more targeted panel that analyzes a handful of genes associated with a specific phenotype (e.g., a hereditary gastrointestinal cancer panel).

WHOLE-EXOME SEQUENCING

Whole-exome sequencing (WES) refers to an MPS-based DNA sequencing method that specifically targets the coding regions (exons) and directly adjacent intronic regions for the majority of the approximately 20,000 genes known to exist in the human genome. On average, WES is able to identify an underlying genetic alteration in approximately 25% to 50% of patients (depending on the inclusion criteria). This makes WES an effective diagnostic tool for patients with a phenotype that suggests an inherited disorder that does not fit the clinical characteristics of previously described syndromes.

WES has been implemented widely in clinical diagnostic laboratories in the United States, and thousands of patients have been tested by this method. WES should be considered in patients with a phenotype suspected to be caused by a mutation in a single gene when known single-gene disorders have been eliminated. Careful consideration should always be given to the patient's presentation before determining whether or not WES is the appropriate test for a given phenotype. In many cases, a single gene test or a targeted MPS-based gene panel with multiple genes may be the appropriate first test to order.

Over the past several years, WES has proven to be an invaluable research tool in the discovery of the underlying genetic alterations for many Mendelian disorders. Identification of the underlying molecular alteration (and molecular pathogenesis) that results in a specific disease is the first step in developing treatment modalities. Because of the discoveries made by WES, the next decade will see a vast improvement in the treatment and underlying cause of many inherited disorders.

POINTS TO REMEMBER

- Autosomal dominate, autosomal recessive, and X-linked genetic diseases follow specific modes of inheritance, but imprinting disorders, and mitochondrial and complex diseases do not
- For most genes, pathogenic mutations are located throughout the gene
- For some disease genes, the type of mutation can predict the clinical phenotype and identify targeted therapy
- Diagnostic DNA testing is complicated by genetic and allelic heterogeneity
- Genetic counseling is an important component of patient care and management

GENETICS OF HEMATOPOIETIC MALIGNANCIES

Neoplasia of blood cells is often easier to diagnose and monitor than solid tumors because a simple blood draw usually allows access to tumor cells. Genetic aspects of hematologic malignancies include numerical and structural chromosomal abnormalities and smaller-scale genetic changes found within single genes.

Cytogenetics is very useful in the diagnosis of hematologic disorders. Conventional cytogenetics refers to visual analysis of a karyotype composed of a complete set of metaphase-arrested chromosomes typically stained by Giemsa (G-banding). G-banding highlights light and dark zones corresponding to A-T-rich and G-C-rich regions of chromosomes. This decades-old technique remains very important for the comprehensive bone marrow evaluation of patients with acute leukemia or a myelodysplastic syndrome (MDS). G-banded karyotyping enabled the discovery of many of the recurring translocations characteristic of specific hematologic diseases. Karyotyping provides a genome-scale perspective and is well suited for uncovering large-scale chromosomal aberrations (resolution >5 Mb) including reciprocal translocations, large deletions, and aneuploidy. Cryptic rearrangements and small insertions or deletions are likely to be missed by conventional karyotyping and require more sensitive techniques for detection. Karyotyping is a labor-intensive process that requires considerable expertise for performance and interpretation. In addition, because cell growth in culture is required, turnaround times often range from 3 to 7 days.

NUMERICAL CHROMOSOMAL ABNORMALITIES

Numerical chromosomal abnormalities are found in hematopoietic malignancies. For example, in B-cell acute lymphoblastic leukemia (ALL), the malignant cells can have variable numbers of chromosomes. Cases with greater than 50 chromosomes are referred to as hyperdiploid, while those with less than 45 chromosomes are designated hypodiploid. B-ALL with hyperdiploidy is seen in 25% of childhood disease and decreases in incidence with age. The prognosis is very good, and achievement of cure is highly likely with standard therapy. The number of chromosomes may be less important than the specific chromosomes involved. Patients with three copies of chromosomes 4, 10, and 17, so-called "triple-trisomy ALL," have an excellent prognosis. Hypodiploid B-ALL is seen in less than 5% of patients overall. Prognosis is generally poor and appears to be correlated to the degree of chromosome loss, with near-diploid (44 chromosomes) and high-hypodiploid (40 to 43 chromosomes) patients faring relatively well. Low-hypodiploid (32 to 39 chromosomes) cases have a high frequency of mutations in the tumor suppressor gene, *TP53*.

Intrachromosomal amplification of parts of chromosome 21 can also occur as a primary genetic event in B-cell ALL and arises from the effect of "gene dosage" due to copy-number alterations of potentially hundreds of linked genes. This abnormality in B-ALL is associated with older age at presentation and a low white blood cell count. It also fares poorly when treated according to standard-risk protocols and benefits from high-risk ALL therapy.

RECURRENT TRANSLOCATIONS

Many hematopoietic malignancies have structural chromosomal abnormalities, including balanced and unbalanced translocations and large-scale structural changes such as deletions or duplications. Some of these abnormalities are characteristic of certain disease subtypes and are thus repeatedly, or recurrently, found in different individuals with the same disease. Recurrent genetic abnormalities of all types are found across the spectrum of hematopoietic malignancies.

Chromosomal abnormalities may be detectable by conventional cytogenetics or may require more specialized molecular techniques. Fluorescence in situ hybridization (FISH) is a powerful tool that overcomes several limitations of conventional G-banded karyotype. FISH uses fluorescently labeled DNA probes designed to hybridize to specific sequences on metaphase or interphase chromosomal preparations. Resolution is improved to approximately 2 Mb. FISH allows the detection of cryptic chromosomal abnormalities, such as translocations or deletions, which may not be identifiable by routine karyotype. Many FISH probes are commercially available for recurrent abnormalities. Paraffin-embedded formalin-fixed samples can be used because FISH assays do not require viable cells or growth in culture. Break-apart FISH probes allow for identification of gene rearrangements without a priori knowledge of the translocation partner, of which there may be numerous possibilities. Global assessment of copy number changes is not possible by FISH because only a limited number of probes are utilized.

While some genetic lesions are disease defining, many have utility in diagnosis, prognosis, and clinical management. The nomenclature for chromosomal abnormalities is highly developed. Fig. 49.3 details the general format used to represent chromosomal abnormalities. As an example of a hematopoietic

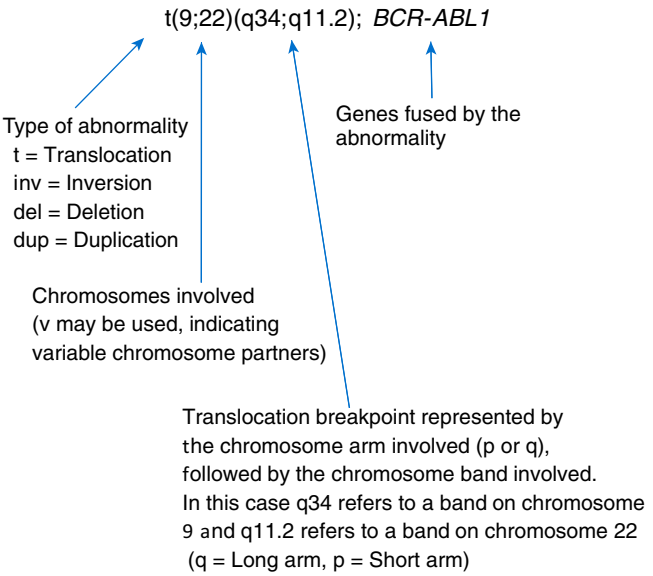


Fig. 49.3 Format used to represent chromosomal abnormalities.

translocation, consider the 9;22 translocation found in some ALL and almost all chronic myeloid leukemias (CML).

B-cell ALL with t(9;22)(q34;q11.2) is a high-risk disease characterized by the production of a fusion protein with constitutively active ABL1 tyrosine kinase activity. This is the most common recurrent genetic abnormality in adult B-cell ALL patients and confers poor prognosis among all age groups. Patients may benefit from tyrosine kinase inhibitor (TKI) therapy in addition to traditional high-dose chemotherapy. While t(9;22)(q34;q11.2) is reliably detected by FISH, the type of *BCR-ABL1* fusions present should be confirmed by quantitative RT-PCR for the purposes of ongoing monitoring during treatment and minimal residual disease detection. The p190 kDa fusion protein is common in pediatric disease, while adult ALL patients typically have either the p190 form or the larger p210 form. The latter is also seen in almost all cases of chronic myelogenous leukemia (CML).

The 9;22 translocation was first discovered in 1960 as a recurrent chromosomal abnormality in patients with CML. This was followed 13 years later by finding a consistent reciprocal translocation between 9q34 and 22q11.2, resulting in a der(22q) (now known as the Philadelphia chromosome) found in virtually all cases of CML. Another 10 years would pass before the precise breakpoints were cloned, and it was shown that the translocation results in juxtaposition of the *ABL1* proto-oncogene on 9q34 to *BCR* on 22q11.2 with formation of a novel *BCR-ABL1* fusion gene. This results in constitutive ABL1 protein tyrosine kinase activity and dysregulated cellular proliferation. Documentation of t(9;22)(q34;q11.2); *BCR-ABL1* fusion is necessary for diagnosis and is readily demonstrated by metaphase cytogenetics, FISH and/or RT-PCR. Distinct *BCR-ABL1* fusion transcripts, which correspond to variably sized fusion proteins, are encountered based on the translocation breakpoints.

Quantitative RT-PCR is important in the clinical diagnosis and management of CML (Fig. 49.4). Routine molecular monitoring of response to TKI therapy in CML is the standard of care. Testing should be performed at diagnosis and periodically throughout the course of treatment. A three-log or greater reduction in *BCR-ABL1* transcript levels within 18 months of initiation of therapy is designated a major molecular response (MMR) and portends a high likelihood of a favorable long-term outcome. Quantitative PCR assays, therefore, should be designed and validated with clinically important levels of sensitivity in mind.

Patients with either CML or Ph+ B-cell ALL are treated with *ABL1* TKIs. A subset of patients may develop acquired mutations in the tyrosine kinase domain of *ABL1* that impede drug binding. When this occurs, patients will manifest signs of resistance typically evident in rising levels of *BCR-ABL1* transcripts detected by a routine quantitative RT-PCR. Many mutations have been identified, and dozens are recurrent across the spectrum of patients treated with TKIs. The identification of one or more mutations is important for clinical treatment because the sensitivity of *BCR-ABL1* to the available TKIs differs depending on which mutation is present.

In a subset of patients, particularly those who have been treated with multiple sequential TKIs, more than one mutation may arise in the same clone (i.e., on the same *BCR-ABL1* allele). These are commonly referred to as compound mutations, and they may impart resistance profiles that differ substantially from those seen when either mutation alone is present. Compound mutations tend to center around a dozen or so key residues that are frequently found mutated alone. Thus the typical sequence of events in the development of a compound mutation is likely similar in most patients. First, the acquisition and outgrowth of an initial single mutation (X) in a key residue occurs. This is followed by the acquisition of a second mutation (Y) in a clone harboring mutation X, thereby generating a new clone with two mutations (X + Y). At this time, both clone X and clone X + Y coexist in the patient, although clone X + Y may quickly dominate due to enhanced resistance properties. Strategies for analyzing and evaluating the complex clonal architecture in patients with *BCR-ABL1*-positive leukemias who have developed resistance mutations are often imperfect. Sanger sequencing is

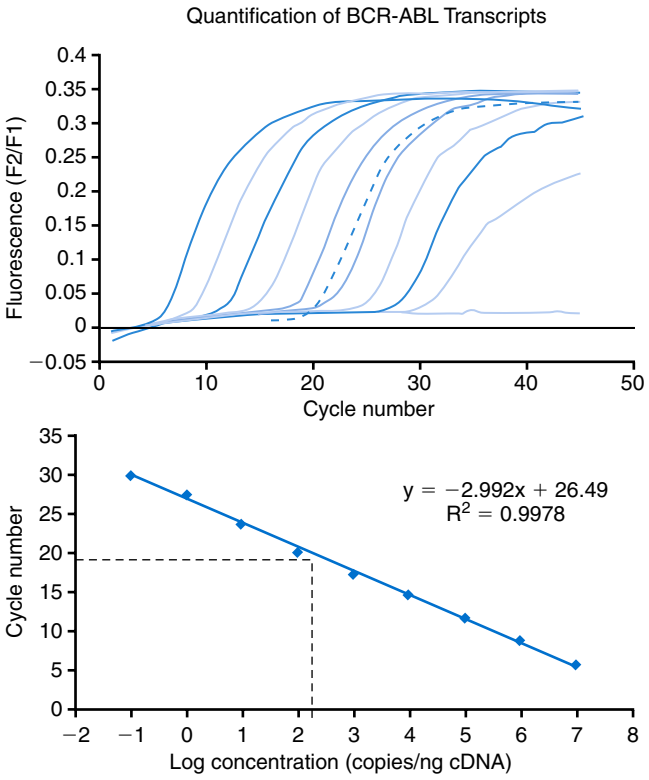


Fig. 49.4 An example of quantitative reverse transcription polymerase chain reaction for enumeration of *BCR-ABL1* transcripts. The upper panel shows amplification curves for standards containing serial log dilutions of *BCR-ABL1* fusion transcripts ranging from 10^0 to 10^9 copies *(solid lines)* and an unknown patient sample *(dotted line)*. The lower panel shows the standard curve for the reaction that plots cycle number versus the log of the template concentration. The quantification cycle is then used to quantitate the *BCR-ABL1* transcript copy number in the patient sample. For normalization purposes, total *ABL1* transcripts can be analyzed separately in an identical fashion to generate a copy number ratio of *BCR-ABL1* to *ABL1* referred to as the *normalized copy number*. The normalized copy number can then be converted to the international standard scale.

not typically useful because the data generated consist of a mixture of sequences and clonal relationships are mostly lost. Other techniques, such as massively parallel sequencing, offer more promise because the sequence data comprises single-molecule sequences and not mixtures, and therefore the clonal architecture is retained. Despite dramatic improvement in long-term survival, treatment of CML patients with TKIs does not typically result in cure.

Although we have focused only on the 9;22 translocation, many other recurrent translocations are found in other hematopoietic malignancies, including T-cell ALL, Burkitt lymphoma, follicular lymphoma, diffuse large B-cell lymphoma, mantle cell lymphoma, extranodal marginal zone lymphoma, multiple myeloma, anaplastic large cell lymphoma, AML, and MDS.

GENE MUTATIONS IN HEMATOPOIETIC MALIGNANCIES

Many different types of genetic abnormalities may be present in a neoplasm that are not detectable by karyotyping or FISH-based assays, but rather by molecular techniques such as PCR and sequencing. In some disorders, such mutations may be the only detectable genetic abnormality and are likely sufficient in and of themselves to promote development of the neoplasm. In other cases multiple single-gene mutations are present and may occur along with structural chromosomal abnormalities and/or translocations detectable by classical karyotyping or FISH and with epigenetic changes that must be queried by yet additional tests. Furthermore, copy number changes may be present and typically require additional, array-based technologies for detection.

Mutations can result from exposure to mutagenic agents (i.e., ionizing radiation or chemotherapeutic agents) or randomly during the process of DNA replication during cell division. Thus mutation rates at any particular nucleotide position depend on a variety of factors, including nucleotide stability due to genome context (e.g., homopolymer tracts or repetitive sequences) or chemical modification (e.g., methylation), the error rate of DNA synthesis, and the accuracy and effectiveness of DNA repair mechanisms. Mutations may therefore leave a signature with clues about the original mechanism that resulted in the error.

Mutations that correlate most closely with aging include C to T transition mutations that arise in CpG dinucleotides due to spontaneous deamination of 5-methylcytosine to thymine. Indeed, there is a positive correlation between the number of such mutations and the age of the patient at the time of cancer diagnosis. Chemotherapeutic alkylating agents such as cyclophosphamide also have a characteristic mutation signature involving C to T mutations. Interestingly, and perhaps not unexpectedly, there is also a positive correlation between the number of divisions of normal tissue stem cells and the rate of malignancies among tissue types. This finding suggests that the majority of malignant processes, including those involving the hematopoietic system, are derived from the evolution of chance mutations that occur during the process of DNA replication and not from some environmental or toxic insult. Mutations that arise in a stem cell may be noncontributory to the process of malignant transformation. However, they remain a component of the malignant clone and are detectable if queried in an appropriate manner; for example, by whole genome sequencing. These have come to be known as "passenger" mutations and can sometimes be segregated from true disease "driver" mutations that occur in coding regions, at splice sites, or in RNA genes because they are not recurrent over large disease-specific patient cohorts. As more data are generated, the differentiation between passenger and driver will become easier.

Mutations in discrete cellular pathways and regulatory mechanisms are important in hematopoietic malignancies, particularly of myeloid origin. Although many genes may be recurrently mutated, there is a large degree of overlap among the various disease entities and there are few, if any, true disease-defining mutations. Generally the affected pathways may be classified into those involving signal transduction, epigenetic modifiers, RNA splicing, sister chromatid separation during meiosis and mitosis, transcription factors, and chromatin modifiers. The order in which mutations arise also appears to have consequences with respect to the overall clinical and pathophysiologic features of a disease process. There are many important genes in cellular and regulatory pathways with mutations that contribute to malignancy; only *BRAF* and *TP53* are covered here as examples.

Signal Transduction Pathways

Signal transduction pathways mediate cellular communication cell surface receptors. These receptors are often linked to tyrosine and serine/threonine kinases that serve to propagate the message, ultimately to the nucleus where the modulation of gene expression results in some change in cellular activity. Mutations that occur in signaling pathways tend to affect tyrosine kinase mediated signal transduction and often result in constitutively active tyrosine kinases. Activating mutations like these tend to result from recurrent so-called "hot spots" often in regulatory or enzymatic (i.e., tyrosine kinase) domains of these important signaling molecules.

BRAF p.Val600Glu mutations are present in nearly all cases of classic hairy cell leukemia and are absent in other B cell lymphoproliferative disorders, including those that may cause diagnostic confusion with hairy cell leukemia. The mutation promotes constitutive signaling through a multienzyme tyrosine kinase pathway that start at a cell surface receptor and ends with altered mRNA expression. Hairy cell leukemia is an indolent B cell lymphoproliferative disorder typically diagnosed based on immunophenotypic findings assessed by flow cytometry and/or immunohistochemistry along with characteristic neoplastic cells with "hairy" cytoplasmic projections. Although the diagnosis is typically straightforward based on the clinical and laboratory findings, there are occasional cases that can cause diagnostic difficulty. For these cases, identification of the presence of the *BRAF* p.Val600Glu mutation is diagnostically important. BRAF inhibitor therapies are available and have been used with some success in other tumors

with BRAF- (p. Val600Glu) activating mutation such as melanoma. In patients with hairy cell leukemia, the inhibition of BRAF and other protein kinases with drugs appears effective at inducing apoptosis.

TP53

The tumor suppressor p53 regulates a variety of critical anticancer functions and is nicknamed the "guardian of the genome." Its functions include regulation of cell cycle entry and exit at various checkpoints, maintenance of cellular senescence, modulation of autophagy, and control of apoptosis. Patients with germline *TP53* mutations (Li-Fraumeni syndrome) have a very high risk for the development of cancers, often sarcoma or breast cancer. In general, dysregulation of p53-dependent signaling networks is a common feature of aggressive malignancies, including those of the hematopoietic system. Genetic abnormalities that result in loss of p53 activity include chromosomal deletions of the *TP53* gene locus on chromosome 17 (17p), loss of function mutations in the *TP53* gene, and epigenetic suppression of *TP53* expression.

In myeloid malignancies, *TP53* mutations are seen most frequently in patients with AML, MDS, and primary myelofibrosis. *TP53* mutations are loss-of-function variants, including insertions, deletions, nonsense mutations, and splice site mutations that result in the production of truncated protein products. In AML, *TP53* mutations may be associated with complex (three or more) karyotypic abnormalities and have an extremely poor prognosis. *TP53* mutations are commonly seen in a subtype of myeloid cancer that develops following exposure to cytotoxic chemotherapy for another malignancy called therapy-related AML or therapy-related MDS. The myeloid clones that expand following chemotherapy represent rare hematopoietic stem and progenitor cells with age-related *TP53* mutations, rather than newly induced *TP53* mutated clones.

Massively Parallel Sequencing in Hematopoietic Malignancies

Massively parallel sequencing as a testing strategy is becoming more commonplace in the clinical hematology laboratory and offers numerous advantages to traditional techniques. Currently, testing of gene panels, rather than whole exome or whole genome sequencing combine a relatively small number of gene targets, at least compared to the breadth of sequence covered by whole genome or whole exome sequencing. Panels are composed of genes that are recurrently mutated in a particular disease or class of disease. For example, a panel may comprise a few dozen genes frequently mutated in myeloid malignancies. To enhance the representation of specific gene targets for sequencing, PCR- or hybrid capture-based enrichment is commonly employed. The end result is very high coverage (usually greater than 1000×) of a clinically important set of targets. This degree of coverage depth yields test sensitivities that may be as low as 1% to 2% variant allele frequency, far better than what is possible by Sanger sequencing, but not as good as what is achievable by targeted allele-specific PCR assays or digital PCR.

The spectrum of genetic abnormalities present in hematopoietic malignancies requires the use of many complex testing strategies for comprehensive evaluation. These testing strategies include detection of translocations, copy number changes, gene mutations, and epigenetic abnormalities. As the number and type of clinically actionable genetic variants continue to grow, the laboratory evaluation of these patients will continue to be a multidisciplinary effort. Traditional clinical, laboratory, and pathologic methods remain central to patient diagnosis and management and are unlikely to be supplanted in the near future by genetic testing.

CIRCULATING TUMOR CELLS AND CIRCULATING TUMOR CELL FREE DNA

Classic tissue biopsies or surgical resections are invasive procedures that capture only a single snapshot in the evolution of a cancer in a patient. In contrast, a blood-based test or "liquid biopsy" has the potential to characterize the evolution of a solid tumor in real time by extracting molecular information from CTCs or ctDNA. Molecular characterization of CTCs and ctDNA holds considerable promise for the identification of therapeutic targets and resistance mechanisms, and for real-time monitoring of the efficacy of systemic therapies. The major potential advantage of CTC and ctDNA analysis is that they are minimally invasive and can be repeated during patient follow-up. With respect to the clinical laboratory, the development of targeted molecular assays as companion diagnostics, for disease monitoring, and even for early cancer detection are all possibilities at various stages of progression and during treatment regimens.

Cancer genomes are not static but change over the course of therapy. The term *liquid biopsy* refers to blood-based cancer testing and often involves detailed molecular analysis of the tumor from circulating genetic material in the peripheral blood. This genetic information is derived mainly from CTCs and ctDNA (Fig. 49.5). Liquid biopsy offers a simple and noninvasive insight into a patient's cancer using blood samples at any time.

The blood biopsy as a diagnostic, prognostic, and theranostic tool is appealing because it is minimally invasive and is easily performed in a serial manner. However, there are still several concerns about its routine clinical use: (1) numerous technologies are available for the detection of cancer blood biomarkers, (2) more specific biomarkers are needed to assess treatment response and for evaluation of CTCs and ctDNA at specific stages of the cancer for accuracy, (3) well-designed comparison studies between liquid biopsy (CTCs and/or ctDNA) with conventional tissue biopsy samples are needed for validation, (4) it can be challenging to control the pre-analytical phase to obtain robust and reproducible results in some assay systems, (5) currently available techniques are costly and not easily reimbursable by healthcare, particularly CTCs, (6) multiple ctDNA assays have been developed and are available but none are FDA-cleared, while for CTCs, only one assay (CellSearch, Menarini Inc.) has been FDA-cleared, and (7) the TATs vary depending on the assay and can be longer than a week.

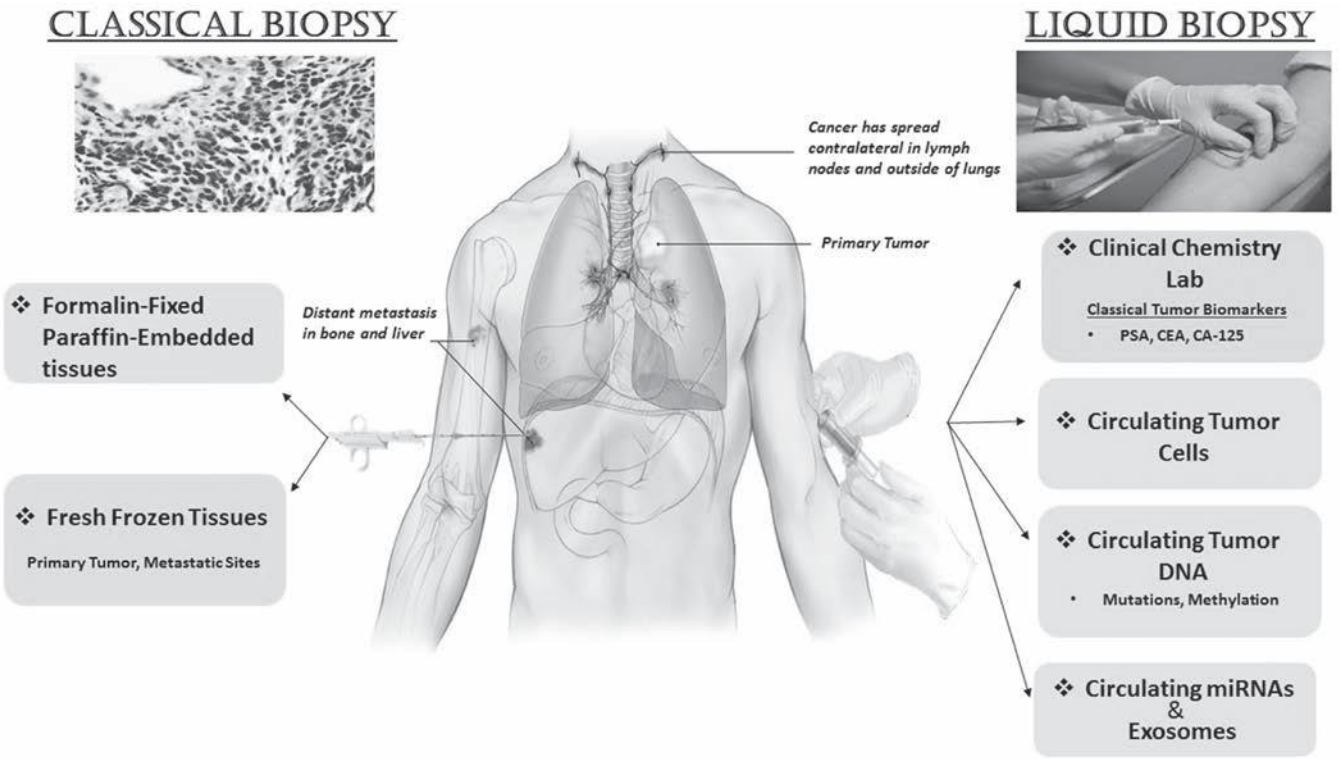


Fig. 49.5 Classic versus liquid biopsy approaches.

Circulating Tumor Cells

CTCs are rare cells that originate from primary or metastatic tumors that have managed to get into the circulation and that may extravasate to different organs. Only a small fraction of CTCs will develop into "true" metastasis. CTCs are a major player in the liquid biopsy cellular approach and may provide real-time information on a patient's disease status. Cancer metastasis is the main cause of cancer-related death, and dissemination of tumor cells through the blood circulation is an important intermediate step that also exemplifies the switch from localized to systemic disease spread occurring.

The FDA has cleared CTC analysis for breast (2004), colorectal (2008), and prostate (2008) cancer using the CellSearch system based on the critical role that CTCs play in metastasis. However, this system is approved only for prognostic information, not for treatment decisions. CTCs are targets for understanding tumor biology and tumor cell dissemination in humans. Their molecular characterization offers an exciting approach to understanding resistance to established therapies and elucidating the complex biology of metastasis. Further research on the molecular characterization of CTCs should contribute to a better understanding of the biology of metastatic development in cancer patients and the identification of novel therapeutic targets. This approach may provide individualized targeted treatments and spare cancer patients unnecessary and triage into ineffective therapies.

Circulating Tumor Cells: Analytical Techniques

The isolation and further analysis of highly pure CTCs are difficult because these cells are extremely rare in the peripheral blood, particularly in the early stages of cancer, therefore enumeration-based assays have been inconsistent and not robust. Typically, CTCs are isolated along with a background of peripheral blood mononuclear cells. The combination of high-throughput and automated CTC isolation technologies with generally accepted and validated downstream molecular assays is necessary for the routine use of CTC-based diagnostics.

CTC analysis includes isolation/enrichment, detection, enumeration, and molecular characterization. CTCs are highly heterogeneous, even within the same individual. This has been verified by using reliable single-CTC isolation followed by massively parallel sequencing to clearly reveal differences among CTCs. MPS technologies reveal that intra-tumor heterogeneity may reflect tumor evolution and adaptations that hinder personalized medicine strategies developed from results on single-tumor biopsy samples. CTCs present in blood provide a better picture for individualized treatments because they represent cells that have disseminated from both the primary tumor and cells from secondary metastases.

Circulating Tumor Cells and Breast Cancer

The clinical significance of CTCs has been extensively evaluated in patients with breast cancer. Many clinical studies have shown that CTC detection is associated with overall survival and progression-free survival in both early and metastatic disease. The detection of CTCs is a reliable prognostic factor in patients both with early-stage and metastatic breast cancer. Metastatic breast cancer patients with basal counts of 5 CTCs/7.5 mL of blood or greater have a poor prognosis, and enumeration of CTCs during treatment

predicts the progression of disease earlier than conventional imaging tests. In early breast cancer, CTC numbers are low, and molecular assays are most successful for detection. Quantitative RT-PCR (RT-qPCR) for cytokeratin expression in the peripheral blood of node-negative breast cancer patients is of prognostic value. Elimination of these cytokeratin mRNA-positive CTCs during adjuvant chemotherapy reflects successful treatment.

The main goal of adjuvant therapy is to prevent distant recurrences by targeting residual disseminated tumor cells. However, almost 70% of deaths in patients with early breast cancer occur after 5 years, showing that residual disease can be dormant for very long periods. Differences between primary tumors and CTCs could be of crucial importance for therapeutic decisions. In breast cancer, molecular characterization of CTCs may help identify therapeutic targets and resistance mechanisms and to stratify breast cancer patients.

In breast cancer, anti-HER-2 therapies are prescribed according to the HER-2 status of the primary tumor, as assessed by immunohistochemistry or FISH. However, a continuously growing body of evidence indicates that CTC HER-2 status can be different from that of the corresponding primary tumor. Moreover, CTC HER-2 status can change over time, particularly during disease recurrence or progression.

The HER-2 status of CTCs was correlated to the clinical response of HER-2-targeted therapies in a "liquid biopsy trial" in 2012. Patients who were HER-2 negative in the primary tumor but positive for HER2 in CTCs were randomized into two groups; one group received the anti-HER-2 drug trastuzumab, while the other received a placebo. After trastuzumab administration, the risk of disease recurrence was reduced, and the median disease-free survival was longer.

Endocrine treatment is the preferred systemic treatment in metastatic breast cancer patients who have an estrogen receptor (ER)-positive primary tumor or metastatic lesions. However, 20% of these patients do not benefit from this therapy and demonstrate further metastatic progress. A possible explanation for failure of endocrine therapy might be the heterogeneity of ER expression in CTCs. Similar to the HER-2 story, there is a growing body of evidence that hormone receptor status can change over time, especially during disease recurrence or progression in breast cancer patients. In this context, reevaluation of hormone receptor status by molecular characterization of CTCs is a strategy with potential clinical application. Optimal individualized treatment can be determined by characterizing ER status in CTCs and comparing it to the primary tumor.

Overall, CTC analysis is highly correlative to systemic disease and outcome in breast cancer patients. Monitoring of CTCs can predict early subclinical metastasis in disease-free patients and monitor progression during and after treatment. Many clinical studies support the potential utility of CTC detection in breast cancer. In addition, CTC analysis has been FDA approved for prostate and colon cancer diagnosis only, and shows promise in other cancers, including lung, melanoma, ovarian, pancreatic, head and neck, liver, and bladder. CTC detection in disease-free post-surgical resection can be used to determine subclinical recurrence during follow-up in early and late stages of melanoma.

CIRCULATING TUMOR CELL FREE DNA

ctDNA isolation, detection, and analysis have significantly improved over the past two decades. ctDNA can be actively or passively released by tumor cells and enter different fluid compartments of the body such as lymphatics, urine, blood, semen, saliva, and cerebral spinal fluid. ctDNA may also be from CTCs that are disrupted in the bloodstream. ctDNA may have physiological functions that influence normal cells, particularly those adjacent to the tumor. The release of ctDNA may have direct effects on tumor microenvironment immune cells, as well as distant organ sites.

Types of ctDNA

Multiple forms of ctDNA exist in the blood of cancer patients (Fig 49.6). The most frequent ctDNA results reported are mutations. Tumor mutations in several cancers have drawn more attention because of the availability of respective targeted therapies that are FDA and CLIA approved in clinical oncology. However, in addition to mutations (single-nucleotide variants, small insertions and small deletions), there are additional types of genomic aberrations in individual cancers with potential utility. These potential targets include methylation of gene promoter regions of coding and noncoding genomic sequences, microsatellite loss of heterozygosity, DNA integrity, gene fusion, copy number variation, and cancer viral DNA.

These different types of ctDNA vary among individual cancer types, although some genomic aberrations can overlap, particularly with cancers of similar embryonic or carcinogenic origin. The clinical utility of one type of ctDNA in a specific cancer may not be the same for other cancers. For example, *BRAF* mutations are found in several cancers, including melanoma, colorectal, and lung, thyroid, and non-small cell lung cancer, but outside of melanoma they are only partially responsive in a subset of patients to anti-BRAF small molecule inhibitor therapies. BRAF inhibitors in melanoma have the highest response overall but duration is usually less than 2 years at best. Thus ctDNA BRAF mutations may potentially be useful for several types of solid tumor cancers. Several epidermal growth factor receptor (EGFR) mutation ctDNA assays have been FDA approved and are used to triage anti-EGFR treatment and monitor non-small cell lung cancer patients during treatment. Of all the ctDNA assays approved to date, EGFR mutation assays have been the most widely used and informative particularly when used with the different EGFR mutation-targeted therapies in carcinomas. This will in the future set precedent to develop new theranostic mutations of specific cancers.

Tumor-Specific ctDNA Mutations

Tumor-specific mutations in ctDNA are most commonly studied in cancer patients. Tumor mutations in ctDNA become

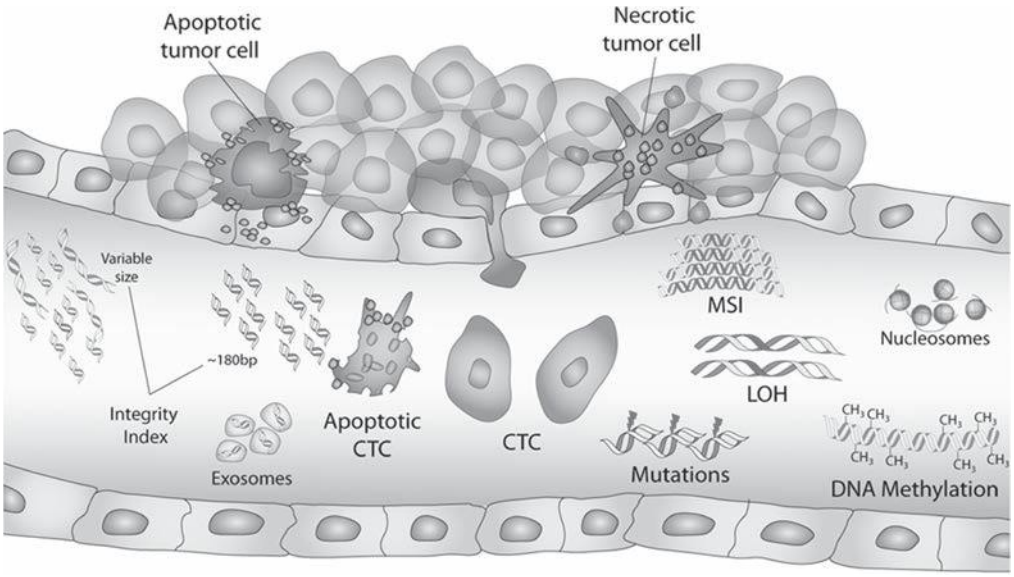


Fig. 49.6 Schematic representation of circulating tumor cell free DNA (ctDNA) in the systemic circulation. ctDNA arises from apoptosis or necrosis of primary and metastatic solid tumors and from degradation of circulating tumor cells *(CTC)*. In the circulation, ctDNA is of variable size, can be quantified by an integrity index, and is often enriched with multiples of 180 bp fragments that wind around nucleosomes. In the circulation, ctDNA may be contained within vesicles, attached to nucleosomes or free in solution. Analysis of ctDNA can identify tumor mutations, microsatellite instability *(MSI)*, loss of heterozygosity *(LOH)*, and methylation patterns. Apoptosis of normal cells also results in circulating DNA that dilutes the variant fraction of abnormal DNA.

increasingly important when the mutations have approved targeted therapies as companion diagnostics and are associated with therapeutic response. ctDNA mutations can easily be assessed with high sensitivity and specificity by qPCR or MPS. Most mutations can now be detected in ctDNA as well as tumor tissue. Although concordance is still a logistic issue.

Assessment of ctDNA mutations in blood during treatment and follow-up provide information to the clinician about new mutations that were not present in the original tumor or start of therapy. New strategies for initiating targeted therapies based on ctDNA analysis (Fig. 49.7) may replace biopsy of metastases for new gene mutations arising during follow-up or use in situations where biopsy is not practical. Currently, clinical trials are underway to validate this logistic.

Methylated ctDNA

Methylated ctDNA is detectable in cancer patients and has utility in both early detection and treatment monitoring. Typically, gene promoter regions contain CpG sites that are hypermethylated or hypomethylated to control gene transcription. In cancer, the promoter regions of tumor suppressor genes may be hypermethylated to limit transcription, and oncogene transcription may be increased by hypomethylation. In the blood of cancer patients, methylated CpG ctDNA can be detected and used for prognosis and as biomarkers of treatment response. During tumor progression, gene promoters in tumor cells typically become hypermethylated. Normal cell regulatory genes can also change in methylation status and be used for methylated ctDNA analysis. Currently, only one assay for methylation of Septin gene is FDA approved for the detection of colon cancer.

Viral ctDNA

Several cancers have an etiology strongly linked to viral DNA integration. In these cancers, ctDNA can be tested for specific cell-free viral DNA. The advantage of assessing viral ctDNA is the high specificity and detection sensitivity because of the unique viral sequences. The most notable and extensive assessment for clinical utility of viral ctDNA is Epstein-Barr virus (EBV) detection in nasopharyngeal cancer, one of the most common malignancies in southern China and Southeast Asia. Patients with metastatic disease have a poor prognosis. EBV ctDNA is associated with treatment outcome prediction. Plasma EBV ctDNA is useful in screening early-stage nasopharyngeal cancer and is relatively low in cost. Radiotherapy and combination chemoradiotherapy are major treatment modalities, whereby in the setting of these treatments, plasma EBV ctDNA levels can help assess treatment response and outcome. Additional examples of viral ctDNA include human papilloma virus for head and neck cancer, and cervical cancer, and hepatitis virus B for hepatocellular carcinoma.

ctDNA Integrity

Another approach to evaluating ctDNA is to examine its size, often referred to as "integrity." DNA fragment size can be used to determine the origin of cell-free DNA from either apoptotic or necrotic cells. Programmed cell death under normal conditions is accompanied by an organized degradation of the DNA, which results in small and uniform DNA fragments (approximately 140 to 200 bp). In contrast, cell necrosis, which is a frequent event in solid tumors, generates a wide spectrum of DNA fragment lengths, including those greater than 200 bp as a consequence of digestion. The

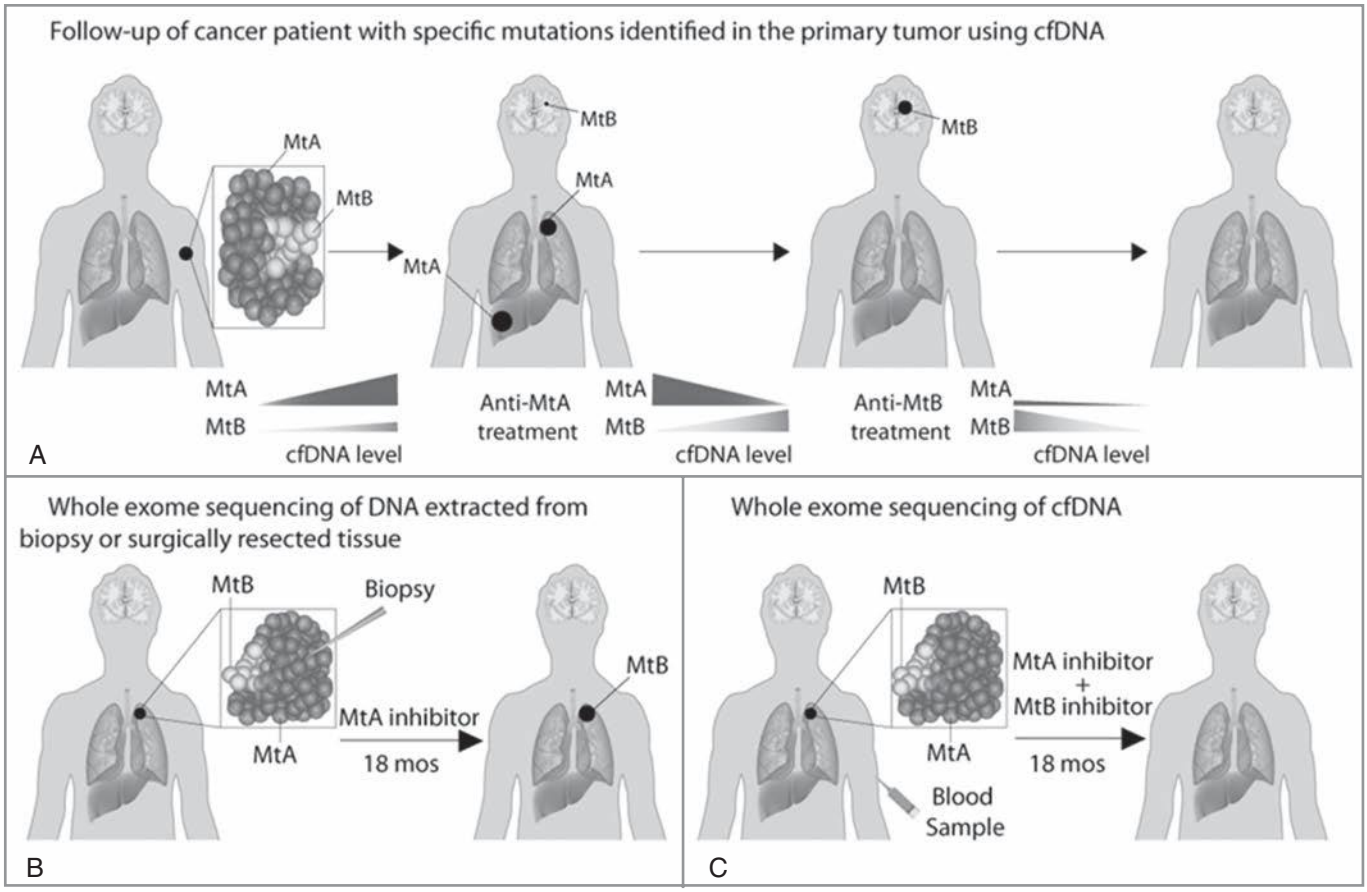


Fig. 49.7 Examples of circulating tumor cell free DNA analysis in cancer patients with multiple mutations. (A) A melanoma patient presents with a heterogeneous primary tumor with both mutation profile A *(MtA)* and mutation profile B *(MtB)*. Initial circulating free DNA *(cfDNA)* assessment indicates a higher level of MtA as a consequence of being the predominant tumor subclone. After starting with anti-MtA therapy, MtA clones decrease, while MtB may increase. If the MtB burden increases, anti-mtB treatment can be initiated, potentially clearing both clones by personalized treatment. (B) A resected tumor of a patient with non-small cell lung carcinoma is analyzed, revealing the targetable mutation L858R in *EGFR* but missing a minor clone harboring *BRAF* V600E. After 1 year of anti-EGFR treatment, the patient relapses with a tumor harboring *BRAF* V600E. (C) Whole genome sequencing of cfDNA identifies both L858 *EGFR* and an underrepresented subclone with a V600E *BRAF* mutation. In this diagnostic setting, both targetable mutations could be considered for treatment selection and potential cure.

increased DNA integrity of cell-free DNA in different cancers can be quantified as an integrity index that compares long to short cell-free DNA, such as ALU sequences.

Multimarker ctDNA Assays

The approach of multimarker versus single-marker assays is clearly essential in most cancers. Because of tumor heterogeneity and continual genomic/epigenomic changes during tumor progression, multiple-marker assays are often necessary for diagnosis and monitoring. Furthermore, therapeutics can modify the molecular profile of tumors, particularly when drug resistance develops. Most patients with distant metastasis have inter- and intra-tumor molecular heterogeneity between clinical and subclinical metastasis. Multiple molecular biomarkers and platforms may be necessary to study different genomic/epigenomic aberrations and/or multiple targets efficiently within the tumor load of a cancer patient. Combination drug therapy is often required because monotherapy in many advanced-stage cancers is not sufficient to control long-term tumor progression as in cutaneous melanoma, breast cancer, and colorectal cancer. Because signal pathways can change during tumor progression and resistance, more complete profiling of specific gene aberrations may be needed. Cancer genomic changes are complex. It is not surprising that multiple biomarkers and assays are needed for adequate patient monitoring and understanding disease progression. As we know, cancer is quite heterogeneous, so in genomic changes it is more important to have a panel of markers for efficient detection. For example, very few cancers have a single mutation at higher than 50% frequency. Other molecular biomarkers, such as circulating microRNA, are on the horizon and may improve the efficiency of circulating nucleic acid assays.

Comparison Between Circulating Tumor Cells and ctDNA

Both CTC and ctDNA technology and applications have greatly advanced in the last decade. Both assays provide some overlapping information as well as unique information for

specific cancers. CTC detection provides real-time information on tumor spread. If CTCs are sampled repetitively over days to weeks and are increasing in number, this may indicate that the tumor is progressing and active metastasis is occurring. Assessing the presence of CTCs in blood may be important before, during, and after therapy because their presence suggests potential metastasis and colonization in distant sites. Continued development of better molecular biomarkers for CTCs that predict survival and establishment of metastasis is important. Molecular characterization of CTCs in addition to quantification will provide more guidance for treatment decisions, particularly for targeted therapies. For example, the analysis of gene expression in a blood biopsy is only feasible in CTCs and not in ctDNA; this may be important for the detection of crucial therapy response biomarkers, such as androgen receptor splice variants and biomarkers indicating response to immunotherapy with checkpoint inhibitors.

ctDNA detection indicates there may be tumor present, but unlike CTCs, it does not indicate that metastasis is occurring. Consistent detection or increasing concentrations over time suggest that the tumor is active (progressing or rapidly renewing cell turnover). ctDNA assays require less blood, typically less than 2 mL, whereas most CTC assays require 7 to 10 mL. When the demand for other blood tests is high, the additional volume required by CTC assays can be a concern. Assays that require less blood will have better patient compliance, particularly for repetitive phlebotomy during treatment and follow-up. Table 49.4 compares the relative merits of CTC and ctDNA analyses.

The development of both CTC and ctDNA liquid biopsies provides many new opportunities for implementation over the next 5 years. These assays, upon validation, may provide better clinical management in treatment monitoring and understanding of tumor progression for many different types of cancers. As new biosensors, molecular procedures, and molecular devices develop in both CTC and ctDNA analysis, the sensitivity and specificity, as well as logistics, will improve. With the resource of the cancer gene atlas (TCGA NIH USA: http://cancergenome.nih.gov/), there is a significant amount of sequencing data of multiple cancer types that can be translated into new CTC and ctRNA/DNA targets in the future.

TABLE 49.4 Comparison Between Circulating Tumor Cell Free DNA and Circulating Tumor Cells

ctDNA	CTCs
Minimal amount of blood (<2 mL)	Larger blood draw (>7 mL)
Logistically easier to assay due to purity of ctDNA	Logistically difficult to assay due to small number of cells and purity
More informative	Less informative for the sample used
Easily quantifiable due to high-sensitivity molecular assays	Difficult to quantify due to heterogeneity and cell recovery
Easier to assess tumor heterogeneity	Tumor heterogeneity is difficult to assess
Identifies presence of tumor	Identifies presence of tumor
Cannot identify metastasis occurring	Can identify metastasis occurring
Highly diagnostic	Diagnosis not dependable clinically
Prognostic	Prognostic
Can identify tumor resistance during therapy	Poor in identifying tumor resistance during therapy

CTC, Circulating tumor cells; *ctDNA*, circulating tumor cell free DNA.

CIRCULATING NUCLEIC ACIDS FOR PRENATAL DIAGNOSTICS

Traditional methods of sampling fetal genetic material for a definitive diagnosis, such as chorionic villus sampling and amniocentesis, are invasive and expose the fetus to the risk of spontaneous miscarriage. The discovery of cell-free fetal DNA (cffDNA) in the circulation of pregnant women has led to noninvasive genetic and chromosomal assessment of the fetus through the sampling of maternal peripheral blood. The clinical introduction of cffDNA tests has substantially reduced the number of invasive procedures performed worldwide.

The risk for Down syndrome, with an incidence rate of 1 in 800 pregnancies, is the predominant reason for women seeking prenatal diagnosis. Strategies have been devised to identify high-risk pregnancies without invasive procedures by combining maternal age, serum biochemical markers, and ultrasonographic findings. However, false-positive rates are high. To improve the sensitivity and specificity of prenatal screening, methods for direct detection of the core pathology of the condition were needed; for example, trisomy 21 for Down syndrome. To this end, noninvasive methods have been developed to provide access to fetal DNA for analysis.

Cell-Free Fetal DNA in Maternal Plasma

Although it is hard to capture the rare, intact fetal cells in the maternal circulation, fetal DNA in the cell-free portion of maternal blood; namely, plasma and serum is more prevalent. Chromosome Y DNA was first detected in the plasma and serum of women who were pregnant with male fetuses, but not in women with female fetuses, in 1997. Many properties and uses of circulating cffDNA have now been investigated. cffDNA analysis is currently used for routine prenatal testing.

BIOLOGY OF CIRCULATING CELL-FREE FETAL NUCLEIC ACIDS IN MATERNAL PLASMA

Every milliliter of maternal plasma contains thousands of genome-equivalents of cell-free DNA, with the fetus contributing a minor proportion. In other words, the majority of the cell-free DNA in maternal plasma is derived from the mother. The median amount of cffDNA as a proportion of the total DNA in maternal plasma, termed the "fetal fraction," is around 10% in the first and second trimesters and around 20% during the third trimester. The absolute concentration of

cffDNA increases with gestational age, probably as a consequence of the increase in placental tissue mass.

Fetal DNA molecules are first detectable in maternal plasma around the 10th week of gestation. cffDNA is cleared from maternal plasma very rapidly after delivery with a half-life of about an hour, revealing that fetal DNA molecules in maternal plasma have a high turnover rate with efficient clearance. Renal clearance studies show that renal excretion is only a minor component of fetal DNA clearance. Despite the rapid clearance kinetics, cffDNA amounts to some 10% to 20% of the total DNA in maternal plasma, suggesting that substantial amounts of fetal DNA may be released continuously.

Cell-free DNA is a metabolic by-product of cell death and is therefore present in the circulation as short fragments in a "cell-free" form. The size profile of cell-free DNA molecules indicates that fragments are generally shorter than 200 bp (Fig 49.8). The most frequently represented size in plasma is 166 bp in length, and 166 bp corresponds to the length of DNA that is wound around a histone core with a linker. This characteristic size suggests that cell-free DNA is mainly derived from mononucleosomes. Interestingly, cffDNA molecules are somewhat shorter, and the most frequently represented size is 142 bp in length. This shorter length corresponds to the length of DNA wound around a histone core without a linker. This implies that the fetal DNA molecules may have undergone further steps of degradation compared to maternal DNA in maternal plasma. In addition to these dominant peaks, there are smaller amounts of cell-free DNA that are successively shorter in 10-bp increments. This observation suggests that DNase participates in the degradation of cell-free DNA because DNase cutting sites are located every 10 bases around the nucleosomal DNA.

Fetal Chromosomal Aneuploidy Screening

Down syndrome is one of the key reasons for couples to consider prenatal diagnosis. Therefore the key to achieving noninvasive prenatal detection of Down syndrome is to provide evidence that increased copies of chromosome 21 are present in maternal plasma. The majority of the DNA in maternal plasma, including DNA from chromosome 21, originates from the mother who has a normal amount of chromosome 21. If the fetus has trisomy 21, it would contribute additional amounts of chromosome 21 into maternal plasma. Therefore the additional amount of cell-free chromosome 21 DNA molecules in maternal plasma is dependent on the fetal fraction (i.e., the percent of fetal DNA in the background of maternal DNA in the plasma). The higher the fetal fraction, the easier the identification of trisomy 21–associated changes in maternal plasma.

Methodological Approaches

To precisely quantify and detect this small additional amount of chromosome 21 DNA, most protocols utilize massively parallel sequencing. One approach is based on random or shotgun sequencing of maternal plasma DNA. If a random fraction of all cell-free DNA fragments are sequenced, the relative amount of

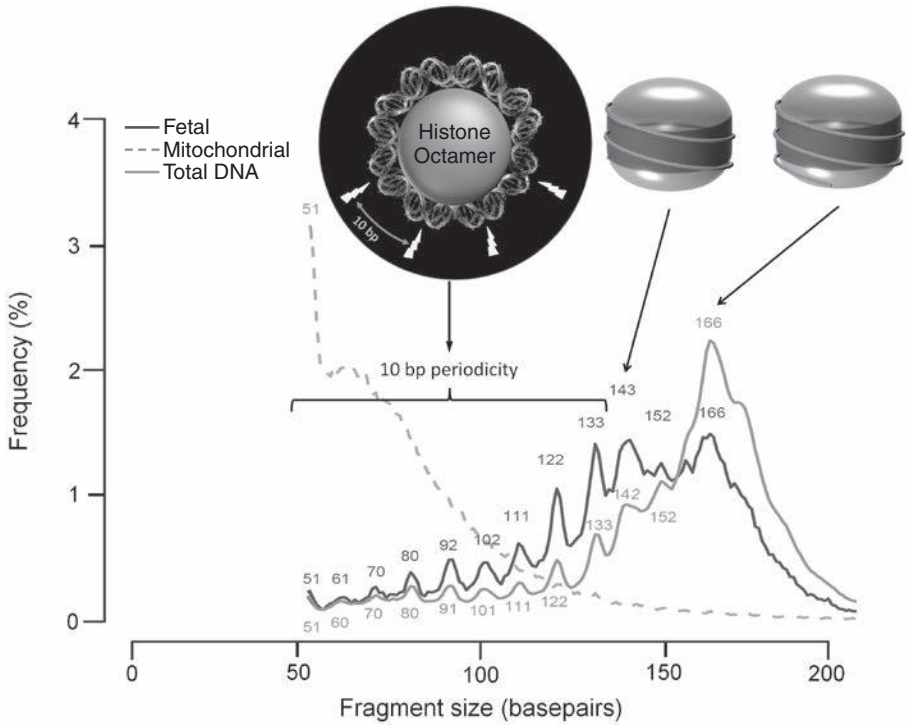


Fig. 49.8 Size distribution of fetal, mitochondrial, and total DNA. Numbers denote the DNA size in bps at the peaks. Schematic illustrations of the structural organization of a nucleosome are shown above the graph. From left to right, DNA double helix wound around a nucleosomal core unit with the sites for nuclease cleavage shown; a nucleosome core unit with approximately 146 bp of DNA wound around it; and a nucleosomal core unit with an approximately 20-bp linker intact.

DNA obtained from each segment across the genome should reflect the relative DNA contribution of that segment in the genome. To determine the relative amount of DNA from each segment—say, each chromosome—one determines the number of DNA molecules sequenced from that chromosome as a proportion of all molecules sequenced from the sample. The relative amount, or genomic representation, is then compared with the expected amount for the same chromosome among a control group of samples representing euploid pregnancies (Fig. 49.9). If a sample shows a genomic representation of chromosome 21 that is significantly different (e.g., more than three standard deviations) from the control group, the amount of chromosome 21 is considered elevated and therefore suggestive of trisomy 21.

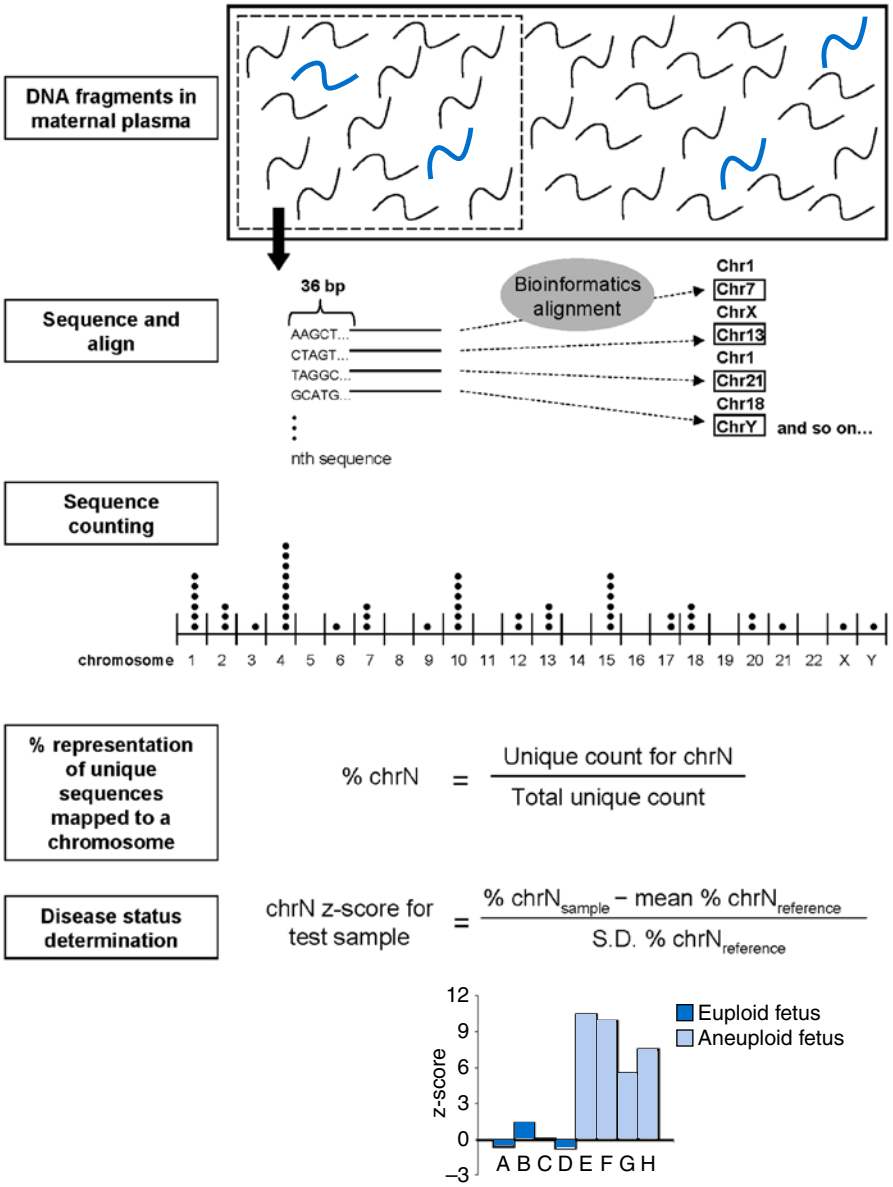


Fig. 49.9 Schematic illustration of the procedural framework for using massively parallel genomic sequencing for the noninvasive prenatal detection of fetal chromosomal aneuploidy. Fetal DNA *(thick blue fragments)* circulates in maternal plasma as a minor population among a high background of maternal DNA *(black fragments)*. A sample containing representative DNA molecules in maternal plasma is obtained. Plasma DNA molecules are sequenced, and the chromosomal origin of each molecule is identified through mapping to the human reference genome by bioinformatic analysis. The number of sequences mapped to each chromosome is counted and then expressed as a percentage of all unique sequences generated for the sample, termed *%chrN* for chromosome N. Z-scores for each chromosome and each test sample are calculated using the formula shown. The z-score of a potentially aneuploid chromosome is expected to be higher for pregnancies with an aneuploid fetus (cases E to H) than those with a euploid fetus (cases A to D). (From Chiu, R. W. K., Allen Chan, K. C., Gao, Y., Lau, V. Y. M., Zheng, W., Leung, T. Y., et al. [2008]. Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. *Proc Natl Acad Sci U S A, 105*, 20458–20463.)

Because this approach is based on random whole-genome sequencing, it could in principle be applied to other chromosomal aneuploidies, such as trisomy 18 and trisomy 13, and the sex chromosome aneuploidies, such as Turner syndrome, 45 X0, and Klinefelter syndrome, 47 XXY. The protocol could also be repurposed for the detection of microdeletion and microduplication syndromes.

Clinical Implementation

cffDNA-based noninvasive prenatal screening for fetal chromosomal aneuploidy became clinically available in 2011. Within 3 years, service availability extended to over 60 countries. In general, cffDNA-based noninvasive prenatal screening achieved detection rates of about 99% for Down syndrome with less than 1% false-positive rates. Depending on the specific protocol used, the detection rates for trisomy 18 and trisomy 13 are about 90% or greater with less than 1% false-positive rates. cffDNA-based prenatal screening has achieved better detection sensitivities and specificities than screening by maternal serum biochemistry and/or fetal ultrasonography. However, false-positive cases still occur, and therefore cffDNA assessment is still a screening procedure and not a definitive diagnosis. All clinical guidelines recommend that cffDNA tests with positive findings suggestive of chromosomal aneuploidy should be confirmed by tests on fetal genetic material collected by conventional invasive methods such as chorionic villus sampling or amniocentesis.

ANALYTICAL ASPECTS

Cell-free nucleic acids exist in plasma in the form of short fragments and are present in low abundance. The key to successful analysis of cffDNA lies in the preservation and maximization of the fetal fraction in the maternal blood sample. Thus attention to a number of preanalytical and analytical details is important to ensure the quality of the analysis.

Maternal Blood Sample Collection and Processing

One advantage of cffDNA analysis for prenatal screening is that, unlike maternal serum biochemistry testing, the use of the cffDNA tests is not restricted to a specific gestational period. This is because the cffDNA tests target the detection of the core pathologies of the fetus that exist throughout the pregnancy. However, during the earliest stage of pregnancy, there may not be a sufficient amount of fetal DNA in the maternal circulation. While cffDNA has been detected in maternal plasma in individual pregnant women from as early as 18 days of pregnancy, they become quantitatively sufficient for most testing purposes from about the 11th week of pregnancy. Between 9 to 13 weeks of gestation, 2% of pregnancies show fetal DNA fractions that are below 5% and are prone to potential false-negativity.

Maximizing the fetal DNA fraction in the maternal sample is a key factor to many cffDNA tests. With such a consideration, maternal plasma is preferred over maternal serum. Clotting of the maternal blood cells results in the release of more maternal DNA into serum as compared to plasma. While the absolute fetal DNA amounts in plasma and serum are similar, the fetal DNA fraction is significantly reduced in serum.

After the maternal peripheral blood samples are collected, plasma is harvested via centrifugation. The goal of centrifugation is to minimize the residual maternal blood cells in the sample. Reduction in maternal DNA contamination results in maximization of the fetal fraction. Therefore two-step centrifugation protocols are recommended. The plasma supernatant is carefully harvested after the first centrifugation with care not to disturb the underlying maternal blood cell layer. With the second centrifugation, the remaining cells in the form of a cell pellet are separated from the final supernatant. The resultant plasma can be stored frozen until further analysis.

Circulating Cell-Free Fetal DNA Analysis

A significant amount of cffDNA has a length shorter than 100 bp. For plasma DNA extraction, a protocol that is efficient in preserving the small DNA molecules is advantageous. Researchers have used size-selection methods to increase the proportional representation of short DNA within the sample to improve the detection of fetal DNA. In addition, the detection of short DNA from maternal plasma can be used to detect fetal disorders. The rationale is that when the fetus has an aneuploidy, such as trisomy 21, there should be an additional dose of short DNA from chromosome 21. Therefore the overall size profile of the chromosome 21 DNA molecules is shorter than that of other nonaneuploid chromosomes.

Measuring the Fetal DNA Fraction

Analysis of the fetal DNA fraction is an important quality control parameter for cffDNA analysis. The statistics developed to determine the presence or absence of a significant difference in chromosome dosage or allelic ratio is dependent on the sample containing at least a certain amount of fetal DNA. Therefore, for the tests based on massively parallel sequencing, the proportion of chromosome Y sequences is often used as a fetal fraction measurement for chromosome-Y-positive samples. A panel of placenta-specific DNA methylation markers has also been developed to quantify the fetal DNA. The proportion of short DNA in a sample is also a reasonable measure of the fetal DNA fraction.

Massively Parallel Maternal Plasma DNA Sequencing

The advantage of massively parallel sequencing-based cffDNA assessment is that many maternal plasma DNA molecules can be analyzed per sample, regardless of whether the DNA fragment is originating from the fetus or the mother. For quantitative sequencing applications, such as for fetal chromosomal aneuploidy detection, the key to success is to maximize the signal-to-noise ratio of the sequencing data covering the genetic abnormality of interest. Trying to maximize the fetal

fraction improves the chance of detecting the fetal abnormality. Increasing the sequencing depth either for random or targeted sequencing improves the precision, and thus reduces the noise, for genomic representation or allelic ratio assessments. The size of the genomic locus of interest also matters. At the same sequencing depth, larger loci, such as whole chromosome aneuploidies, are much easier to detect than subchromosomal aneuploidies. Thus protocols for the detection of subchromosomal aneuploidies require higher sequencing depths. The current massively parallel sequencing protocols also suffer from a certain extent of GC bias. To reduce the imprecision surrounding the quantitative measurement of cffDNA by sequencing, GC normalization steps are typically included in the bioinformatics analysis of the data. This is especially important for the detection of aneuploidies in GC-rich chromosomes, such as chromosomes 13 and 18.

On the other hand, current massively parallel sequencing protocols have a sequencing error rate that cannot be ignored. This has an impact on the detection of single-nucleotide variants of the fetus, such as point mutations, polymorphic alleles, and de novo mutations. Because cffDNA is the minor species of DNA in maternal plasma, high-depth sequencing is needed to detect fetal single-nucleotide variants. Yet, the higher the total number of bases sequenced, the higher the chance for sequencing errors. Thus extra care is needed in designing protocols for the detection of fetal single-nucleotide variants by sequencing. For example, targeted capture of loci of interest followed by sequencing is one feasible option. Targeted sequencing allows a high depth to be achieved without sequencing a high total number of bases, thereby reducing the number of sequencing errors detected and improving the signal-to-noise ratio.

ACKNOWLEDGMENT

The authors wish to acknowledge the original contributions of Aaron D. Bossler and Angela M. Caliendo, on which portions of the Infectious Disease section of this chapter are based.

REVIEW QUESTIONS

1. What is generally considered a significant change in a patient's HIV-1 viral load?
   a. 0.2 $\log_{10}$ copies/mL
   b. 0.3 $\log_{10}$ copies/mL
   c. 0.5 $\log_{10}$ copies/mL
   d. 1.0 $\log_{10}$ copies/mL
2. For nucleic acid testing of CT and NG, how long do NAATs remain positive after successful treatment of patients?
   a. 1 day
   b. 1 week
   c. 3 weeks
   d. 3 months
3. Which is a gene target used extensively in taxonomic metagenomics studies?
   a. 5S rRNA
   b. 16S rRNA
   c. 23S rRNA
   d. 30S rRNA
4. Which of the following is the most common mutation identified in the *CFTR* gene?
   a. p.Arg68Cys
   b. p.Arg117His
   c. p.Ile507del
   d. p.Phe508del
5. Which of the following *HTT* genotypes would be diagnostic for Huntington disease? CAG repeats of:
   a. 17 and 34
   b. 19 and 45
   c. 12 and 15
   d. 14 and 27
6. Whole exome sequencing is most useful in patients:
   a. with a constellation of findings associated with a known microdeletion syndrome
   b. with a phenotype suspected to be the result of a balanced translocation
   c. with a constellation of findings that are not associated with any known disorder
   d. with a phenotype suspected to be related to a mutation in a single gene
7. What is the most common ctDNA used for clinical tests in cancer patients?
   a. methylated DNA
   b. integrated DNA
   c. amplified genomic DNA
   d. mutation in DNA of gene exons
   e. viral DNA
8. In nasopharyngeal carcinoma patients, what viral ctDNA is of clinical utility?
   a. Herpes virus
   b. Papilloma virus
   c. EB virus
   d. HIV virus
9. In non-small cell lung cancer patients, what cancer mutation and detection method is most frequent and FDA approved for ctDNA monitoring of the respective targeted therapy?
   a. BRAF
   b. EGFR
   c. KRAS
   d. P53
10. What are the main benefits of CTCs analysis?
   a. It is easy, standardized, and high throughput
   b. It can give valuable prognostic and predictive information for the presence of active metastasis in cancer patients both at the DNA, and at the gene expression level
   c. CTCs are directly comparable with the primary tumor
   d. CTCs are all identical and thus easy to detect at very low levels

11. Cell free-fetal DNA
    a. is predominantly cleared via the maternal kidneys
    b. is predominantly shorter than the maternally derived cell-free DNA in maternal plasma
    c. shows a peak size of 246 bp
    d. is the predominant species of cell-free DNA in maternal plasma
    e. is predominantly derived from the fetal liver
12. Analysis of cell-free fetal DNA
    a. using maternal serum is preferred over maternal plasma
    b. is only applicable during the second trimester of pregnancy
    c. could only be achieved by the detection of chromosome Y
    d. is important to assess the fetal fraction
    e. cannot be based on massively parallel sequencing due to its scarce quantity
13. Noninvasive fetal chromosomal aneuploidy assessment by circulating cell-free fetal DNA analysis
    a. is only applicable to trisomy 13
    b. could be preceded by conventional maternal serum biochemistry screening
    c. shows equal performance for chromosome 21 and the sex chromosomes
    d. is not applicable to the deletion of fetal microdeletion
    e. is restricted to the third trimester of pregnancy
14. The most sensitive method for routine molecular monitoring of response to TKI therapy in chronic myelogenous leukemia patients is
    a. fluorescence in situ hybridization
    b. massively parallel sequencing
    c. RT-qPCR
    d. karyotyping
15. *TP53* loss of function mutations confer a poor prognosis in hematopoietic malignancies as *TP53* generally plays the role of
    a. receptor tyrosine kinase
    b. tumor suppressor
    c. RNA splicing molecule
    d. chromatin modifier
16. *BRAF* V600E mutation is an example of a mutation which affects signal transduction pathways and is commonly detected in
    a. Acute myeloid leukemia
    b. Chronic myelogenous leukemia
    c. Myelodysplastic syndrome
    d. Hairy cell leukemia

SUGGESTED READINGS

Alix-Panabières, C., & Pantel, K. (2016). Clinical applications of circulating tumor cells and circulating tumor DNA as liquid biopsy. *Cancer Discovery*, *6*, 479–491.

Bianchi, D. W., Parker, R. L., Wentworth, J., et al. (2014). DNA sequencing versus standard prenatal aneuploidy screening. *The New England Journal of Medicine*, *370*, 799–808.

Bidard, F. C., Fehm, T., Ignatiadis, M., et al. (2013). Clinical application of circulating tumor cells in breast cancer: Overview of the current interventional trials. *Cancer Metastasis Reviews*, *32*, 179–188.

Centers for Disease Control and Prevention. (2013). Testing for HCV infection: An update of guidance for clinicians and laboratorians. *MMWR. Recommendations and Reports*, *62*, 362–365.

Centers for Disease Control and Prevention. (2014). Recommendations for the laboratory-based detection of *Chlamydia trachomatis* and *Neisseria gonorrhoeae*. *MMWR. Recommendations and Reports* 63, (RR–2):1–19.

Chiu, R. W. K., Chan, K. C. A., Gao, Y., et al. (2008). Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 20458–20463.

Cho, I., & Blaser, M. J. (2012). The human microbiome: At the interface of health and disease. *Nature Reviews Genetics*, *3*, 260–270.

Jaiswal, S., Fontanillas, P., Flannick, J., et al. (2014). Age-related clonal hematopoiesis associated with adverse outcomes. *The New England Journal of Medicine*, *371*, 2488–2498.

Lo, Y. M. D., Chan, K. C. A., Sun, H., et al. (2010). Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Science Translational Medicine* 2, 61ra91.

Lo, Y. M. D., Corbetta, N., Chamberlain, P. F., et al. (1997). Presence of fetal DNA in maternal plasma and serum. *Lancet*, *350*, 485–487.

MacLeod, R., Tibben, A., Frontali, M., et al. (2013). Recommendations for the predictive genetic test in Huntington's disease. *Clinical Genetics*, *83*, 221–231.

Mardis, E. R., Ding, L., Dooling, D. J., et al. (2009). Recurring mutations found by sequencing an acute myeloid leukemia genome. *The New England Journal of Medicine*, *361*, 1058–1066.

Marzese, D. M., Hirose, H., & Hoon, D. S. (2013). Diagnostic and prognostic value of circulating tumor-related DNA in cancer patients. *Expert Review of Molecular Diagnostics*, *13*, 827–844.

Patrinos, G. P., Danielson, P. B., & Ansorge, W. J. (Eds.). (2017). *Molecular Diagnostics* (3rd ed.). London: Elsevier.

Persing, D.H., Tenover, F. C., Hayden, R. T., et al. (Eds.). (2016). *Molecular Microbiology: Diagnostic Principles and Practice* (3rd ed.). Washington DC: ASM Press.

Seto, J. T., Bengtsson, N. E., & Chamberlain, J. S. (2014). Therapy of genetic disorders-novel therapies for Duchenne muscular dystrophy. *Current Pediatrica Report*, *2*, 102–112.

Soverini, S., Hochhaus, A., Nicolini, F. E., et al. (2011). BCR-ABL kinase domain mutation analysis in chronic myeloid leukemia patients treated with tyrosine kinase inhibitors: Recommendations from an expert panel on behalf of European Leukemianet. *Blood*, *118*, 1208–1215.

Tomasetti, C., & Vogelstein, B. (2015). Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, *347*, 78–81.

Yang, Y., Muzny, D. M., Reid, J. G., et al. (2013). Clinical whole-exome sequencing for the diagnosis of Mendelian disorders. *The New England Journal of Medicine*, *369*, 1502–1511.

Reference Information for the Clinical Laboratory

Khosrow Adeli, Ferruccio Ceriotti, Victoria Higgins

OVERVIEW

Laboratory tests are ubiquitously used in clinical practice to screen, diagnose, and monitor patients for several medical conditions. To correctly interpret laboratory test results, physicians often use reference intervals (i.e., normal ranges) as a health-associated benchmark by which to compare patient test results. If a laboratory test result falls within the reference interval, the patient is considered to be within a healthy range for that particular test. On the other hand, if a laboratory test result falls outside the reference interval, the patient may require follow-up testing to confirm the presence of a particular medical condition. Therefore, population-based reference intervals are a fundamental tool in laboratory medicine to assess health status.

Although the concept and utility of reference intervals appear straightforward, their establishment can be challenging and complex, particularly for the pediatric population. Determining robust reference intervals requires a sufficiently large reference population (120 individuals per partition according to CLSI EP28-A3c) comprised of representative, healthy individuals. Recruiting a healthy pediatric population can pose additional challenges due to parental consent, small blood sample volume, and the large number of healthy subjects required. Due to the profound stages of growth and development throughout childhood and adolescence, these populations often additionally require several age and sex partitions.

Given the complexity of reference interval establishment, it is difficult for most laboratories to undertake this task, and instead, they often use reference intervals established from other laboratories, textbooks, or manufacturer product inserts. Unfortunately, several of these sources often do not use appropriate sampling methods (e.g., sample from patient populations) and/or statistical procedures to establish reference intervals, leading to inaccurate test result interpretation. In addition to the lack of evidence-based reference intervals in clinical use, many laboratories use reference intervals that were established using a different analytical instrument or population. Prior to adopting a reference interval, each clinical laboratory must ensure that the analytical method used to establish the reference interval is comparable to the method used for patient testing. Moreover, it is not yet clear to what extent ethnicity and/or local environmental factors influence reference intervals for specific analytes, and therefore reference intervals developed based on one reference population may not be applicable to a laboratory serving a different population.

It is important to note that despite the application of exclusion criteria to reference populations (i.e., populations used to establish reference intervals), they may still include individuals with subclinical conditions. Therefore, although reference intervals are a useful tool for clinicians, they should not be used as definitive limits of health and disease.

The reference intervals presented in the following tables are for general informational purposes only. Each individual laboratory should generate its own set of reference intervals or validate published reference intervals based on the CLSI EP28-A3c guidelines (CLSI document EP28-A3c. Wayne, P. A.: Clinical and Laboratory Standards Institute, 2010; p. 30-34).

Reference Interval and Critical Value Tables

- Table A1. Pediatric and Adult Reference Intervals for Biochemical Markers (Serum, Plasma, Urine, and Whole Blood)
- Table A2. Pediatric and Adult Critical Values (Risk Thresholds) for Biochemical Markers

CHAPTER OUTLINE

Tables A1 and A2 include up-to-date reference intervals and critical values obtained from recent studies of pediatric and adult reference values for several biochemical markers. Wherever possible, data from recent, large a priori reference interval studies based on healthy populations have been used to update each table with robust reference information. Where no new data were available, reference intervals from the *Tietz Textbook of Clinical Chemistry and Molecular Diagnostics*, 6th edition tables are presented.

Table A1 provides pediatric and adult reference intervals for several biochemical markers. The majority of pediatric reference intervals presented in this table are based on recent data published by the Canadian Laboratory Initiative on Pediatric Reference Intervals (CALIPER), a national research project that has established age- and sex-specific pediatric reference intervals based on thousands of healthy community children and adolescents. As noted in the table, the majority of CALIPER reference intervals were determined on Abbott Architect assays. For specific biochemical and endocrine markers, both pediatric and adult reference

values were obtained through the recent Canadian Health Measures Survey (CHMS), a study of approximately 12,000 children, adults, and the elderly. CHMS reference intervals were determined based on numerous platforms, including Ortho Clinical Diagnostics and Siemens, as noted in the table. Other reference intervals were adopted primarily from reference tables in *Tietz Textbook of Clinical Chemistry and Molecular Diagnostics,* sixth edition.

Table A2 provides a table of pediatric and adult critical risk values. Notifying clinical staff of critical results is an important post-analytical process in all acute care clinical laboratories. Data from numerous sources (including CLSI Guideline GP47. Wayne, P. A.: Clinical and Laboratory Standards Institute, 2015; p. 100) are included in this table of critical risk values to assist laboratories in establishing and updating their critical risk result management policy.

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers

			REFERENCE INTERVALS		
Analyte	**Specimen**	**Condition**	**Conventional Units**	**Conversion Factor**	**SI Units**
α1-Acid glycoprotein	S		mg/dL	0.01	g/L
Pediatric values (0–19 years) based on Abbott Architect method		0–<6 months	21–85		0.2–0.9
		6 months–<5 years	48–201		0.5–2.0
		5–<19 years	48–114		0.5–1.1
		Adult (20–60 years)	50–120		0.5–1.2
Adrenocorticotropic hormone	P, EDTA		pg/mL	0.22	pmol/L
ACTH reference intervals may vary depending on immunoassay methods		Cord	50–570		11–125
		Newborn	10–185		2.2–41
		Adult (0800–0900)	<120		<26
		Adult (24 h, supine)	<85		<19
Alanine	P		mg/dL	112.2	μmol/L
		Premature, 1 day	2.44–4.24		274–476
		Newborn, 1 day	2.10–3.65		236–410
		1–3 months	1.19–3.71		134–416
		2–6 months	1.58–3.68		177–413
		9 months–2 years	0.88–2.79		99–313
		3–10 years	1.22–2.71		137–305
		6–18 years	1.72–4.85		193–545
		Adult	1.87–5.88		210–661
	U, 24 h		mg/day	11.2	μmol/day
		10 days–7 weeks	4.1–9.3		46–104
		3–12 years	9.1–39.2		102–439
		Adult	7.9–48.3		88–541
			μmol/g creatinine	0.113	μmol/mol creatinine
		0–1 month	554–2,957		62.6–334.1
		1–6 months	613–2,874		59.3–324.8
		6 months–1 year	428–2,064		48.4–233.2
		1–2 years	389–1,497		44.0–169.2
		2–3 years	255–1,726		28.8–195.0
Alanine aminotransferase	S		U/L	0.017	μkat/L
With pyridoxal phosphate		0–<1 year	5–51		0.08–0.85
Values (0–<3 years) based on Abbott Architect method		1–<3 years	11–30		0.18–0.50
Values (3–79 years) based on Ortho Clinical Diagnostics (OCD) VITROS method		3–5 years	15–33		0.26–0.56
		6–8 years	16–37		0.27–0.63
		9–11 years	18–39		0.31–0.66
		12–17 years M	17–50		0.29–0.85
		12–49 years F	14–41		0.24–0.70
		18–49 years M	18–78		0.31–1.33

Continued

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
		50–79 years M	20–62		0.34–1.05
		50–79 years F	16–44		0.27–0.75
Without pyridoxal phosphate	S	0–<1 year	5–33		0.08–0.56
		1–<13 years	9–25		0.15–0.43
		13–19 years M	9–24		0.15–0.41
		13–19 years F	8–22		0.14–0.37
Albumin (BCG)	S		g/dL	10	g/L
Values (3–79 years) based on OCD VITROS method		0–<15 days	3.3–4.5		33–45
		15 days–<1 year	2.8–4.7		28–47
		1–<3 years	3.8–4.7		38–47
		3–5 years	3.9–5.0		39–50
		6–15 years	4.1–5.1		41–51
		16–29 years M	4.6–5.3		46–53
		16–54 years F	3.9–5.0		39–50
		30–54 years M	4.4–5.1		44–51
		55–79 years	4.2–5.0		42–50
		79–90 years	3.2–4.6		32–46
		>90 years	2.9–4.5		29–45
	U		mg/L	1	mg/L
		3–5 years	1.5–21.5		1.5–21.5
		6–8 years	1.5–37.8		1.5–37.8
		9–19 years	1.5–169.6		1.5–169.6
		20–29 years	1.5–74.6		1.5–74.6
		30–39 years	1.5–36.8		1.5–36.8
		40–79 years	1.5–47.2		1.5–47.2
Albumin-creatinine ratio (NKDEP guidelines)			mg/mmol		mg/g creatinine
		M	<2.5		<22
		F	<3.5		<30
Aldosterone	S		ng/dL	0.0277	nmol/L
		Cord blood	40–200		1.11–5.54
		Premature infants	19–141		0.53–3.91
Full-term infants					
		3 days	7–184		0.19–5.10
		1 week	5–175		0.03–4.85
		1–12 months	5–90		0.14–2.49
		1–2 years	7–54		0.19–1.50
		2–10 years (supine)	3–35		0.08–0.97
		2–10 years (upright)	5–80		0.14–2.22
		10–15 years (supine)	2–22		0.06–0.61
		10–15 years (upright)	4–48		0.11–1.33
Adults					
		(supine)	3–16		0.08–0.44
		(upright)	7–30		0.19–0.83
	U, 24 h		μg/day	nmol/day	μg/g creatinine
		Newborns (1–3 days)	0.5–5	2.771–14	20–140
		Prepubertal children			
		4–10 years	1–8	3–22	4–22
		Adults	3–19	8–51	1.5–20
Aluminum	S, P		μg/L	0.0371	μmol/L
			<5.51		<0.2
		Patients on hemodialysis	20–550		0.74–20.4
		All medication	<30		<1.11
	U		5–30		0.19–1.11
Amylase	S		U/L	0.017	μkat/L
Values (0≤19 years) based on Abbott Architect method		0–14 days	3–10		0.05–0.17

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
		15 days–<13 weeks	2–22		0.03–0.37
		13 weeks–<1 year	3–50		0.05–0.83
		1–<19 years	25–101		0.42–1.68
IFCC, 37°C		Adult	31–107		0.52–1.78
Androgen index, free (FAI)					%
		0–<1 year F			0.04–1.32
		1–<9 years F			0.04–1.32
		9–<14 years F			0.12–2.63
		14–19 years F			0.59–6.50
		0–<1 year M			0.02–32.72
		1–<9 years M			0.03–0.60
		9–<14 years M			0.15–34.68
		14–<19 years M			3.58–83.30
		*Please note: FAI should not be used in men where testosterone levels exceed sex hormone binding globulin (SHBG) binding capacity			
Androstenedione (LC-MS/MS)	S		ng/L		
		6–24 months M	25–150		
		2–3 years M	<110		
		4–5 years M	23–170		
		6–7 years M	10–290		
		7–9 years M	30–300		
		10–11 years M	70–390		
		12–13 years M	100–640		
		14–15 years M	180–940		
		16–17 years M	300–1,130		
		18–40 years M	330–1,340		
		40–67 years M	230–890		
		6–24 months F	<150		
		2–3 years F	<160		
		4–5 years F	20–210		
		6–7 years F	20–280		
		7–9 years F	40–420		
		10–11 years F	90–1,230		
		12–13 years F	240–1,730		
		14–15 years F	390–2,000		
		16–17 years F	350–2,120		
		Menstrual Status			
		Before menarche	48–1,080		
		After menarche, ≤18 years	330–2,130		
		Premenopausal, >18 years	260–2,140		
		Postmenopausal	130–820		
Vasopressin (antidiuretic hormone ADH)	P, EDTA	mOsm/kg	ng/L	0.926	pmol/L
		270–280	<1.5		<1.4
		280–285	<2.5		<2.3
		285–290	1–5		0.9–4.6
		290–294	2–7		1.9–6.5
		295–300	4–12		3.7–11.1
Antistreptolysin-O (ASO)	S		IU/mL	1	kIU/L
Values (0–<19 years) based on Abbott Architect method		0–<6 months	0–0		0–0
		6 months–<1 year	0–30		0–30
		1–<6 years	0–104		0–104
		6–<19 years	0–331		0–331
		Adult	<20–364		<20–364

Continued

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
α1-Antitrypsin	S		mg/dL	0.01	g/L
Values (0–<19 years) based on Abbott Architect method		0–<19 years	110–181		1.1–1.8
		Adult (20–60 years)	90–200		0.9–2.0
Apolipoprotein AI	S		mg/dL	0.01	g/L
Values (0–<6 years) based on Abbott Architect method		0–14 days F	71–97		0.7–1.0
		0–14 days M	62–91		0.6–0.9
		15 days–<1 year	53–175		0.5–1.8
		1–<6 years	80–164		0.8–1.6
Values (6–79 years) based on OCD VITROS method		6–15 years	100–180		1.0–1.8
		16–39 years M	80–170		0.8–1.7
		16–39 years F	100–200		1.0–2.0
		40–79 years M	120–200		1.2–2.0
		40–79 years F	110–230		1.1–2.3
Apolipoprotein B-100	S		mg/dL	0.01	g/L
Values (0–<6 years) based on Abbott Architect method		0–14 days	9–67		0.1–0.7
		15 days–<1 year	19–123		0.2–1.2
		1–<6 years	41–93		0.4–0.9
Values (6–79 years) based on OCD VITROS method		6–13 years	50–100		0.5–1.0
		14–29 years	40–110		0.4–1.1
		30–79 years	60–140		0.6–1.4
Arginine	P		mg/dL	57.4	μmol/L
		Premature, 1 day	0.17–1.57		10–90
		Newborn, 1 day	0.38–1.53		22–88
		1–3 months	0.38–1.30		22–74
		2–6 months	0.98–2.47		56–142
		9 months–2 years	0.19–1.13		11–65
		3–10 years	0.40–1.50		23–86
		6–18 years	0.77–2.26		44–130
		Adult	0.37–2.40		21–138
	U, 24 h		mg/day	5.74	μmol/day
		10 days–7 weeks	<1.2		<7
		3–12 years	<5.1		<29
		Adult	<50.2		<288
			mg/g creatinine	0.65	mmol/mol creatinine
		Adult	0–4		0–2.7
Arsenic	WB (Hep)		μg/L	0.0113	μmol/L
		Not exposed	2–23		0.03–0.31
		Chronic poisoning	100–500		1.33–6.65
		Acute poisoning	600–9,300		7.98–124
			μg/day	0.0133	μmol/day
	U, 24 h		5–50		0.07–0.67
Ascorbic acid (see vitamin C)					
Asparagine	P		mg/dL	75.7	μmol/L
		1–3 months	0.08–0.44		6–33
		3 months–6 years	0.95–1.90		72–144
		6–18 years	0.42–0.82		32–62
		Adult	0.40–0.91		30–69
	U, 24 h		mg/day	7.57	μmol/day
		Adult	4.5–13.2		34–100
			mg/g creatinine	0.86	mmol/mol creatinine
		Adult	2–10		1.8–8.6

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
Aspartate aminotransferase	S		U/L	0.017	µkat/L
with pyridoxal phosphate		0–14 days	23–186		0.38–3.10
Values (0–<3 years) based on Abbott Architect method		15 days–<1 year	23–83		0.38–1.38
		1–<3 years	26–55		0.43–0.92
Values (3–79 years) based on OCD VITROS method		3–5 years	28–52		0.48–0.88
		6–11 years M	25–47		0.43–0.80
		6–11 years F	23–44		0.39–0.75
		12–17 years M	18–36		0.31–0.61
		12–19 years F	15–34		0.26–0.58
		18–54 years M	18–54		0.31–0.92
		20–54 years F	18–34		0.31–0.58
		55–79 years	18–39		0.31–0.66
Without pyridoxal phosphate	S	0–14 days	32–162		0.54–2.75
Values (0–<19 years) based on Abbott Architect method		15 days–<1 year	20–67		0.34–1.14
		1–<7 years	21–44		0.36–0.75
		7–<12 years	18–36		0.31–0.61
		12–19 years M	14–35		0.24–0.60
		12–19 years F	13–26		0.22–0.44
Aspartic acid	P		mg/dL	75.1	µmol/L
		Premature, 1 day	0–0.39		0–30
		Newborn, 1 day	<0.21		<16
		1–3 months	0–0.15		0–8
		9 months–2 years	<0.12		<9
		19 months–10 years	<0.27		<20
		6–18 years	<0.19		<14
		Adult	<0.32		<24
	U, 24 h		mg/day	7.51	µmol/day
		3–12 years	<5.1		<38
		Adult	<26.2		<197
			mg/g creatinine	0.85	mmol/mol creatinine
		Adult	0–4		0.1–3.7
Bilirubin, direct (conjugated)	S		mg/dL	17.1	µmol/L
Values (0–<19 years) based on Abbott Architect method		0–14 days	0.33–0.71		5.7–12.1
		15 days–<1 year	0.05–0.30		0.8–5.2
		1–<9 years	0.05–0.20		0.8–3.4
		9–<13 years	0.05–0.29		0.8–5.0
		13–<19 years M	0.11–0.42		1.9–7.1
		13–<19 years F	0.10–0.39		1.7–6.7
		Adult	0.0–0.2		0.0–3.4
Bilirubin, total	S		mg/dL	17.1	µmol/L
Values (0–<3 years) based on Abbott Architect method		0–14 days	0.19–16.6		3.3–283.8
		15 days–<1 year	0.05–0.68		0.8–11.7
		1–<3 years	0.05–0.40		0.8–6.8
Values (3–79 years) based on OCD VITROS method					
		3–5 years	0.1–0.5		1.0–8.8
		6–15 years	0.1–0.9		1.0–15.6
		16–48 years M	0.2–1.1		3.0–18
		16–48 years F	0.1–0.9		1.0–16
		49–79 years M	0.1–1.2		2.0–19.9
		49–79 years F	0.1–1.0		1.0–16.6

Continued

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
	U		Negative		Negative
Biotin	WB				nmol/L
		Healthy			0.5–2.20
		Deficiency			<0.5
Cadmium	WB (Hep)		µg/L	8.897	nmol/L
		Nonsmokers	0.3–1.2		2.7–10.7
		Smokers	0.6–3.9		5.3–34.7
	U, 24 h		µg/L		µmol/L
		Toxic range	100–3,000		0.9–26.7
Calcitonin	S, P		pg/mL	1	ng/L
		M	<8.8		<8.8
		F	<5.8		<5.8
		Athyroidal	<0.5		<0.5
Calcium, ionized (free)	S, P (Hep)		mg/dL	0.25	mmol/L
		Adults	4.6–5.3		1.15–1.33
Calcium, total	S, P (Hep)		mg/dL	0.25	mmol/L
Values (0–<3 years) based on Abbott Architect method		0–<1 year	8.5–11.0		2.13–2.74
		1–<3 years	9.2–10.5		2.29–2.63
Values (3–79 years) based on OCD VITROS method		3–5 years	9.4–10.6		2.35–2.64
		6–15 years	9.3–10.5		2.33–2.62
		16–19	9.2–10.4		2.3–2.60
		20–39 years M	9.1–10.4		2.28–2.60
		20–39 years F	9.0–10.1		2.24–2.53
		40–79 years	9.0–10.2		2.24–2.56
β-Carotene HPLC	S		µg/dL	0.0186	µmol/L
			10–85		0.19–1.58
Cancer antigen 15–3	S		U/mL	1	kU/L
Values (0–<19 years) based on Abbott Architect method		0–<1 week	3.4–24		3.4–24
		1 week–<1 year	4.9–33		4.9–33
		1–<19 years	3.9–21		3.9–21
		Adult	<30		<30
Cancer antigen 19–9	S		U/mL	1	kU/L
Values (0–<19 years) based on Abbott Architect method		0–<1 year	<2.0–64		<2.0–64
		1–<19 years	<2.0–41		<2.0–41
		Adult	<37		<37
Cancer antigen 125	S		U/mL	1	kU/L
Values (0–<19 years) based on Abbott Architect method		0–<4 months	2.4–22		2.4–22
		4 months–<5 years	7.7–33		7.7–33
		5–<11 years	4.7–30		4.7–30
		11–<19 years	5.4–28		5.4–28
		Adult	<35		<35
Carbon dioxide, partial pressure PCO_2	WB, arterial (Hep)		mm Hg	0.133	kPa
		Newborn	27–40		3.59–5.32
		Infant	27–41		3.59–5.45
		Adult M	35–48		4.66–6.38
		Adult F	32–45		4.26–5.99
Carbon dioxide, total (tCO_2)			mEq/L	1	mmol/L
Values (0–<6 years) based on Abbott Architect method		0–14 days	5–20		5–20

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
		15 days–<1 year	10–24		10–24
		1–<5 years	14–24		14–24
		5–<6 years	17–26		17–26
Values (6–79 years) based on OCD VITROS method		6–79 years	19–26		19–26
	WB				
	Arterial		19–24		19–24
	Venous		22–26		22–26
Carcinoembryonic antigen (CEA)	S		ng/mL	1	µg/L
Values based on the Roche method		0–<1 week	8.1–62		8.1–62
		1 week–<2 years	<0.5–4.7		<0.5–4.7
		2–<19 years	<0.5–2.6		<0.5–2.6
		Adult, nonsmokers	<3		<3
		Adult, smokers	<5		<5
Catecholamines					
Epinephrine	P	Adults	pg/mL	5.46	pmol/L
		Supine (30 min)	<50		<273
		Sitting (15 min)	<60		<328
		Standing (30 min)	<90		<491
Norepinephrine	P	Adults	pg/mL	5.91	pmol/L
		Supine (30 min)	110–410		650–2,423
		Sitting (15 min)	120–680		709–4,019
		Standing (30 min)	125–700		739–4,137
Dopamine	P	Adults	pg/mL	6.53	pmol/L
		Supine (30 min)	<87		<475
		Sitting (15 min)	<87		<475
		Standing (30 min)	<87		<475
Ceruloplasmin	P		mg/L	0.001	g/L
		Cord (term)	50–330		0.05–0.33
Pediatric values (0–<19 years) based on Abbott Architect method		0–<2 months	74–237		0.07–0.24
		2–<6 months	135–329		0.13–0.33
		6 months–<1 year	137–389		0.14–0.39
		1–<8 years	217–433		0.22–0.43
		8–<14 years	205–402		0.21–0.40
		14–<19 years M	170–348		0.17–0.35
		14–<19 years F	208–432		0.21–0.43
		Adult M	220–400		0.22–0.40
		Adult F (no contraceptive)	250–600		0.25–0.60
		Adult F (contraceptives [estrogen])	270–660		0.27–0.66
		Adult F (pregnant)	300–1,200		0.3–1.20
			mg/dL	0.01	g/L
		Adult (20–60 years)	20–60		0.2–0.6
Chloride (Cl)	S, P		mEq/L	1	mmol/L
Values (3–79 years) based on OCD VITROS method		3–5 years	100–107		100–107
		6–11 years	101–107		101–107
		12–29 years M	101–106		101–106
		12–29 years F	100–107		100–107
		30–79 years	102–108		102–108
	U, 24 h		mEq/day	1	mmol/day

Continued

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
		Infant	2–10		2–10
		Child <6 years	15–40		15–40
		6–10 years M	36–110		36–110
		6–10 years F	18–74		18–74
		10–14 years M	64–176		64–176
		10–14 years F	36–173		36–173
		Adult	110–250		110–250
		>60 years	95–195		95–195
Cholesterol	S		mg/dL	0.0259	mmol/L
Reference limits					
Values (0–<3 years) based on Abbott Architect method		0–14 days M	42–109		1.09–2.82
		0–14 days F	46–125		1.19–3.24
		15 days–<1 year	64–237		1.66–6.14
		1–<3 years	112–208		2.90–5.39
Values (3–79 years) based on OCD VITROS method		3–5 years	120–216		3.11–5.59
		6–15 years	116–205		3.00–5.31
		16–19 years	100–182		2.59–4.71
		20–29 years	116–228		3.00–5.91
		30–39 years	147–266		3.81–6.89
		40–79 years	139–274		3.60–7.10
		*Please note: The more recent guidelines suggest that risk stratification should rely only upon the 10-year atherosclerotic cardiovascular disease risk calculation (2013 ACC/AHA Guideline on the Treatment of Blood Cholesterol to Reduce Atherosclerotic Cardiovascular risk in Adults).			
Clinical decision limits		Coronary heart disease risk, child			
		Desirable	<170		<4.40
		Borderline high	170–199		4.40–5.15
		High	>199		>5.15
		Coronary heart disease risk, adult			
		Desirable	<200		<5.18
		Borderline high	200–239		5.18–6.19
		High	>239		>6.19
		*Please note: Cholesterol should not be used on its own for risk prediction			
Cholinesterase (37°C)	S		U/L	0.017	μkat/L
		M	40–78		0.68–1.33
		F	33–76		0.56–1.29
Cholinesterase activity with dibucaine	S		U/L	0.017	μkat/L
inhibitor (ChEDi)		0–<1 month	797–2,478		13–41
Pediatric values (0–<19 years) based on Abbott Architect method		1 month–<19 years	1,523–3,280		25–55
Chorionic gonadotropin intact molecule	S		mIU/mL	1	IU/L
		Male and nonpregnant female	<5.0		<5.0
		Female, pregnancy (weeks of gestation)			
		4 weeks	5–100		5–100
		5 weeks	200–3,000		200–3,000
		6 weeks	10,000–80,000		10,000–80,000

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
		7–14 weeks	90,000–500,000		90,000–500,000
		15–26 weeks	5,000–80,000		5,000–80,000
		27–40 weeks	3,000–15,000		3,000–15,000
			*Values based on the Second International Standard for hCG		
		Trophoblastic disease	>100,000		>100,000
	U		Negative		Negative
			One half of pregnancies are detected on the first day of the missed menstrual period		One half of pregnancies are detected on the first day of the missed menstrual period
Chromium			µg/L	19.23	nmol/L
	WB (Hep)		0.7–28.0		14–538
	S		0.1–0.2		2–3
			µg/day	19.23	nmol/day
	U, 24 h		0.1–2.0		1.9–38.4
			µg/L	19.23	nmol/L
	RBC		20–36		384–692
Citric acid	U				mmol/mol creatinine
		0–1 month			<1,046
		1–6 months			104–268
		6 months–5 years			0–656
		>5 years			87–639
Cobalt			µg/L	16.97	nmol/L
	S		0.11–0.45		1.9–7.6
	U		1–2		17.0–34.0
			µg/kg		nmol/kg
	RBC		16–46		272–781
Complement C3	S		mg/dL	0.01	g/L
Values (0–<19 years) based on Abbott Architect method		0–14 days	50–121		0.5–1.2
		15 days–<1 year	51–160		0.5–1.6
		1–<19 years	83–152		0.8–1.5
		Adult (20–60 years)	90–180		0.9–1.8
Complement C4	S		mg/dL	0.01	g/L
Values (0–<19 years) based on Abbott Architect method		0–<1 year	7–30		0.1–0.3
		1–<19 years	13–37		0.1–0.4
		Adult (20–60 years)	10–40		0.1–0.4
Copper	S		µg/dL	0.157	µmol/L
		Birth–6 months	20–70		3.1–11.0
		Deficiency	<30		<5
		6 years	90–190		14.1–29.8
		12 years	80–160		12.6–25.1
		Adult M	70–140		11.0–22.0
		Adult F	80–155		12.6–24.3
		Deficiency	50		8
		Pregnancy, at term	118–302		18.5–47.4
		Blacks	Blacks 8–12% higher		Blacks 8–12% higher
	U, 24 h		µg/24 h	0.0157	µmol/24 h

Continued

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
		Adults	<60		<1.0
		Wilson's disease	>200		>3
Cortisol, free	S		μg/dL	27.6	nmol/L
		0800 h	0.6–1.6		17–44
		1600 h	0.2–0.9		6–25
	U, 24 h		μg/day	2.76	nmol/day
		1–10 years	2–27		6–74
		2–11 years	1–21		3–58
		11–20 years	5–55		14–152
		12–16 years	2–38		6–105
		Adult	μg/day	2.76	nmol/day
		Extracted	20–90		55–248
		Unextracted (HPLC)	75–270		207–745
Cortisol, total	S		μg/dL	27.6	nmol/L
		Cord blood	5–17		138–469
Pediatric values (2 days–<19 years) based on Abbott Architect method		2–<15 days	1–12		13–340
		15 days–<1 year	1–17		14–458
		1–<9 years	2–11		48–297
		9–<14 years	2–13		60–349
		14–<17 years	3–16		77–453
		17–<19 years	4–18		97–506
			μg/dL	27.6	nmol/L
		Child (1–16 years), 0800 h	3–21		83–580
		Adult			
		0800 h	5–23		138–635
		1600 h	3–16		83–441
		2000 h	<50% of 0800 h values		<50% of 0800 h values
C-reactive protein (CRP) high sensitivity	S		mg/L	1	mg/L
Values (0–<3 years) based on Abbott Architect method		0–14 days	0.3–6.1		0.3–6.1
		15 days–<3 years	0.1–1.0		0.1–1.0
Values (3–79 years) based on OCD VITROS method		3–5 years	0.1–2.4		0.1–2.4
		6–11 years	0.1–5.9		0.1–5.9
		12–13 years	0.1–1.9		0.1–1.9
		14–16 years	0.1–2.9		0.1–2.9
		17–39 years M	0.1–6.0		0.1–6.0
		17–39 years F	0.1–12.1		0.1–12.1
		40–79 years	0.1–8.8		0.1–8.8
		American M	0.3–8.6		0.3–8.6
		White American M	0.2–12.3		0.2–12.3
		African American M	0.1–8.2		0.1–8.2
		Mexican American M	0.2–6.3		0.2–6.3
		European M	0.3–8.6		0.3–8.6
		Japanese M	<7.8		<7.8
		American F	0.2–9.1		0.2–9.1
		European F	0.3–8.8		0.3–8.8
Creatine kinase (CK)	S		U/L	0.017	μkat/L
IFCC, 37°C		M	46–171		0.78–2.90
		F	34–145		0.58–2.47

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
CK isoenzymes	S	Fraction 2 (MB)	<5.0 μg/L	1	<5.0 μg/L
		Relative index MB/ total	<3.9%	0.01	<0.039 Fractional activity
Creatinine			mg/dL	88.4	μmol/L
Enzymatic	S	0–14 days	0.32–0.92		28–81
Values (0–<3 years) based on Abbott Architect method		15 days–<2 years	0.10–0.36		9–32
		2–<3 years	0.20–0.43		18–38
Values (3–79 years) based on OCD VITROS method		3–5 years	0.31–0.51		28–45
		6–7 years	0.36–0.56		32–49
		8–9 years	0.37–0.63		32–56
		10–11 years	0.43–0.68		38–60
		12–15 years M	0.47–0.91		42–81
		12–16 years F	0.48–0.84		42–74
		16–79 years M	0.71–1.16		63–102
		17–79 years F	0.56–0.96		49–85
	U		mg/dL	0.0884	mmol/L
Values (3–79 years) based on OCD VITROS method		3–5 years	14.71–151.58		1–13
		6–11 years	13.57–195.70		1–17
		12–13 years	21.49–214.93		2–19
		14–29 years	19.23–305.43		2–27
		30–79 years M	14.71–294.12		1–26
		30–79 years F	12.44–229.64		1–20
Jaffe			mg/dL	88.4	μmol/L
	S	Cord	0.60–1.20		53–106
		0–14 days	0.42–1.05		37–93
		15 days–<1 year	0.31–0.53		28–47
		1–<4 years	0.39–0.55		34–48
		4–<7 years	0.44–0.65		39–57
		7–<12 years	0.52–0.69		46–61
		12–<15 years	0.57–0.80		50–71
		15–<17 years M	0.65–1.04		58–92
		15–<17 years F	0.59–0.86		52–76
		17–<19 years M	0.69–1.10		61–97
		17–<19 years F	0.60–0.88		53–78
		18–60 years M	0.90–1.30		80–115
		18–60 years F	0.60–1.10		53–97
		60–90 years M	0.80–1.30		71–115
		60–90 years F	0.60–1.20		53–106
		>90 years M	1.00–1.70		88–150
		>90 years F	0.60–1.30		53–115
Jaffe, manual	U, 24 h		mg/kg per day	8.84	μmol/kg per day
		Infant	8–20		71–177
		Child	8–22		71–194
		Adolescent	8–30		71–265
		Adult M	14–26		124–230
		Adult F	11–20		97–177
Creatinine clearance (see Glomerular filtration rate)					
C-Telopeptide	S		ng/L	1	ng/L
		M	<1,009		<1,009
		Premenopausal female	<574		<574
	U		mg/mol creatinine	1	mg/mol creatinine

Continued

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
		M	0–505		0–505
		Premenopausal female	0–476		0–476
Cyanide	WB (Ox)		mg/L	38.5	µmol/L
		Nonsmokers	<0.2		<7.7
		Smokers	<0.4		<15.4
		Nitroprusside therapy	Up to 100 without toxicity		Up to 3,850
		Toxic	>1		>38.5
Cystatin C	S		mg/L	1	mg/L
Pediatric values (0–<19 years) based on Abbott Architect method		0–<1 month	1.49–2.85		1.49–2.85
		1–<5 months	1.01–1.92		1.01–1.92
		5 months–<1 year	0.75–1.53		0.75–1.53
		1–<2 years M	0.77–1.85		0.77–1.85
		1–<2 years F	0.60–1.20		0.60–1.20
		2–<19 years	0.62–1.11		0.62–1.11
		Adult F	0.61–1.05		0.61–1.05
		Adult M	0.71–1.21		0.71–1.21
Cystine	S		mg/dL	83.3	µmol/L
		Premature, 1 day	0.54–1.02		45–85
		Newborn, 1 day	0.43–1.01		36–84
		1–3 months	0.15–1.15		13–96
		2–6 months	0.64–0.97		53–81
		3–10 years	0.54–0.92		45–77
		6–18 years	0.43–0.70		36–58
		Adult	0.40–1.40		33–117
	U, 24 h		mg/day	8.33	µmol/day
		10 days–7 weeks	2.16–3.37		18–28
		3–12 years	4.9–30.9		41–257
		Adult	<38.1		<317
			mg/g creatinine	0.94	mmol/mol creatinine
		Adult	2–14		1.9–13.1
Dehydroepiandrosterone, unconjugated	S		ng/dL	0.0347	nmol/L
		Children			
		6–9 years M	13–187		0.45–6.49
		6–9 years F	18–189		0.62–6.55
		10–11 years M	31–205		1.07–7.11
		10–11 years F	112–224		3.88–7.77
		12–14 years M	83–258		2.88–8.95
		12–14 years F	98–360		3.40–12.5
		Adult M	180–1,250		6.25–43.4
		Adult F	130–980		4.51–34.0
Dehydroepiandrosterone sulfate	S		µg/dL	0.027	µmol/L
Pediatric values (0–<19 years) based on Abbott Architect method		0–<2 months	1,070–>1,507		28.9–>40.7
		2–<6 months	26–578		0.7–15.6
		6 months–<1 year	7–178		0.2–4.8
		1–<6 years	4–111		0.1–3.0
		6–<9 years	4–152		0.1–4.1
		9–<13 years	33–270		0.9–7.3
		13–<16 years	56–463		1.5–12.5
		16–<19 years M	126–674		3.4–18.2

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
		16–<19 years F	148–574		4.0–15.5
		Pubertal levels, Tanner Stage			
		1, M	5–265		0.1–7.2
		1, F	5–125		0.1–3.4
		2, M	15–380		0.4–10.3
		2, F	15–150		0.4–4.0
		3, M	60–505		1.6–13.6
		3, F	20–535		0.5–14.4
		4, M	65–560		1.8–15.1
		4, F	35–485		0.9–13.1
		5, M	165–500		4.4–13.5
		5, F	75–530		2.0–14.3
		Adults			
		18–30 years M	125–619		3.4–16.7
		18–30 years F	45–380		1.2–10.3
		31–50 years M	5–532		1.6–12.2
		31–50 years F	12–379		0.8–10.2
		51–60 years M	20–413		0.5–11.1
		61–83 years M	10–285		0.3–7.7
		Postmenopausal F	30–260		0.8–7.0
Dihydrotestosterone	S		ng/dL	0.0334	nmol/L
		Child, prepubertal	<3		<0.10
		Adult M	30–85		1.03–2.92
		Adult F	4–22		0.14–0.76
Dopamines	P, S		pg/mL		nmol/L
L-Dopa (1-dodecenoylcarnitine)		Normotensive adults	1,042–2,366	0.0051	5.3–12.0
DOPAC (3,4-dihydroxyphenylacetic acid)			674–2,636	0.0059	4.0–15.7
DHPG (3,4-dihydroxyphenylglycol)			797–1,208	0.0059	4.7–7.1
DU-PAN-2 (Duke pancreatic cancer associated antigen)			U/mL	1	kU/L
			<401		<401
Estradiol	S		pg/mL	3.69	pmol/L
Values (15 days–<19 years) based on Abbott Architect method		15 days–<1 year	<25		<92
		1–<9 years F	<10		<37
		9–<11 years F	<48		<176
		11–<12 years F	<94		<345
		12–<14 years F	11–172		39–631
		14–19 years F	<255		<936
		1–<11 years M	<13		<46
		11–<13 years M	<26		<95
		13–<15 years M	<28		<102
		15–<19 years M	<38		<141
		Adult M	10–50		37–184
		Adult F			
		Early follicular phase	20–150		73–550
		Late follicular phase	40–350		147–1,285
		Midcycle	150–750		550–2,753
		Luteal phase	30–450		110–1,652
		Postmenopausal	<21		<74
		Pubertal levels, Tanner Stage			

Continued

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
Tanner (values based on Abbott Architect method)		1, M	<19		<68
		1, F	<20		<74
		2, M	<18		<67
		2, F	<26		<96
		3, M	<21		<76
		3, F	<86		<317
		4, M	<35		<128
		4, F	13–141		49–517
		5, M	17–34		64–126
		5, F	19–208		69–762
Estriol, free (unconjugated, uE3)	S		ng/mL	3.47	nmol/L
		Males and nonpregnant females	<2.0		<6.9
		Pregnancy, week of gestation			
		16	0.30–1.05		1.04–3.64
		18	0.63–2.30		2.19–7.98
		34	5.3–18.3		18.4–63.5
		35	5.2–26.4		18.0–91.6
		36	8.2–28.1		28.4–97.5
		37	8.0–30.1		27.8–104.0
		38	8.6–38.0		29.8–131.9
		39	7.2–34.3		25.0–119.0
		40	9.6–28.9		33.3–100.3
Estriol, total (E3)	S		ng/mL	3.47	nmol/L
		Pregnancy, week of gestation			
		34	38–140		132–486
		35	31–140		108–486
		36	35–330		121–1,145
		37	45–260		156–902
		38	48–350		167–1,215
		39	59–570		205–1,978
		40	95–460		330–1,596
	U, 24 h		µg/day	3.47	nmol/day
		M	1.0–11.0		3.5–38.2
		F			
		Follicular phase	0–15.0		0–52.0
		Ovulatory phase	13.0–54.0		45.1–187.4
		Luteal phase	8.0–60.0		27.8–208.2
		Postmenopausal	0–11.0		0–38.2
		Pregnancy			
		First trimester	0–800		0–2,776
		Second trimester	800–12,000		2,776–41,640
		Third trimester	5,000–50,000		17,350–173,500
Estrone	S		pg/mL	3.69	pmol/L
		M	15–65		55–240
		F			
		Early follicular phase	15–150		55–555
		Late follicular phase	100–250		370–925
		Luteal phase	15–200		55–740
		Postmenopausal	15–55		55–204

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

			REFERENCE INTERVALS		
Analyte	**Specimen**	**Condition**	**Conventional Units**	**Conversion Factor**	**SI Units**
Ethanol	WB (Ox)		mg/dL	0.217	mmol/L
		Impairment	50–100		11–22
		Depression of CNS	>100		>21.7
		Fatalities reported	>400		>86.8
Ferritin	S		ng/mL	1	µg/L
Values (4 days–<3 years) based on Abbott Architect method		4–<15 days	99.6–717.0		99.6–717.0
		15 days–<6 months	14.0–647.2		14.0–647.2
		6 months–<1 year	8.4–181.9		8.4–181.9
		1–<3 years	5.3–99.9		5.3–99.9
Values (3–79 years) based on the Siemens IMMULITE method		3–5 years	10.7–85.2		10.7–85.2
		6–16 years M	16.2–106.7		16.2–106.7
		6–24 years F	9.6–81.9		9.6–81.9
		17–37 years M	39.3–439.4		39.3–439.4
		25–49 years F	6.5–147.1		6.5–147.1
		38–79 years M	45.8–714.8		45.8–714.8
		50–79 years F	6.0–362.6		6.0–362.6
α-fetoprotein	S		mg/dL	0.01	g/L
		Fetal, first trimester	200–400		2.0–4.0
		Cord blood	<5		<0.05
Pediatric values based on Abbott Architect method			ng/mL	1	µg/L
		0–<1 month	>2,000		>2,000
		1–<3 months	9.80–1,359.0		9.80–1,359.0
		3–<6 months	4.15–274.70		4.15–274.70
		6 months–<1 year	2.66–148.21		2.66–148.21
		1–<3 years	2.88–20.94		2.88–20.94
		3–<19 years	0.89–4.48		0.89–4.48
		Adult (85% of population)	<8.5		<8.5
		Adult (100% of population)	<15		<15
	Maternal serum		ng/mL (median)	1	µg/L (median)
		Weeks of gestation			
		14	25.6		25.6
		15	29.9		29.9
		16	34.8		34.8
		17	40.6		40.6
		18	47.3		47.3
		19	55.1		55.1
		20	64.3		64.3
		21	74.9		74.9
	S	Tumor marker	ng/mL	1	µg/L
		Early marker	10–20		10–20
		Cancer	>1,000		>1,000
Fluoride	S		mg/L	52.6	µmol/L
			0.2–3.2		10.5–168
Folate			ng/mL	2.265	nmol/L
Values (5 days–<6 years) based on Abbott Architect method	S	5 days–<1 year	>10.6		>23.9
		1–<3 years	>3.9		>8.7
		3–<6 years	>11.9		>27.0

Continued

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
Values (6–79 years) based on Siemens IMMULITE method		6–18 years	8.2–30.6		18.6–69.3
		19–79 years	9.5–39.0		21.5–88.4
	Erythrocyte	3–5 years	294.7–883.4		703.1–2,012.9
		6–79 years	228.2–998.7		541.4–2,110.6
		*Please note: Reference limits for erythrocytes depend on the level of supplementation in the country			
	S Deficiency		<1.4		<3.2
	Erythrocyte deficiency		<110		<252
Follicle stimulating hormone	S		mIU/mL	1	IU/L
Values (30 days–<19 years) based on Abbott Architect method		30 days–<1 year F	0.4–10.4		0.4–10.4
		1–<9 years F	0.4–5.5		0.4–5.5
		9–<11 years F	0.4–4.2		0.4–4.2
		11–19 years F	0.3–7.8		0.3–7.8
		30 days–<1 year M	0.1–2.4		0.1–2.4
		1–<5 years M	<0.9		<0.9
		5–<10 years M	<1.6		<1.6
		10–<13 years M	0.4–3.9		0.4–3.9
		13–<19 years M	0.8–5.1		0.8–5.1
		Pubertal levels, Tanner Stage			
		1, M	<1.5		<1.5
		1, F	0.6–4.1		0.6–4.1
		2, M	<3.0		<3.0
		2, F	0.3–5.8		0.3–5.8
		3, M	0.4–6.2		0.4–6.2
		3, F	0.1–7.2		0.1–7.2
		4, M	0.6–5.1		0.6–5.1
		4, F	0.3–7.0		0.3–7.0
		5, M	0.8–7.2		0.8–7.2
		5, F	0.4–8.6		0.4–8.6
		Adult (23–70 years) M	1.4–15.4		1.4–15.4
		Adult (23–70 years) F			
		Follicular phase	1.4–9.9		1.4–9.9
		Midcycle peak	0.2–17.2		0.2–17.2
		Luteal phase	1.1–9.2		1.1–9.2
		Postmenopausal	19.3–100.6		19.3–100.6
Fructosamine	S	Child	5% below adult levels		
		Adult	205–285 μmol/L		205–285 μmol/L
Glomerular filtration rate (endogenous)	Categories		mL/min/1.73 m^2	0.00963	mL/s/m^2
—based on KDIGO	G1	Normal or high	≥90		≥0.87
	G2	Mildly decreased	60–89		0.58–0.86
	G3a	Mildly to moderately decreased	45–59		0.43–0.57
	G3b	Moderately to severely decreased	30–44		0.29–0.42
	G4	Severely decreased	15–29		0.14–0.28
	G5	Kidney failure	<15		<0.14
Glucagon	P (Hep or EDTA)	Adult	ng/L		ng/L
			70–180		70–180

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
Glucose	S, fasting		mg/dL	0.0555	mmol/L
		Cord	45–96		2.5–5.3
		Premature	20–60		1.1–3.3
		Neonate	30–60		1.7–3.3
		Newborn			
		1 day	40–60		2.2–3.3
		>1 day	50–80		2.8–4.5
		Child	60–100		3.3–5.6
		Adult	74–100		4.1–5.6
		>60 years	82–115		4.6–6.4
		>90 years	75–121		4.2–6.7
Clinical decision limits	U		1–15		0.1–0.8
			g/day	5.55	mmol/day
	U, 24 h		<0.5		<2.8
Glutamic acid	P		mg/dL	68	µmol/L
		Premature, 1 day	0–1.98		0–135
		Newborn, 1 day	0.29–1.57		20–107
		6 months–3 years	0.28–1.47		19–100
		3–10 years	0.34–3.68		23–250
		6–18 years	0.10–0.96		7–65
		Adult	0.21–2.82		14–192
	U 24 h		mg/day	6.8	µmol/day
		10 days–7 weeks	0.3–1.5		2–10
		Adult	<33.8		<230
			mg/g creatinine	0.77	mmol/mol creatinine
		Adult	2–6		1.5–4.7
Glutamine	P		mg/dL	68.5	µmol/L
		3 months–6 years	6.93–10.89		475–746
		6–18 years	5.26–10.80		360–740
		Adult	5.78–10.38		396–711
	U, 24 h		mg/day	6.85	µmol/day
		10 days–7 weeks	12.4–25.8		85–177
		3–12 years	20.4–113.7		140–779
		Adult	43.8–151.8		300–1,040
			mg/g creatinine	0.77	mmol/mol creatinine
		Adult	2–78		2–60
γ-Glutamyltransferase	S		U/L	0.017	µkat/L
Values (0–<3 years) based on Abbott Architect method		0–<14 days	23–219		0.38–3.65
		15 days–<1 year	8–127		0.13–2.12
		1–<3 years	6–16		0.10–0.27
Values (3–79 years) based on OCD VITROS method		3–5 years	11–20		0.19–0.34
		6–14 years M	10–26		0.17–0.44
		6–17 years F	9–24		0.15–0.41
		15–19 years M	10–33		0.17–0.56
		18–35 years F	12–38		0.20–0.65
		20–35 years M	12–62		0.20–1.05
		36–79 years M	13–109		0.22–1.85
		36–79 years F	10–54		0.17–0.92
Glycated hemoglobin (HbA1c)	WB (EDTA, Hep or Ox)		%		mmol/mol (IFCC)
Values (6–79 years) based on OCD VITROS method		6–39 years	4.9–6.1		30–43
		40–79 years	5.0–6.3		31–45

Continued

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
		Cut off for diagnosis	≥6.5 (NGSP)		≥48
Glycine	P		mg/dL	133.3	µmol/L
		Premature 1 day	0–7.57		0–1,010
		Newborn 1 day	1.68–3.86		224–514
		1–3 months	0.79–1.67		106–222
		2–6 months	1.31–2.22		175–296
		9 months–2 years	0.42–2.31		56–308
		3–10 years	0.88–1.67		117–223
		6–18 years	1.18–2.27		158–302
		Adult	0.90–4.16		120–554
	U, 24 h		mg/day	13.3	µmol/day
		10 days–7 weeks	14.6–59.2		194–787
		3–12 years	12.4–106.8		165–1,420
		Adult	59.0–294.6		785–3,918
			mg/g creatinine	1.51	mmol/mol creatinine
		Adult	12–108		18.2–163
Growth hormone	S		ng/mL	1	µg/L
		Basal	2–5		2–5
		Insulin tolerance test	>10		>10
		Arginine	>7.5		>7.5
		L-Dopa	>7.5		>7.5
Haptoglobin	S		mg/dL	0.01	g/L
Values (0–<19 years) based on Abbott Architect method		0–14 days	0–10		0–0.1
		15 days–<1 year	7–221		0.1–2.2
		1–<12 years	7–163		0.1–1.6
		12–<19 years	7–179		0.1–1.8
		Adult (20–60 years)	30–200		0.3–2.0
High-density lipoprotein cholesterol (HDL-C)	S		mg/dL	0.0259	mmol/L
Reference Intervals					
Values (0–<3 years) based on Abbott Architect method		0–14 days	15–42		0.4–1.1
		15 days–<1 year	12–71		0.3–1.9
		1–<3 years	32–63		0.8–1.6
Values (3–79 years) based on OCD VITROS method		3–5 years	31–73		0.8–1.9
		6–14 years	35–81		0.9–2.1
		15–79 years M	31–70		0.8–1.8
		15–79 years F	35–89		0.9–2.3
Clinical decision limits		Pediatric	mg/dL	0.0259	mmol/L
		Acceptable	>45		>1.2
		Borderline	40–45		1.0–1.2
		Low	<40		<1.0
	ATP III Classification				
	S	Low	<40		<1.0
		High	>59		>1.5
Histidine	P		mg/dL	64.5	µmol/L
		Premature, 1 day	0.16–1.40		10–90
		Newborn, 1 day	0.76–1.77		49–114
		1–3 months	0.66–1.30		43–83
		2–6 months	1.49–2.12		96–137
		9 months–2 years	0.37–1.74		24–112
		3–10 years	0.37–1.32		24–85

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
		6–18 years	0.99–1.64		64–106
		Adult	0.50–1.66		32–107
	U, 24 h		mg/day	6.45	µmol/day
		10 days–7 weeks	16.0–38.6		103–249
		3–12 years	47.4–199.2		306–1,285
		Adult	72.9–440.8		470–2,843
			mg/g creatinine	0.73	mmol/mol creatinine
		Adult	1–141		1–103
Homocysteine, total	S, P		µg/mL	7.397	µmol/L
		Folate supplemented diet			
		<15 years	<1.08		<8
		15–65 years	<1.62		<12
		>65 years	<2.16		<16
		No folate supplementation			
		<15 years	<1.35		<10
		15–65 years	<2.03		<15
		>65 years	<2.70		<20
Values (5 days–<6 years) based on Abbott Architect method		5 days–<1 year	0.39–1.35		2.9–10.0
		1–<6 years	0.37–1.03		2.8–7.6
Values (6–79 years) based on OCD VITROS method		6–12 years	0.23–0.93		1.7–6.9
		13–25 years M	0.49–1.43		3.6–10.6
		13–39 years F	0.39–1.28		2.9–9.5
		26–79 years M	0.70–1.91		5.2–14.1
		40–79 years F	0.50–1.47		3.7–10.9
Homovanillic acid	U, 24 h		mg/day	5.49	µmol/day
		3–6 years	1.4–4.3		8–24
		6–10 years	2.1–4.7		12–26
		10–16 years	2.4–8.7		13–48
		16–83 years	1.4–8.8		8–48
	U		mg/g creatinine	0.571	mmol/mol creatinine
		0–6 months	<40		<23
		6 months–5 years	<10		<6
		3–6 years	5.4–15.5		3.4–9.6
		6–10 years	4.4–11.5		2.7–7.1
		10–16 years	3.3–10.3		2.0–6.4
17–Hydroxyprogesterone			ng/dL	0.03	nmol/L
		Cord blood	900–5,000		27.3–151.5
		Premature	26–568		0.8–17.0
Values (4 days–<19 years) based on Abbott Architect method		4 days–<1 year F	<132		<4.0
		30 days–<1 year M	<66		<2.0
		1–<10 years	<35		<1.1
		10–<15 years	13–85		0.4–2.7
		15–<19 years F	20–1,026		0.6–32.6
		15–<19 years M	16–57		0.5–1.8
Tanner values based on Abbott Architect method		Puberty–Tanner Stage			
		1, M	<44		<1.4
		1, F	<28		<0.9
		2, M	<44		<1.4
		2, F	13–41		0.4–1.3

Continued

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
		3, M	<50		<1.6
		3, F	16–47		0.5–1.5
		4, M	<41		<1.3
		4, F	19–72		0.6–2.3
		5, M	13–50		0.4–1.6
		5, F	<1,082		<34.4
		Adult M	27–199		0.8–6.0
		Adult F			
		Follicular phase	15–70		0.4–2.1
		Luteal phase	35–290		1.0–8.7
		Pregnancy	200–1,200		6.0–36.0
For LC-MSMS Reference Intervals please see: *Clin Chem* 2006; *52*(8), 1559–1567		Post ACTH	<320		<9.6
		Postmenopausal	<70		<2.1
Immunoglobulin A	S, P		mg/dL	0.01	g/L
Values (0–<19 years) based on Abbott Architect method		0–<1 year	1–29		0.0–0.3
		1–<3 years	4–90		0.0–0.9
		3–<6 years	26–147		0.3–1.5
		6–<14 years	47–221		0.5–2.2
		14–<19 years	53–287		0.5–2.9
		Adult (20–60 years)	70–400		0.7–4.0
		Adult (>60 years)	90–410		0.9–4.1
Immunoglobulin D	S		IU/mL	1	kIU/L
		Adult (20–60 years)	0–160		0–160
			ng/mL	1	µg/L
			0–384		0–384
Immunoglobulin E	S		kIU/L	2.4	µg/L
Values (0–<19 years) based on Abbott Architect method		0–<7 years	<25–440		<60–1,057
		7–<19 years	<25–450		<60–1,079
		Adult (20–60 years)	0–160		0–380
Immunoglobulin G	S		mg/dL	0.01	g/L
Values (0–<19 years) based on Abbott Architect method		0–14 days	320–1,407		3.2–14.1
		15 days–<1 year	108–702		1.1–7.0
		1–<4 years	316–1,148		3.2–11.5
		4–<10 years	542–1,358		5.4–13.6
		10–<19 years	658–1,534		6.6–15.3
		Adult (20–60 years)	700–1,600		7.0–16.0
		Adult (>60 years)	600–1,560		6.0–15.6
Immunoglobulin M	S		mg/dL	0.01	g/L
Values (0–<19 years) based on Abbott Architect method		0–14 days	5–35		0.1–0.4
		15 days–13 weeks	12–71		0.1–0.7
		13 weeks–<1 year	16–86		0.2–0.9
		1–<19 years M	39–151		0.4–1.5
		1–<19 years F	48–186		0.5–1.9
		Adult (20–60 years)	40–230		0.4–2.3
		Adult (>60 years)	30–360		0.3–3.6
Inhibin A	S		pg/mL	1	ng/L
		M	1.0–3.6		1.0–3.6
		F (Cycling; days of cycle)			

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
		Early follicular phase (−14 to −10 days)	5.5–28.2		5.5–28.2
		Midfollicular phase (−9 to −4 days)	7.9–34.5		7.9–34.5
		Late follicular phase (−3 to −1 day)	19.5–102 .3		19.5–102 .3
		Midcycle (day 0)	49.9–155.5		49.9–155.5
		Early luteal (1–3 days)	35.9–132.7		35.9–132.7
		Midluteal (4–11 days)	13.2–159.6		13.2–159.6
		Late luteal (12–14 days)	7.3–89.9		7.3–89.9
		IVF, peak levels	354–1,690		354–1,690
		PCOS, ovulatory	5.7–16.0		5.7–16.0
		Postmenopausal	1.0–3.9		1.0–3.9
Insulin	S		μIU/mL	7	pmol/L
Values (0–<6 years) based on Abbott Architect method		0–<1 year	1.0–23.4		7-164
		1–<6 years	1.3–40.2		9-281
Values (6–79 years) based on Siemens IMMULITE ADVIA Centaur methods		6–10 years	0.4–13.0		3-91
		11–19 years	2.1–19.5		15-137
		20–79 years	2.4–21.8		17-153
Insulin-like growth factor-1	S		ng/mL	1	μg/L
		1–2 years M	31–160		31–160
		1–2 years F	11–206		11–206
		3–6 years M	16–288		16–288
		3–6 years F	70–316		70–316
		7–10 years M	136–385		136–385
		7–10 years F	123–396		123–396
		11–12 years M	136–440		136–440
		11–12 years F	191–462		191–462
		13–14 years M	165–616		165–616
		13–14 years F	286–660		286–660
		15–18 years M	134–836		134–836
		15–18 years F	152–660		152–660
		19–25 years M	202–433		202–433
		19–25 years F	231–550		231–550
		Adult (25–85 years) M	135–449		135–449
		Adult (25–85 years) F	135–449		135–449
Insulin-like growth factor-II	S		ng/mL	1	μg/L
		Child			
		Prepubertal	334–642		334–642
		Pubertal	245–737		245–737
		Adult	288–736		288–736
		GH deficiency	51–299		51–299
Iodine	U		μg/dL	0.079	μmol/L
Values based on manual microplate analysis		3–5 years	5–83		0.39–6.58
		6–79 years	1–49		0.09–3.88
Iron			μg/dL	0.179	μmol/L
Values based on Abbott Architect method		0–<14 years	16–128		2.8–22.9
		14–<19 years M	31–168		5.5–30.0
		14–<19 years F	20–162		3.5–29.0

Continued

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

			REFERENCE INTERVALS		
Analyte	**Specimen**	**Condition**	**Conventional Units**	**Conversion Factor**	**SI Units**
Isoleucine	P		mg/dL	76.3	µmol/L
		Premature, 1 day	0.26–0.78		20–60
		Newborn, 1 day	0.35–0.69		27–53
		1–3 months	0.59–0.95		45–73
		2–6 months	0.50–1.61		38–123
		9 months–2 years	0.34–1.23		26–94
		3–10 years	0.37–1.10		28–84
		6–18 years	0.50–1.24		38–95
		Adult	0.48–1.28		37–98
	U		mg/day	7.62	µmol/day
		10 days–7 weeks	Trace-0.4		Trace-3
		3–12 years	2–7		15–53
		Adult	5–24		38–183
			mg/g creatinine	0.86	mmol/mol creatinine
		Adult	1–5		0.8–4.4
L-Lactate	WB (Hep)		mg/dL	0.111	mmol/L
		At bed rest	5–12		0.56–1.39
		Venous	3–7		0.36–0.75
		Arterial	16–17		1.78–1.88
	U, 24 h	Adult			
					mmol/mol creatinine
		0–1 month			46–348
		1–6 months			57–346
		6 months–5 years			21–38
		>5 years			20–101
Lactate dehydrogenase (LD)	S		U/L	0.017	µkat/L
		0–14 days	309–1,222		5.3–20.8
		15 days–<1 year	163–452		2.8–7.7
		1–<6 years	192–321		3.3–5.5
		6–10 years	405–646		6.9–11.0
		11–15 years M	319–612		5.4–10.4
		11–15 years F	284–575		4.8–9.8
		16–20 years	259–472		4.4–8.0
		21–39 years M	267–477		4.5–8.1
		21–39 years F	258–467		4.4–7.9
		40–79 years	271–497		4.6–8.4
Lead	WB (Hep)		µg/dL	0.0483	µmol/L
		Child	<25		<1.21
		Adult	<25		<1.21
		Toxic	>99		>4.78
	U, 24 h		µg/L		µmol/L
			<80		<0.39
Leucine	P		mg/dL	76.3	µmol/L
		Premature, 1 day	0.26–1.58		20–120
		Newborn, 1 day	0.62–1.43		47–109
		1–3 months	10.58–2.14		44–164
		9 months–2 years	0.59–2.03		45–155
		3–10 years	0.73–2.33		56–178
		6–18 years	1.03–2.28		79–174
		Adult	0.98–2.29		75–175
	U, 24 h		mg/day	7.624	µmol/day
		10 days–7 weeks	0.9–2.0		7–15
		3–12 years	3–11		23–84
		Adult	2.6–8.1		20–62
			mg/g creatinine	0.86	mmol/mol creatinine

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
		Adult	0–8		0–6.8
Lipase,	S		U/L	0.017	µkat/L
Values (0–<19 years) based on Abbott Architect method		0–<19 years	4–39		0.07–0.65
37°C		Adult	<38		<0.65
Low-density lipoprotein cholesterol (LDL-C) (Measured)	S		mg/dL	0.0259	mmol/L
Reference Intervals		6–24 years	46–143		1.2–3.7
Values (6–79 years) based on OCD VITROS method		25–49 years M	62–189		1.6–4.9
		25–49 years F	50–178		1.3–4.6
		50–79 years	73–189		1.9–4.9
		*Please note: The more recent guidelines suggest that risk stratification should rely only upon the 10-year atherosclerotic cardiovascular disease risk calculation (2013 ACC/AHA Guideline on the Treatment of Blood Cholesterol to Reduce Atherosclerotic Cardiovascular risk in Adults).			
Clinical decision limits			mg/dL	0.0259	mmol/L
		Risk of coronary heart disease, Child			
		Acceptable	<110		<2.8
		Borderline	110–129		<3.3
		High	≥130		≥3.4
		Risk coronary heart disease, Adults			
		Optimal	<100		<2.59
		Near/above optimal	100–129		2.59–3.34
		Borderline high	130–159		3.37–4.12
		High	160–189		4.15–4.90
		Very high	>189		>4.90
Luteinizing hormone (LH)			mIU/mL	1	IU/L
Values (4 days–<19 years) based on Abbott Architect method		4 days–<3 months F	<2.4		<2.4
		4 days–<3 months M	0.2–3.8		0.2–3.8
		3 months–<1 year F	<1.2		<1.2
		3 months–<1 year M	<2.9		<2.9
		1–<10 years	<0.3		<0.3
		10–<13 years	<4.3		<4.3
		13–<15 years F	0.4–6.5		0.4–6.5
		13–<15 years M	<4.1		<4.1
		15–<17 years F	<13.1		<13.1
		15–<17 years M	0.8–4.8		0.8–4.8
		17–<19 years F	<8.4		<8.4
		17–<19 years M	0.9–7.1		0.9–7.1
		Pubertal levels, Tanner Stage			
Tanner values based on Abbott Architect method		1, M	<1.2		<1.2
		1, F	<0.1		<0.1
		2, M	<1.2		<1.2
		2, F	<2.3		<2.3
		3, M	<2.3		<2.3
		3, F	<7.4		<7.4
		4, M	<4.9		<4.9
		4, F	0.3–6.7		0.3–6.7
		5, M	0.6–5.9		0.6–5.9

Continued

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
		5, F	0.4–21.2		0.4–21.2
		Adult (23–70 years) M	1.2–7.8		1.2–7.8
		Adult (23–70 years) F			
		Follicular phase	1.7–15.0		1.7–15.0
		Midcycle peak	21.9–56.6		21.9–56.6
		Luteal phase	0.6–16.3		0.6–16.3
		Postmenopausal	14.2–52.3		14.2–52.3
Lysine	P		mg/dL	68.5	µmol/L
		Premature,1 day	1.01–4.53		70–310
		Newborn, 1 day	1.66–3.93		114–269
		1–3 months	0.54–2.46		37–169
		9 months–2 years	0.66–2.10		45–144
		3–10 years	1.04–2.20		71–151
		6–18 years	1.58–3.40		108–233
		Adult	1.21–3.47		83–238
	U, 24 h		mg/day	6.85	µmol/day
		10 days–7 weeks	5.7–10.9		39–75
		3–12 years	9.3–93.7		64–642
		Adult	3.1–153.0		21–1,048
			mg/g creatinine	0.77	mmol/mol creatinine
		Adult	4–12		3.2–9.2
α2-Macroglobulin	S		mg/dL	0.01	g/L
		Adult (20–60 years)	130–300		1.3–3.0
Magnesium, free	S		mmol/L	1.0	mmol/L
			0.45–0.60		0.45–0.60
Magnesium, total (enzymatic)	S		mg/dL	0.4114	mmol/L
		0–14 days	1.99–3.94		0.82–1.62
		15 days–<1 year	1.97–3.09		0.81–1.27
		1–<19 years	2.09–2.84		0.86–1.17
Manganese			µg/L	18.0	nmol/L
	WB (Hep)		5–15		90–270
	S		0.5–1.3		9–24
	U, collect in metal free container		0.5–9.8		9.1–178
		Toxic conc.	>19		>342
Mercury			µg/L	4.99	nmol/L
	WB (EDTA)		0.6–59		3.0–294.4
	U, 24 h		<20		<99.8
		Toxic conc.	>150		>748.5
		Lethal conc.	>800		>3,992
Methemoglobin (MetHb)	WB (EDTA, Hep or ACD)		g/dL	155	µmol/L
			0.06–0.24		9.3–37.2
			% of total Hb	0.01	Mass fraction of total Hb
			0.04–1.52		0.0004–0.0152
Methionine			mg/dL	67.7	µmol/L
	P	Premature, 1 day	0.38–0.66		25–45
		Newborn, 1 day	0.13–0.61		9–41
		1–3 months	0.05–0.57		3–39
		2–6 months	0.24–0.73		16–49
		9 months–2 years	0.04–0.43		3–29
		3–10 years	0.16–0.24		11–16
		6–18 years	0.24–0.55		16–37

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
		Adult	0.09–0.60		6–40
			mg/day	6.7	µmol/day
		10 days–7 weeks	0.1–1.9		0.7–13
		3–12 years	3–14		20–95
		Adult	<9.1		<63
			mg/g creatinine	0.76	mmol/mol creatinine
		Adult	0–9.5		0–7.2
β2-Microglobulin	S		mg/dL	10	mg/L
Values (0–<19 years) based on Abbott Architect method		0–<3 months M	0.19–0.47		1.9–4.7
		0–<3 months F	0.19–0.58		1.9–5.8
		3 months–<2 years	0.13–0.45		1.3–4.5
		2–<19 years	0.12–0.23		1.2–2.3
			mg/dL (mean)	10	mg/L (mean)
		0–59 years	0.19		1.9
		60–69 years	0.21		2.1
		>70 years	0.24		2.4
Molybdenum			µg/L	10.42	nmol/L
	S		0.1–3.0		1.0–31.3
			µg/day	10.42	nmol/day
	U, 24 h		40–60		416–625
Niacin	U, 24 h		mg/day	7.3	µmol/day
			2.4–6.4		17.5–46.7
Nickel			µg/L	17	nmol/L
	S or P (Hep)		0.14–1.0		2.4–17.0
	WB		1.0–28.0		17–476
			µg/day	17	nmol/day
	U, 24 h		0.1–10		2–170
N-telopeptide (BCE = bone collagen equivalents)	S		nmol BCE/L	1.0	nmol BCE/L
		M	5.4–24.2		5.4–24.2
		Premenopausal female	6.2–19.0		6.2–19.0
			nmol BCE/mmol creatinine	1.0	nmol BCE/mmol creatinine
	U	M	3–63		3–63
		Premenopausal female	5–65		5–65
Orotic acid					mmol/mol creatinine
	U	0–1 month			1.4–5.3
		1–6 months			1.0–3.2
		6 months–5 years			0.5–3.3
		>5 years			0.4–1.2
Osteocalcin	S		ng/mL	1.0	µg/L
		Adult M	3.0–13.0		3.0–13.0
		Adult F			
		Premenopausal	0.4–8.2		0.4–8.2
		Postmenopausal	1.5–11.0		1.5–11.0
Oxalic acid					mmol/mol creatinine
	U	0–1 month			51–931
		1–6 months			7–567
		6 months–5 years			7–352
		>5 years			<188
Oxygen, partial pressure (PO_2)	Cord blood		mm Hg	0.133	kPa
	Arterial		5.7–30.5		0.8–4.0
	Venous		17.4–41.0		2.3–5.5

Continued

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
	WB, arterial	Birth	8–24		1.06–3.19
		5–10 minutes	33–75		4.39–9.96
		30 minutes	31–85		4.12–11.31
		1 hour	55–80		7.32–10.64
		1 day	54–95		7.18–12.64
		2 days–60 years	83–108		11.04–14.36
		>60 years	>80		> 10.64
		>70 years	>70		>9.31
		>80 years	>60		>7.98
		>90 years	>50		>6.65
Oxygen, saturation (sO_2)	WB, arterial		Percent saturation	0.01	Fraction saturation
		Newborn	40–90		0.40–0.90
		Thereafter	94–98		0.94–0.98
Oxytocin	P, EDTA		μU/mL	1.0	mU/L
		M	1.1–1.9		1.1–1.9
		F			
		Nonpregnant	1.0–1.8		1.0–1.8
		Second stage of labor	3.2–5.3		3.2–5.3
Parathyroid hormone, intact	S		pg/mL	0.106	pmol/L
Values (6 days–<3 years) based on Abbott Architect method		6 days–<1 year	6–89		0.7–9.4
		1–<3 years	16–63		1.7–6.7
Values (3–79 years) based on DiaSorin LIAISON method		3–5 years	7–29		0.7–3.1
		6–11 years	7–30		0.7–3.2
		12–15 years	8–36		0.8–3.8
		16–29 years	8–32		0.8–3.4
		30–79 years	9–42		1.0–4.4
Parathyroid hormone, (1–84)	S		pg/mL	1.0	ng/L
			6–40		6–40
pH (37°C)	WB, arterial		pH	1.0	pH
		Cord blood			
		Arterial	7.18–7.38		7.18–7.38
		Venous	7.25–7.45		7.25–7.45
		Newborn			
		Premature, 48 hours	7.35–7.50		7.35–7.50
		Full term			
		Birth	7.11–7.36		7.11–7.36
		5–10 minutes	7.09–7.30		7.09–7.30
		30 minutes	7.21–7.38		7.21–7.38
		1 hour	7.26–7.49		7.26–7.49
		1 day	7.29–7.45		7.29–7.45
		Children, adults			
		Arterial	7.35–7.45		7.35–7.45
		Venous	7.32–7.43		7.32–7.43
		Adults			
		60–90 years	7.31–7.42		7.31–7.42
		>90 years	7.26–7.43		7.26–7.43
Phenylalanine			mg/dL	60.5	μmol/L
	Dry blood spot		<2.1		<122
	P	Premature	2.0–7.5		121–454
		Newborn	1.2–3.4		73–205
		Phenylketonuric 2–3 days	>4.5		>272

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
		Phenylketonuric untreated	15–30		907–1,815
		Adult	0.8–1.8		48–109
			mg/day	6.05	µmol/day
	U, 24 h	10 days–7 weeks	1.2–1.7		7–10
		3–13 years	4.0–17.5		24–106
		Adult	<16.5		<100
			mg/g creatinine	0.68	mmol/mol creatinine
		Adult	2–10		1.3–6.9
Phosphate	S		mg/dL	0.323	mmol/L
Values (0–<3 years) based on Abbott Architect method		0–14 days	5.6–10.5		1.80–3.40
		15 days–<1 year	4.8–8.4		1.54–2.72
		1–<3 years	4.3–6.8		1.38–2.19
Values (3–79 years) based on OCD VITROS method		3–5 years	4.4–6.0		1.41–1.94
		6–10 years	4.4–5.7		1.41–1.85
		11–15 years M	3.8–5.9		1.24–1.91
		11–15 years F	3.6–5.6		1.16–1.81
		16–47 years	2.9–4.7		0.95–1.52
		48–79 years M	2.8–4.7		0.89–1.52
		48–79 years F	3.1–4.8		0.99–1.54
	U, 24 h		g/day	32.3	mmol/day
		Adults	0.4–1.3		12.9–42.0
Phosphatase, acid tartrate resistant 37°C	S		U/L	0.017	µkat/L
		Children	3.4–9.0		0.05–0.15
		Adult	1.5–4.5		0.03–0.08
Phosphatase, alkaline IFCC, 37°C	S		U/L	0.017	µkat/L
Values (0–<3 years) based on Abbott Architect method		0–14 days	90–273		1.50–4.55
		15 days–<1 year	134–518		2.23–8.63
		1–<3 years	156–369		2.60–6.15
Values (3–79 years) based on OCD VITROS method		3–5 years	144–327		2.45–5.56
		6–10 years	153–367		2.60–6.24
		11–15 years M	113–438		1.92–7.45
		11–15 years F	64–359		1.09–6.10
		16–21 years M	56–167		0.95–2.84
		16–29 years F	44–107		0.75–1.82
		22–79 years M	50–116		0.85–1.97
		30–79 years F	46–122		0.78–2.07
Phosphatase, alkaline (bone specific, by immunoabsorption)	S		U/L	1.0	U/L
		M	15.0–41.3		15.0–41.3
		Premenopausal female	11.6–29.6		11.6–29.6
Phosphatase, alkaline isoenzymes					
Percentage of total activity	<1 year	1–15 years	Adult	Pregnant female	Postmenopausal female
Biliary	3–6	2–5	1–3	1–3	0–12
Liver	20–34	22–34	17–35	5–17	17–48
Bone	20–30	21–30	13–19	8–14	8–21
Placental	8–19	5–17	13–21	53–69	7–15

Continued

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

			REFERENCE INTERVALS		
Analyte	**Specimen**	**Condition**	**Conventional Units**	**Conversion Factor**	**SI Units**
Renal	1–3	0–1	0–2	3–6	0–2
Intestinal	0–2	0–1	0–1	0–1	0–1
Fraction activity	<1 year	1–15 years	Adult	Pregnant female	Postmenopausal female
Biliary	0.03–0.06	0.02–0.05	0.01–0.03	0.01–0.03	0.0–0.12
Liver	0.20–0.34	0.22–0.34	0.17–0.35	0.05–0.17	0.17–0.48
Bone	0.20–0.30	0.21–0.30	0.13–0.19	0.08–0.14	0.08–0.21
Placental	0.08–0.19	0.05–0.17	0.13–0.21	0.53–0.69	0.07–0.15
Renal	0.01–0.03	0.0–0.01	0.0–0.02	0.03–0.06	0.0–0.02
Intestinal	0.0–0.02	0.0–0.01	0.0–0.01	0.0–0.01	0.0–0.01
Porphobilinogen	U, 24 h		mg/L	4.42	μmol/L
			<2.26		<10
Porphyrins, total	U, 24 h				nmol/L
					20–320
	Feces				nmol/L g dry wt
					10–200
	Erythrocytes				μmol/L erythrocytes
					0.4–1.7
Potassium (K)			mEq/L	1.0	mmol/L
	S	Premature cord	5.0–10.2		5.0–10.2
		Premature, 48 hours	3.0–6.0		3.0–6.0
		Newborn cord	5.6–12.0		5.6–12.0
		Newborn	3.7–5.9		3.7–5.9
		Infant	4.1–5.3		4.1–5.3
Values (3–79 years) based on OCD VITROS method		3–5 years	3.9–4.6		3.9–4.6
		6–79 years	3.8–4.9		3.8–4.9
			mEq/day	1.0	mmol/day
	U, 24 h	6–10 years M	17–54		17–54
		6–10 years F	8–37		8–37
		10–14 years M	22–57		22–57
		10–14 years F	18–58		18–58
		Adult	25–125		25–125
Proinsulin	S		pmol/L	1.0	pmol/L
			1.1–6.9		1.1–6.9
Prolactin	S		ng/mL	21.0	mIU/L
		Cord blood	45–539		945–11,319
Values (4 days–<19 years) based on Abbott Architect method		4–<30 days	13–213		273–4,473
		30 days–<1 year	6–114		126–2,394
		1–<19 years	4–23		84–483
		Puberty, Tanner Stage			
		1, M	3–20		63–420
		1, F	2–20		42–420
		2, M	4–19		84–399
		2, F	4–23		84–483
		3, M	4–23		84–483
		3, F	4–23		84–483
		4, M	6–20		126–420
		4, F	6–23		126–483
		5, M	7–32		147–672
		5, F	5–23		105–483
		Adult M	3.0–14.7		63.0–308.7
		Adult F	3.8–23.0		79.8–483.0
		Pregnancy, third trimester	95–473		1,995–9,933

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
Proline			mg/dL	86.9	µmol/L
	P	Premature, 1 day	0.92–4.36		80–380
		Newborn, 1 day	1.23–3.18		107–277
		1–3 months	0.89–3.73		77–325
		9 months–2 years	0.59–2.13		51–185
		3–10 years	0.78–1.70		68–148
		6–18 years	0.67–3.72		58–324
	U, 24 h	Adult	1.17–3.86		102–336
			mg/day	8.69	µmol/day
		10 days–7 weeks	3.2–11.0		28–96
		3–12 years	Trace		Trace
		Adult	Trace		Trace
			µmol/g creatinine	0.113	µmol/mol creatinine
		0–1 month	70–2,300		7.91–259.9
		1–6 months	<600		<67.8
		6 months–1 year	<300		<33.9
		1–2 years	<270		<30.5
		2–3 years	<220		<24.9
Prostate-specific antigen (PSA)	S		ng/mL	1.0	µg/L
		Free PSA			
		0–<12 years M	<0.008–0.239		<0.008–0.239
		12–<19 years M	<0.008–0.047		<0.008–0.047
		0–<19 years F	<0.008–0.097		<0.008–0.097
		Total PSA			
		1 week–<6 months M	<0.008–0.038		<0.008–0.038
		6 months–<12 years M	<0.008–0.353		<0.008–0.353
		12–<19 years M	<0.008–0.566		<0.008–0.566
		0–<1 week F	<0.008–0.039		<0.008–0.039
		1 week–<1 year F	<0.008–0.010		<0.008–0.010
		1–<19 years F	<0.008–0.015		<0.008–0.015
		Adult M			
		40–49 years	0–2.5		0–2.5
		50–59 years	0–3.5		0–3.5
		60–69 years	0–4.5		0–4.5
		70–79 years	0–6.5		0–6.5
Protein, total			g/dL	10	g/ L
		Cord	4.8–8.0		48–80
	S	Premature	3.6–6.0		36–60
Values (0–<3 years) based on Abbott Architect method		0–14 days	5.3–8.3		53–83
		15 days–<1 year	4.4–7.1		44–71
		1–<3 years	6.1–7.5		61–75
Values (3–79 years) based on OCD VITROS method		3–5 years	6.3–8.1		63–81
		6–19 years	6.8–8.2		68–82
		20–29 years	6.5–8.3		65–83
		30–79 years	6.5–7.8		65–78
	U, 24 h		mg/dL	10	mg/L
		Adult	1–14		10–140
		Excretion	mg/day	0.001	g/day
		Adult	<100		<0.1
		Pregnancy	<150		<0.15
Pyruvic acid			mg/dL	0.114	µmol/L

Continued

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

			REFERENCE INTERVALS		
Analyte	**Specimen**	**Condition**	**Conventional Units**	**Conversion Factor**	**SI Units**
	WB, arterial	Adult	0.2–0.7		0.02–0.08
	WB, venous	Adult	0.3–0.9		0.03–0.10
					mmol/day
	U, 24 h	Adult			<1.1
					mmol/mol creatinine
	U	0–1 month			24–123
		1–6 months			8–90
		6 months–5 years			3–19
		>5 years			6–9
Retinol-binding protein (RBP)			mg/dL	10	mg/L
	S	Birth	1.1–3.4		11–34
		6 months	1.8–5.0		18–50
		Adult	3.0–6.0		30–60
Rheumatoid factor (RF)	S		IU/mL	1	kIU/L
Values (0–<19 years) based on Abbott Architect method		0–14 days	9.0–17.1		9.0–17.1
		15 days–<19 years	9.0–9.0		9.0–9.0
		Adult	<7.5–14		<7.5–14
Riboflavin (vitamin B2)			μg/dL	26.6	nmol/L
	S		4–24		106–638
	Erythrocytes		10–50		266–1,330
			μg/g creatinine	0.3	μmol/mol creatinine
	U		>80		>24
			μg/day	2 .66	nmol/day
	U, 24 h		>100		>266
Selenium	S		μg/L	0.0127	μmol/L
		Neonates	<8.0 (deficiency)		<0.10 (deficiency)
		<2 years	16–71		0.2–0.9
		2–4 years	40–103		0.5–1.3
		4–16 years	55–134		0.7–1.7
		Adults	63–160		0.8–2.0
	WB (Hep)		58–234		0.74–2.97
	U, 24 h		7–160		0.09–2.03
		Toxic conc.	>400		>5.08
Sex hormone–binding globulin (SHBG)	S		nmol/L	1	nmol/L
Values (4 days–<19 years) based on Abbott Architect method		4 days–<1 month	14.4–120.2		14.4–120.2
		1 month–<1 year	36.2–229.0		36.2–229.0
		1–<8 years	41.8–188.7		41.8–188.7
		8–<11 years	26.4–162.4		26.4–162.4
		11–<13 years	14.9–107.8		14.9–107.8
		13–<15 years	11.2–98.2		11.2–98.2
		15–<19 years M	9.7–49.6		9.7–49.6
		15–<17 years F	9.8–84.1		9.8–84.1
		17–<19 years F	10.8–154.6		10.8–154.6
		Puberty, Tanner Stage			
Tanner values based on Abbott Architect method		1 M	23.4–156.8		23.4–156.8
		1, F	21.1–210.1		21.1–210.1
		2, M	27.5–133.4		27.5–133.4
		2, F	29.6–140.7		29.6–140.7
		3, M	17.4–160.1		17.4–160.1
		3, F	23.7–101.7		23.7–101.7

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
		4, M	12.2–79.4		12.2–79.4
		4, F	12.1–125.6		12.1–125.6
		5, M	7.7–49.4		7.7–49.4
		5, F	15.3–92.5		15.3–92.5
		Adult			
		20 years	13.1–53.2		13.1–53.2
		30 years	13.5–57.4		13.5–57.4
		40 years	15.3–65.3		15.3–65.3
		50 years	18.4–75.6		18.4–75.6
		60 years	22.6–87.6		22.6–87.6
		70 years	27.8–101.0		27.8–101.0
		80 years	33.8–115.4		33.8–115.4
Sodium (Na)			mEq/L	1.0	mmol/L
		Premature cord	116–140		116–140
		Premature, 48 hours	128–148		128–148
		Newborn cord	126–166		126–166
		Newborn	133–146		133–146
		Infant	139–146		139–146
Values (3–79 years) based on OCD VITROS method		3–5 years	135–142		135–142
		6–15 years	136–143		136–143
		16–49 years M	137–143		137–143
		16–49 years F	137–142		137–142
		50–79 years	136–143		136–143
	U, 24 h		mEq/day	1.0	mmol/day
		6–10 years M	41–115		41–115
		6–10 years F	20–69		20–69
		10–14 years M	63–177		63–177
		10–14 years F	48–168		48–168
		Adult M	40–220		40–220
		Adult F	27–287		27–287
Testosterone, bioavailable (Vermeulen Equation)	S		ng/dL	0.0347	nmol/L
Values (0–<19 years) based on Abbott Architect method		0–<1 year M	0–121.9		0.01–4.23
		1–<9 years M	0.29–2.88		0.01–0.10
		9–<14 years M	0.58–161.69		0.02–5.62
		14–<19 years M	12.1–346.69		0.42–12.03
		0–<1 year F	0.29–6.05		0.01–0.21
		1–<9 years F	0.29–6.05		0.01–0.21
		9–<14 years F	0.58–10.95		0.02–0.38
		14–<19 years F	3.75–23.05		0.13–0.80
		Adult M	66–417		2.29–14.5
		Adult F	0.6–5.0		0.02–0.17
Testosterone, free (Vermeulen Equation)	S		pg/mL	3.47	pmol/L
		Cord M	5–22		17.4–76.3
		Cord F	4–19		13.9–55.5
Values (0–<19 years) based on Abbott Architect method		0–<1 year M	0.03–57.2		0.1–198.4
		1–<9 years M	0.1–1.2		0.3–4.0
		9–<14 years M	0.4–72.2		1.5–250.6
		14–<19 years M	5.0–142.4		17.4–494.0
		0–<1 year F	0.1–2.6		0.3–9.1
		1–<9 years F	0.1–2.6		0.3–9.1

Continued

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
		9–<14 years F	0.3–4.7		1.0–16.4
		14–<19 years F	1.4–9.9		4.9–34.3
		Adult M	50–210		174–729
		Adult F	1.0–8.5		3.5–29.5
Testosterone, total	S		ng/dL	0.0347	nmol/L
		Cord M	13–55		0.45–1.91
		Cord F	5–45		0.17–1.56
		Premature M	37–198		1.28–6.87
		Premature F	5–22		0.17–0.76
Values (4 days–<19 years) based on Abbott Architect method		4 days–<6 months M	9–299		0.3–10.37
		6 months–<9 years M	<36		<1.24
		9–<11 years M	<23		<0.81
		11–<14 years M	<444		<15.42
		14–<16 years M	36–632		1.25–21.94
		16–<19 years M	148–794		5.13–27.55
		4 days–<9 years F	1–62		0.0–2.15
		9–<13 years F	<28		<0.98
		13–<15 years F	10–44		0.36–1.54
		15–<19 years F	14–49		0.49–1.70
Tanner values based on Abbott Architect method		1, M	<18		<0.62
		1, F	<19		<0.67
		2, M	<25		<0.85
		2, F	<20		<0.69
		3, M	<543		<18.85
		3, F	<42		<1.45
		4, M	9–636		0.30–22.08
		4, F	9–42		0.31–1.44
		5, M	100–760		3.46–26.36
		5, F	4–50		0.13–1.72
		Adult M	260–1,000		9–34.72
		Adult F	15–70		0.52–2.43
Thallium			µg/L	4.89	nmol/L
	WB (Hep)		<5		<24 .5
			mg/L		µmol/L
		Toxic	0.1–8.0		0.5–390
			µg/L	4.89	nmol/L
	U, 24 h		<2.0		<9.8
			mg/L		µmol/L
		Toxic	1.0–20.0		4.9–97.8
Threonine	P		mg/dL	84	µmol/L
		Premature, 1 day	1.14–3.98		95–335
		Newborn, 1 day	1.36–3.99		114–335
		1–3 months	0.75–2.67		64–224
		2–6 months	2.27–4.33		191–364
		3–10 years	0.50–1.13		42–95
		6–18 years	0.88–2.40		74–202
		Adult	0.94–2.30		79–193
	U, 24 h		mg/day	8.40	µmol/day
		10 days–7 weeks	1.5–11.9		13–100
		3–12 years	10.1–29.6		85–249
		Adult	14.3–46.7		120–392
			mg/g creatinine	0.95	mmol/mol creatinine
		Adult	0–28		0–27

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
Thyroglobulin (TG)	S		ng/mL	1.0	µg/L
		Adult euthyroid	3–42		3–42
		Athyroidic patient	<5		<5
Thyrotropin (thyroid-stimulating hormone) (TSH)			µIU/mL	1.0	mIU/L
	S	Premature, 28–36 weeks	0.7–27.0		0.7–27.0
		Cord blood (>37 weeks)	2.3–13.2		2.3–13.2
Values (4 days–<19 years) based on Abbott Architect method		4 days–<6 months	0.7–4.8		0.7–4.8
		6 months–<14 years	0.7–4.2		0.7–4.2
		14–<19 years	0.5–3.4		0.5–3.4
		Adults			
		21–54 years	0.4–4.2		0.4–4.2
		55–87 years	0.5–8.9		0.5–8.9
		Pregnancy	µU/mL	1.0	mU/L
		First trimester	0.1–2.5		0.1–2.5
		Second trimester	0.2–3.0		0.2–3.0
		Third trimester	0.3–3.0		0.3–3.0
	WB (heel puncture)	Newborn screen	<20		<20
Thyroxine-binding globulin			mg/dL	10	mg/L
	S	Cord	3.6–9.6		36–96
		Children			
		4 months–1 year	3.1–5.6		31–56
		1–5 years	2.9–5.4		29–54
		5–10 years	2.5–5.0		25–50
		10–15 years	2.1–4.6		21–46
		Adult M	1.2–2.5		12–25
		Adult F	1.4–3.0		14–30
		Adult F (oral contraceptive)	1.5–5.5		15–55
Thyroxine (T4), total			µg/dL	12.9	nmol/L
Values (7 days–<19 years) based on Abbott Architect method	S	7 days–<1 year	5.9–13.7		76–176
		1–<9 years	6.2–10.3		79–133
		9–<12 years	5.5–9.3		71–120
		12–<14 years M	5.0–8.3		65–107
		12–<14 years F	5.1–8.3		65–107
		14–<19 years M	4.7–8.6		61–111
		14–<19 years F	5.5–13.0		70–167
		Adult (15–60 years) F	4.6–10.5		59–135
		Adult (15–60 years) F	5.5–11.0		65–138
		>60 years	5.0–10.7		65–138
		Newborn screen			
		1–5 days	>7.5		>97
		6 days	>6.5		>84
Thyroxine, free (FT4)	S		ng/dL	12.9	pmol/L
		Newborns (1–4 days)	2.2–5.3		28.4–68.4
Values (5 days–<19 years) based on Abbott Architect method		5–15 days	1.1–3.2		13.5–41.3
		15–<30 days	0.7–2.5		8.7–32.5
		30 days–<1 year	0.9–1.7		11.4–21.9
		1–<19 years	0.9–1.4		11.4–17.6

Continued

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
		Adults (21–87 years)	0.8–2.7		10.3–34.7
		Pregnancy			
		First trimester	0.7–2.0		9.0–25.7
		Second and third trimesters	0.5–1.6		6.4–20.6
Transferrin	S		mg/dL	0.01	g/L
Values (0–<19 years) based on Abbott Architect method		0–<9 weeks	104–224		1.0–2.2
		9 weeks–<1 year	107–324		1.1–3.2
		1–<19 years	220–337		2.2–3.4
		20–60 years	200–360		2.0–3.6
		>60 years	160–340		1.6–3.4
Transthyretin (prealbumin)	S		mg/dL	10	mg/L
Values (0–<19 years) based on Abbott Architect method		0–14 days	2–12		20–120
		15 days–<1 year	5–24		50–240
		1–<5 years	12–23		120–230
		5–<13 years	14–26		140–260
		13–<16 years	18–31		180–310
		16–<19 years M	20–35		200–350
		16–<19 years F	17–33		170–330
		Adult (20–60 years)	20–40		200–400
Triglycerides	S		mg/dL	0.0113	mmol/L
Reference Intervals					
Values (0–<6 years) based on Abbott Architect method		0–14 days	82–259		0.9–2.9
		15 days–<1 year	53–258		0.6–2.9
		1–<6 years	44–197		0.5–2.2
Values (6–79 years) based on OCD VITROS method		6–29 years	35–186		0.4–2.1
		30–79 years M	44–301		0.5–3.4
		30–79 years F	35–212		0.4–2.4
Clinical decision limits		Recommended cutoff points, Child	mg/dL	0.0113	mmol/L
		0–9 years			
		Acceptable	<75		<0.9
		Borderline	75–99		0.9–1.1
		High	≥100		≥1.1
		10–19 years			
		Acceptable	<90		<1.0
		Borderline	90–129		1.0–1.5
		High	≥130		≥1.5
		Recommended cutoff points, Adult	mg/dL	0.0113	mmol/L
		Normal	<150		<1.70
		High	150–199		1.70–2.25
		Hypertriglyceridemic	200–499		2.26–5.64
		Very high	>499		>5.64
Triiodothyronine (T3), free			pg/dL	0.0154	pmol/L
	S	Cord	15–391		0.2–6.0
Values (4 days–<19 years) based on Abbott Architect method		4 days–<1 year	232–487		3.6–7.5
		1–<12 years	279–442		4.3–6.8
		12–<15 years M	289–433		4.4–6.7
		12–<15 years F	247–395		3.8–6.1

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
		15–<19 years M	225–385		3.5–5.9
		15–<19 years F	231–371		3.6–5.7
		Adult	210–440		3.2–6.8
		Pregnancy	200–380		3.1–5.9
Triiodothyronine (T3), total	S		ng/dL	0.0154	nmol/L
		Cord (>37 weeks)	5–141		0.08–2.17
Values (4 days–<19 years) based on Abbott Architect method		4 days–<1 year	85–234		1.33–3.60
		1–<12 years	113–189		1.74–2.91
		12–<15 years	98–176		1.50–2.71
		15–<17 years M	94–156		1.44–2.40
		15–<17 years F	92–142		1.42–2.18
		17–<19 years	90–168		1.38–2.58
		Adults			
		20–50 years	70–204		1.08–3.14
		50–90 years	40–181		0.62–2.79
		Pregnancy			
		First trimester	81–190		1.25–2.93
		Second and third trimesters	100–260		1.54–4.00
Tryptophan			mg/dL	49	μmol/L
	P	Premature, 1 day	0–1.23		0–60
		Newborn, 1 day	<1.37		<67
		1–16 years	0.49–1.61		24–79
		>16 years	0.41–1.94		20–95
			mg/day	4.9	μmol/day
	U, 24 h	Adult	5–39		25–191
			mg/g creatinine	0.55	mmol/mol creatinine
		Adult	<30		<16.5
Tyrosine	P		mg/dL	55.2	mmol/L
		Premature, 1 day	0–5.79		0–320
		Newborn, 1 day	0.76–1.79		42–99
		1–3 months	0.54–2.42		30–134
		2–6 months	1.30–3.91		72–216
		9 months–2 years	0.20–2.21		11–122
		3–10 years	0.56–1.29		31–71
		6–18 years	0.78–1.59		43–88
		Adult	0.40–1.58		22–87
	U, 24 h		mg/day	5.52	μmol/day
		10 days–7 weeks	4.0–7.2		22–40
		3–12 years	7.2–30.4		40–168
		Adult	12.0–55.1		66–304
			mg/g creatinine	0.62	mmol/mol creatinine
		Adult	0–23		0–14.2
Urea	S		mg/dL	0.357	mmol/L
Values (0–<3 years) based on Abbott Architect method		0–<14 days	3–23		1.0–8.2
		15 days–<1 year	3–17		1.2–6.0
		1–<3 years	9–22		3.2–7.9
Values (3–79) based on OCD VITROS method		3–5 years	9–19		3.1–6.9
		6–7 years	8–21		2.8–7.5
		8–19 years M	8–20		2.9–7.0
		20–39 years M	9–22		3.3–7.9
		40–59 years M	10–24		3.5–8.6

Continued

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

Analyte	Specimen	Condition	REFERENCE INTERVALS Conventional Units	Conversion Factor	SI Units
		8–59 years F	8–19		2.7–6.7
		60–79 years	10–26		3.6–9.2
	U, 24 h		g/day	0.0357	mol/day
			10–20		0.43–0.71
Uric acid	S		mg/dL	59.48	µmol/L
Values (0–<3 years) based on Abbott Architect method		0–14 days	2.8–12.7		167–755
		15 days–<1 year	1.6–6.3		95–375
		1–<3 years	1.8–4.9		107–291
Values (3–79 years) based on OCD VITROS method		3–<5 years	2.0–4.9		117–291
		6–8 years	1.9–5.0		116–295
		9–10 years	2.4–5.5		142–326
		11–12 years	2.6–5.8		156–345
		13–79 years M	3.7–7.7		218–459
		13–79 years F	2.5–6.2		147–366
Valine	P		mg/dL	85.5	µmol/L
		Premature, 1 day	0.34–2.70		30–230
		Newborn, 1 day	0.94–2.88		80–246
		1–3 months	1.13–3.4 1		96–292
		9 months–2 years	0.67–3.07		57–262
		3–10 years	1.50–3.31		128–283
		6–18 years	1.83–3.37		156–288
		Adult	1.65–3.71		141–317
	U		mg/day	8.55	µmol/day
		10 days–7 weeks	1.4–3.2		12–27
		3–12 years	1.8–6.0		15–51
		Adult	2.5–11.9		21–102
			mg/g creatinine	0.97	mmol/mol creatinine
		Adult	2–6		1.9–5.9
Vanillylmandelic acid (VMA)	U, 24 h		mg/day	5.05	µmol/day
		3.6 years	1–.26		5–13
		6–10 years	2.0–3.2		10–16
		10–16 years	2.3–5.2		12–26
		16–83 years	1.4–6.5		7–33
			mg/g creatinine	0.571	mmol/mol creatinine
	U	0–1 month	<27		<16
		1–6 months	<19		<11
		6 months–5 years	<13		<8
		3–6 years	4.0–10.8		2.3–6.2
		6–10 years	4.0–7.5		2.3–4.3
		10–16 years	3.0–8.8		1.7–5.0
Vitamin A	S		µg/dL	0.0349	µmol/L
Values (0–<19 years) based on Abbott Architect method		0–<1 year	8–54		0.3–1.9
		1–<11 years	28–44		1.0–1.6
		11–<16 years	25–55		0.9–1.9
		16–<19 years	29–75		1.0–2.6
		Adult	30–80		1.05–2.8
Vitamin B1 (thiamine diphosphate)			nmol/L	1	nmol/L
	WB		90–140		90–140
			ng/g Hb	0.146	µmol/mol Hb
	Erythrocytes		280–590		40.3–85.0
Vitamin B2 (see riboflavin)					

TABLE A1 Pediatric and Adult Reference Intervals for Biochemical Markers—cont'd

			REFERENCE INTERVALS		
Analyte	**Specimen**	**Condition**	**Conventional Units**	**Conversion Factor**	**SI Units**
Vitamin B6	P (EDTA)		ng/mL	4.046	nmol/L
			5–30		20–121
		Deficiency	<5		<20
Vitamin B12	S		ng/L	0.733	pmol/L
Reference Intervals					
Values (5 days–< 3 years) based on Abbott Architect method		5 days–<1 year	259–1,576		191–1,163
		1–<3 years	283–1,613		209–1,190
Values (3–79 years) based on Siemens IMMULITE method		3–5 years	310–988		229–729
		6–8 years	321–985		237–727
		9–11 years	276–969		204–715
		12–79 years	188–908		139–670
Clinical decision limits					
		Acceptable (WHO)	>201		>147
		Deficiency (WHO)	<150		<110
Vitamin C (ascorbic acid)	S		mg/dL	56.78	μmol/L
			0.4–1.5		23–85
		Deficiency	<0.2		<11
			μg/10^8 leukocytes	0.057	fmol/10^8 leukocytes
	Leukocyte		20–53		1.14–3.01
			μg/10^8 leukocytes	0.057	fmol/10^8 leukocytes
		Deficiency	<10		<0.57
Vitamin D 25(OH)D	S		ng/mL	2.5	nmol/L
Values (5 days–<3 years) based on Abbott Architect method		5–<15 days	2–34		4–85
		15 days–<3 months	6–41		15–101
		3 months–<1 year	7–47		17–118
		1–<3 years	13–55		33–137
Values (3–79 years) based on DiaSorin LIAISON method					
	P	3–5 years	13–42		33–104
		6–79 years	8–46		21–116
		Deficiency	<20		<50
$1,25(OH)_2D$			pg/mL	2.4	pmol/L
			15–60		36–144
Vitamin E	S		mg/dL	23.2	μmol/L
		Premature neonates	0.1–0.5		2.3–11.6
Values (0–<19 years) based on Abbott Architect method		0–<1 year	0.2–2.1		4.9–49.6
		1–<19 years	0.6–1.4		14.5–33.0
		Adults	0.5–1.8		12–42
Vitamin K	S		ng/mL	2.22	nmol/L
			0.13–1.19		0.29–2.64
Zinc	S		μg/dL	0.153	μmol/L
			80–120		12–18
		Deficiency	<30		<5
			mg/24 h	15.3	μmol/24 h
	U, 24 h		0.2–1.3		3–21

For additional information regarding the sources of specific presented data refer to *Tietz Textbook of Clinical Chemistry and Molecular Diagnostics*, 6th Edition, 2017.

TABLE A2 Pediatric and Adult Critical Risk Values

	CONVENTIONAL UNITS		SI UNITS	
Parameter	**Lower Limit**	**Upper Limit**	**Lower Limit**	**Upper Limit**
Albumin (children)	g/dL		g/L	
	1.7	6.8	17	68
Aminotransferases	U/L		μkat/L	
	—	1,000	—	16.7
Ammonia	μg/dL		μmol/L	
	—	187	—	110
Anion gap			mmol/L	
			—	20
Bilirubin (newborn)	mg/dL		mmol/L	
	—	15	—	257
Calcium (total)	mg/dL		mmol/L	
	6.6	14	1.65	3.5
Calcium (children)	mg/dL		mmol/L	
	6.5	12.7	1.63	3.18
Calcium (free)	mg/dL		mmol/L	
	3.1	6.3	0.75	1.6
Carbon dioxide, total			mmol/L	
			10	40
Chloride (adult)			mmol/L	
			80	120
Creatinine (adult)	mg/dL		mmol/L	
	—	5	—	442
Creatinine (children)	mg/dL		μmol/L	
	—	3.8	—	336
Creatine kinase	U/L		μkat/L	
	—	1,000	—	16.7
Glucose	mg/dL		mmol/L	
	40	500	2.22	27.8
Glucose (children)	mg/dL		mmol/L	
	46	445	2.56	24.72
Glucose (newborn)	mg/dL		mmol/L	
	30	325	1.67	18.06
Glucose, CSF (adult)	mg/dL		mmol/L	
	40	200	2.22	11.11
Glucose, CSF (children)	mg/dL		mmol/L	
	31	—	1.72	—
Lactate plasma	mg/dL		mmol/L	
	—	45	—	5
Lactate plasma (children)	mg/dL		mmol/L	
	—	36.9	—	4.1
Lactate dehydrogenase	U/L		μkat/L	
	—	1,000	—	16.7
Lipase	U/L		μkat/L	
	—	700	—	11.7
Magnesium	mg/dL		mmol/L	
	1	4.9	0.41	2
Osmolality			mOsm/kg	
			240	330
Osmolar gap			mOsm/kg	
			—	10
Phosphate	mg/dL		mmol/L	
	1	9	0.32	2.9
Potassium			mmol/L	
			2.8	6.2
Potassium (newborn)			mmol/L	
			2.8	7.8

TABLE A2 Pediatric and Adult Critical Risk Values—cont'd

	CONVENTIONAL UNITS		SI UNITS	
Parameter	**Lower Limit**	**Upper Limit**	**Lower Limit**	**Upper Limit**
Protein (children)	g/dL		g/L	
	3.4	9.5	34	95
Protein, CSF (children)	mg/dL		mg/L	
	—	188	—	1,880
Sodium			mmol/L	
			120	160
T4 (free)	ng/dL		pmol/L	
	—	3.5	—	45
Urea nitrogen	mg/dL		mmol/L	
	—	100	—	35.6
Urea	mg/dL		mmol/L	
	—	214	—	35.6
Urea nitrogen (children)	mg/dL		mmol/L	
	—	55	—	19.6
Uric acid	mg/dL		mmol/L	
	—	13	—	0.767
Uric acid (children)	mg/dL			
	—	12	—	0.708
pH				
	7.2	7.6	7.2	7.6
pCO_2	mm Hg		kPa	
	20	70	2.7	9.45
pO_2	mm Hg		kPa	
	40	—	5.3	—
pO_2 children	mm Hg		kPa	
	45	125	6.0	16.7
pO_2 newborn	mm Hg		kPa	
	35	90	4.7	12.0

For additional information regarding the sources of specific presented data refer to *Tietz Textbook of Clinical Chemistry and Molecular Diagnostics,* 6th Edition, 2017.

ABBREVIATIONS

ATP III—Adult Treatment Panel III
Amf—amniotic fluid
CSF—cerebrospinal fluid
EDTA—ethylenediaminetetraacetic acid
KDIGO—Kidney Disease Improving Global Outcomes
F—fluoride ion
Hep—heparin
ICSH—International Council for Standardization in Haematology
NGSP—National Glycohemoglobin Standardization Program
Ox—oxalate
P—plasma
PCOS—polycystic ovary syndrome
RBC—red blood cells
S—serum
U—urine
WB—whole blood
WHO—World Health Organization

SUGGESTED READINGS

Adeli, K., Ceriotti, F., & Nieuwesteeg, M. (2017). Reference information for the clinical laboratory. In N. Rifai, A. R. Horvath, & C. T. Wittwer (Eds.), *Tietz textbook of clinical chemistry and molecular diagnostics* (6th ed.) (pp. 1745–1818). St. Louis: Elsevier.

Adeli, K., Higgins, V., Nieuwesteeg, M., Raizman, J. E., Chen, Y., & Wong, S. L. (2015). Biochemical marker reference values for pediatric, adult and geriatric age groups: Establishment of robust pediatric and adult reference intervals based on the Canadian Health Measures Survey. *Clinical Chemistry*, *61*, 1049–1062.

Adeli, K., Higgins, V., Nieuwesteeg, M., Raizman, J. E., Chen, Y., & Wong, S. L. (2015). Complex reference value distributions for endocrine and special chemistry biomarkers across pediatric, adult and geriatric age: establishment of robust pediatric and adult reference intervals based on the Canadian Health Measures Survey. *Clinical Chemistry*, *61*, 1063–1074.

Bailey, D., Colantonio, D., Kyriakopoulou, L., Cohen, A. H., Chan, M. K., & Armbruster, D. (2013). Marked biological variance in endocrine and biochemical markers in childhood: establishment of pediatric reference intervals using healthy community children from the CALIPER cohort. *Clinical Chemistry*, *59*, 1393–1405.

Bevilacqua, V., Chan, M. K., Chen, Y., Armbruster, D., Schodin, B., & Adeli, K. (2014). Pediatric population reference value distributions for cancer biomarkers and covariate-stratified reference intervals in the CALIPER cohort. *Clinical Chemistry*, *60*, 1532–1542.

Bjerner, J., Biernat, D., Fosså, S. D., & Bjøro, T. (2009). Reference intervals for serum testosterone, SHBG, LH and FSH in males from the NORIP project. *Scandinavian Journal of Clinical and Laboratory Investigation*, *69*, 873–879.

Colantonio, D. A., Kyriakopoulou, L., Chan, M. K., Daly, C. H., Brinc, D., Venner, A. A., et al. (2012). Closing the gaps in pediatric laboratory reference intervals: a CALIPER database of 40 biochemical markers in a healthy and multiethnic population of children. *Clinical Chemistry*, *58*, 854–868.

Fuentes-Arderiu, X., Ferré-Masferrer, M., Gonzàlez-Alba, J. M., Escolà-Aliberas, J., Escolà-Aliberas, J., Balsells-Rosello, D., et al. (2001). Multicentric reference values for some quantities measured with Tina-Quant reagents systems and RD/Hitachi analysers. *Scandinavian Journal of Clinical and Laboratory Investigation*, *61*, 273–276.

Hashim, I. A., & Cuthbert, J. A. (2014). Critical values working group. establishing, harmonizing and analyzing critical values in a large academic health center. *Clinical Chemistry and Laboratory Medicine*, *52*, 1129–1135.

Integrated guidelines for cardiovascular health and risk reduction in children and adolescents. (2013). *NIH National Heart, Lung and Blood Institute*. Accessed March 2017. http://www.nhlbi.nih.gov/health-pro/guidelines/current/cardiovascular-health-pediatric-guidelines.

Ichihara, K., Ceriotti, F., Kazuo, M., Huang, Y. Y., Shimizu, Y., Suzuki, H., et al. (2013). The Asian project for collaborative derivation of reference intervals: (2) results of non-standardized analytes and transference of reference intervals to the participating laboratories on the basis of cross-comparison of test results. *Clinical Chemistry and Laboratory Medicine*, *51*, 1443–1457.

Kelly, J., Raizman, J. E., Bevilacqua, V., Chan, M. K., Chem, Y., Quinn, F., et al. Complex reference value distribution and partitioned reference intervals across the pediatric age for 14 special chemistry and endocrine markers in the CALIPER cohort of healthy community children and adolescents. *Clinica Chimica Acta*, 450, 169–202.

Konforte, D., Shea, J. L., Kyriakopoulou, L., Colantonio, D., Cohen, A. H., Shaw, J., et al. (2013). Complex biological pattern of fertility hormones in children and adolescents: a study of healthy children from the CALIPER cohort and establishment of pediatric reference intervals. *Clinical Chemistry*, *59*, 1215–1227.

Kushnir, M. M., Blamires, T., Rockwood, A. L., et al. Liquid chromatography – tandem mass spectrometry assay for androstenedione, dehydroepiandrosterone, and testosterone with pediatric and adult reference intervals. *Clinical Biochemistry*, 56(7), 1138–1147.

KDIGO Clinical practice guideline for the evaluation and management of chronic kidney disease. (2013). *Kidney International Supplements*, *3*(1). Accessed March 2017. http://www.kdigo.org/clinical_practice_guidelines/pdf/CKD/KDIGO_2012_CKD_GL.pdf.

Raizman, J. E., Cohen, A. H., Teodoro-Morrison, T., Wan, B., Khun-Chen, M., Wilkenson, C., et al. (2014). Pediatric reference value distributions for vitamins A and E in the CALIPER cohort and establishment of age-stratified reference intervals. *Clinical Biochemistry*, *47*, 812–815.

Raizman, J. E., Quinn, F., Armbruster, D. A., & Adeli, K. (2015). Pediatric reference intervals for calculated free testosterone, bioavailable testosterone and free androgen index in the CALIPER cohort. *Clinical Chemistry and Laboratory Medicine*, *53*, e239–e243.

Stagnaro-Green, A., Abalovich, M., Alexander, E., Azizi, F., Mestman, J., Negro, R., et al. (2011). Guidelines of the American Thyroid Association for the diagnosis and management of thyroid disease during pregnancy and postpartum. *Thyroid*, *21*(10), 1081–1125.

Thomas, L. (2003). Critical limits of laboratory results for urgent clinician notification. *eJIFCC*, *14*(1). Accessed March 2017 http://www.ifcc.org/ifccfiles/docs/140103200303.pdf.

World Health Organization. (2011). Use of glycated hemoglobin A1c in the diagnosis of diabetes mellitus. *World Health Organization*, 1–25.

INDEX

Note: Page numbers followed by "f" indicate figures, "t" indicate tables, and "b" indicate boxes.

1000 Genomes Project, 925

A

ABCA1, 405
Abortion, threatened, 869
Absorbance (A), 129–130, 129f, 130t
Absorbance detector, in liquid chromatography, 196–197, 197t
Absorption
  of drugs, 543–544
  in optical methods, 129t
Absorption maximum, 130t
Absorption spectrum, 133
Absorptivity (a), 130t
Accuracy, 14, 14t
Aceruloplasminemia, 490
Acetaminophen, 571
Acetest, 388
Acetone, 570
Acetyl coenzyme A (CoA), 393
Acetylcholine, 574
Acetylcholine acetylhydrolase, 328
Acetylcholinesterase (AChE), 574
Acid(s)
  definition of, 688
  fatty. *see* Fatty acids
Acid citrate dextrose, 83
α_1-Acid glycoprotein, reference intervals for, 989t–1025t
Acid phosphatase, 329
Acid-base balance
  bicarbonate, 688
  classification of, 692t
  definition of, 688–689
  Henderson-Hasselbalch equation, 688–689
  Na^+-H^+ exchange, 691
  pH, 688
  physiology, 688
  pK, 688
  renal mechanisms in regulation of, 691
  respiratory mechanism in regulation of, 690–691
Acid-base disturbance
  metabolic acidosis. *see* Metabolic acidosis
  metabolic alkalosis. *see* Metabolic alkalosis
  respiratory acidosis. *see* Respiratory acidosis
  respiratory alkalosis. *see* Respiratory alkalosis
  salicylate intoxication, 694
Acid-base metabolism, physiology and disorders of, 679–698
Acid-base status, 688–689
Acid-citrate-dextrose (ACD), 963
Acidemia, 688, 885–887
Acidosis, 663
  lactic, 383, 694
  metabolic, 663
    anion gap, 693–694, 693f, 693b
    bicarbonate for, 692–695
    characteristics of, 692t
    compensatory mechanisms in, 695
  renal tubular, 669, 695
  respiratory, 692t, 696, 696b–697b
    compensatory mechanisms in, 696
Acinus, of liver, 702, 703f
Acrodermatitis enteropathica, 493
Acromegaly, 778
ACS. *see* Acute coronary syndrome
Activation energy barrier, 222f
Activator, 226
Active center, 217
Active transport, 682
Activity, 146
Acute abdomen, 741, 741b
Acute coronary syndrome (ACS), 633
  atherosclerosis, 633
  causes of, 633
  clinical history in diagnosis of, 633–634
  definition of, 630
Acute hepatitis, 711–712, 711t
Acute intermittent porphyria (AIP), 530t
Acute kidney injury (AKI), 660, 664–666, 665t, 666b
Acute myocardial infarction (AMI), 630
  definition of, 634b
  diagnosis of, 634, 634b
  electrocardiogram for, 633f
  guidelines and, 640, 642f
  precipitating factors, 633
Acute pancreatitis, 323, 324f, 741–742, 742b
Acute phase response, 305, 306b
Acute photosensitivity, without skin fragility, 534
Acute porphyrias, 525, 531, 532b
  clinical features of, 531, 531t
  differentiation between, 534
  genetic testing in, 538b
  management of, 531
  patient presenting with suspected, 533–534
Acute renal failure (ARF), 660
Acute respiratory viral infection, 965
Acute rhabdomyolysis, 316
Acute tubular necrosis (ATN), 664
Acute viral hepatitis, 712
Acute-phase response, 467
Acylcarnitine, 891
Acylcholesterol acyltransferase, 393, 395f
Acylglycerols, 396–399, 398f
Adalimumab (ADA), for inflammatory bowel disease, 735
Adapter, 956
Addison disease, 793–794, 795f–796f, 795b, 797t
Adenohypophysis
  description of, 774–776, 775t
  hormones produced by, 774
Adenomas, toxic, 827
Adenosine triphosphate
  hydrolysis, 656
  production of, 395
Adiponectin, 442t–444t
Adrenal carcinoma, 802
Adrenal cortex, 787–810, 808b
  hormones secreted by, 788
  laboratory evaluation of
    aldosterone, 806–807
    cortisol, 805–806
    11-deoxycortisol, 807
    specimens for, 804
  suppression tests, 796–798
Adrenal cortex disorders
  congenital adrenal hyperplasia, 801–802
  Cushing syndrome. *see* Cushing syndrome
  description of, 791–803
  hypoaldosteronism, 794–795
Adrenal glands
  anatomy of, 788–789, 789t
  steroid hormones secreted by. *see* Steroid hormones
Adrenal insufficiency, 793–794, 795f–796f, 795b, 797t
Adrenal medullary system, 456
Adrenal tumors, 802
Adrenarche, 844
Adrenocortical steroids
  circulation of, 791
  glucocorticoids. *see* Glucocorticoids
  regulation of, 789–791
Adrenocorticotropic hormone (ACTH)
  in adrenal insufficiency, 793
  biochemistry of, 780
  in Cushing syndrome, 795–796
  deficiency, 796b
  definition of, 774, 780
  reference intervals for, 989t–1025t
  stimulation tests, 793, 794b
Adsorption chromatography, 191f
Adult-onset diabetes, 605
Advanced glycation end products, 624–625
Advanced Medical Technology Association, 5
Adverse drug reactions (ADR), 898–899
Affinity, 238

Affinity biosensors, 159, 165b
Affinity chromatography, 193–194, 193f, 298
  glycated hemoglobin measurements using, 622
Affinity sensors, 163–165
  for DNA analysis
    using electrochemical labels, 164–165, 164f
    using fluorescent labels, 164
Ag/AgCl electrode, 147
Agarose, for electrophoresis, 172
Age, glomerular filtration rate and, 659
Agglutination assays, 249
Agilent 2100 Bioanalyzer (Agilent Technologies Inc., Santa Clara, CA), 173
Alanine, 290t–291t
  reference intervals for, 989t–1025t
Alanine aminotransferase
  in liver disease, 317
  reference intervals for, 989t–1025t
Albumin, 301
  critical risk value, 1026t–1027t
  glycated, 624
    determination of, 624
  measurement of, sample collection for, 360
  reference intervals for, 989t–1025t
  synthesis, by liver, 703
  total protein and, 360–363
  urinary, 661t
    analytical methods and traceability, 362, 363b
Albumin-Creatinine Ratio, reference intervals for, 989t–1025t
Albuminuria, 360, 625–627
  definitions of, 362
  proteinuria and, 363
  reference intervals for, 362
Alcohol, 566t, 568–571
  ethanol. *see* Ethanol
  ingestion of, 566t
  isopropanol, 570
  methanol, 570
Alcoholic hepatitis, 703
Alcoholic liver disease, 717–718
Aldehyde, 378, 378f
Aldolase, 317
Aldoses, 378, 378f
Aldosterone, 442t–444t
  anti-aldosterone antibodies, 806–807
  description of, 789–790
  hypoaldosteronism, 794–795
  measurement of, 806–807
  production of, 687
  reference intervals for, 989t–1025t
Aldosterone-producing adrenal adenoma, 802
Aliquots, 261
Aliquotter, 257
Alkalemia, 688
Alkaline phosphatase, 319–321, 334t
  biochemistry of, 319–320
  in bone disease, 320
  clinical significance of, 320
  electrophoretic separation of, 321
  forms of, 320
  hepatobiliary disease evaluations using, 320, 320t
  immunoassays for, 244
  increase, conditions, 320
  intestinal, 319–320
  isoenzyme content, 320–321
  macro-, 321
  methods of analysis for, 320–321, 321f
  in pregnancy, 320
Alkalosis
  metabolic, 695–696, 695b
    characteristics of, 692t
    chloride-resistant, 695b, 696
    conditions that cause, 693b
  respiratory, 696–697, 697b
    characteristics of, 692t
    compensatory mechanisms in, 695b, 696–697
    in salicylate overdosage, 571–572
Allele, 898–899, 916
Allele-specific PCR, 941
Alloisoleucine, 289–291
Allothreonine, 289–291
Allowable total error, specifications for, 59–60
Allozymes, 218
Alpha-thalassemia silent, 503
Alpha-thalassemias, 503
Alprazolam, 578t
Aluminosis, 592
Aluminum (Al), 592–593
  reference intervals for, 989t–1025t
Alzheimer disease, protein folding and, 298
Amenorrhea
  differential diagnosis of, 847, 848b, 849t
  primary, 847
  secondary, 847–849
American Association for Clinical Chemistry (AACC), 4
AMI. *see* Acute myocardial infarction
Amikacin, 552t
Amino acid(s), 288–312, 311b. *see also specific amino acids*
  acid-base properties of, 291
  analysis of, 294–295
  basic biochemistry of, 289–292, 291f
  in blood, 294
  concentrations of, 294
  essential, 292
  hydrophobicity of, 292
  ionization constants of, 290t–291t
  metabolism of, 293–294, 293f
    disorders, 885–891, 892b
  in plasma, 292
  renal excretion of, 294
  solubility of, 292
Amino acid(s) *(Continued)*
  structure and chemical properties of, 290t–291t, 292f
  supply and transport of, 292–293
Amino acid-derived hormones, 441
Aminoacidopathies, 885–887, 891
Aminoaciduria, 294
Aminoglycosides, 552–553, 552t
5-Aminolevulinate synthase (ALAS), 527
5-Aminolevulinic acid (ALA), 525, 536
5-Aminolevulinic acid dehydratase, 527
Aminotransferases, 317–319, 317f, 318b, 1026t–1027t
  alanine, in liver disease, 317
  aspartate, in liver disease, 317
  biochemistry, 317
  in cirrhosis, 318
  clinical significance of, 318
  methods of analysis for, 319, 319f
  nonalcoholic fatty liver disease effects on, 318
  in tissue, 318t
Amiodarone, 818t
Amitriptyline, 572f
Ammonia, 1026t–1027t
  metabolism, by liver, 705, 705f
  renal production of, 691
Ammonium ions, 691
Amnion, 863
Amniotic fluid
  composition of, 865t
  fetal lung maturity tests, 880
  α-fetoprotein, 873, 877
  functions of, 864–865, 865b
  particulate matter in, 865
  specimen collection, 86
  volume of, 864–865
Amobarbital, 577t
Amperometric detection, enzyme-based biosensors with, 160–162, 160f–161f
Amperometric PO_2 sensors, 156
Amperometry, 153–157, 157b
  applications of, 155–157, 156f–157f
  basic concepts of, 153–155, 154f
Amphetamine, 575
  designer, 575–577
  type stimulants, 575, 575f
Ampholyte, 168
Amplification
  bridge, 957, 958f
  endpoint, 943
  loop-mediated, 942
  multiplex ligation-dependent probe, 948
  polymerase chain reaction. *see* Polymerase chain reaction
  rolling circle, 943, 959
  signal, 938, 943
  strand displacement, 942
  target, 938
  techniques, 938–943, 938t, 938b
  transcription-mediated, 942, 942f
  whole-genome, 943

Amylase, 325–327, 326b
biochemistry of, 325
clinical significance of, 324f, 325, 325t
methods of analysis for, 326–327, 326f
reference intervals for, 989t–1025t
α-Amylase, 325
Amylin, 442t–444t
Amyotrophic lateral sclerosis, 298
Analgesics
acetaminophen and, 571
definition of, 571–572
nephropathy caused by, 670
opioid, 564–565
salicylate, 571–572
Analysis, methods for, 53–54
homogeneity as, 54, 54f
normal distribution as, 53–54
outliers as, 53, 53f
steady state as, 53
Analyte, 13, 117, 562
in chromatography, 185–186, 186f
Analytical assays
accuracy of, 14, 14t
calibration, 13–14, 14f
technical validity of, 13–16, 16b
trueness of, 14, 14t
Analytical bias, 91–93
specifications for, 59
Analytical considerations, 638–639, 639f
Analytical enzymology, 229–234, 234b
Analytical evaluation, of methods, 7–37, 35b
Analytical imprecision, specifications for, 59
Analytical measuring range, 15–16
Analytical performance
criteria, 9
specifications, 58–59
Analytical reagent, 232–233
Analytical reagent-grade chemicals, 110
Analytical results, quality of, 109
Analytical sensitivity, 16
Analytical specificity, interference and, 16
Analytical stability, 94–95
Analytical variation (CV_A), 52
Analytical weight, 116–117
Anaplastic lymphoma receptor tyrosine kinase, 336t
Androgen(s)
adrenal, 789, 791
biochemistry of, 835–838, 836f–837f
blood transport of, 835
chemical structure of, 836f
index, 857–858
metabolism, 788
testicular feminization syndrome, 838–839
Androgen insensitivity syndrome, 838–839
Androstanediol, 836f
Androstenedione, 789
Androstenedione (LC-MS/MS), reference intervals for, 989t–1025t
Anemia
in chronic kidney disease patients, 663
copper and, 490
hemolytic, 327
Anemia of chronic disease (ACD), 508–509, 515
Anemic hypoxia, 432
Anencephaly, 864–865, 873
Angina, 630
Angiotensin converting enzyme (ACE), 655
Angiotensin I, 790, 808
Angiotensin II, 655, 790, 808
Angiotensin-converting enzyme (ACE), 790
Angiotensinogen, 790
Angiotensins, 296
1,5-Anhydroglucitol, 625
Anion exchangers, 110
Anion gap (AG), 1026t–1027t
description of, 565–566
metabolic acidosis, 693–694, 693f, 693b
normal, 694–695
Anion-exchange chromatography, 192
Anodic stripping voltammetry (ASV), 156
ANOVA, 54
Antibiotics, 552–554
antifungal, 554
pharmacokinetic parameters of, 552t
Antibodies. *see also* Immunoglobulins
autoimmune, 611
definition of, 237
enzyme reaction inhibition by, 228
islet cell cytoplasmic, 611
monoclonal, 238
type 1 diabetes mellitus, 605
Antibodies against tissue transglutaminase (TGA), in celiac disease, 731
Antibody to HCV (anti-HCV), 710
Antibody to the HBcAg (anti-HBc), 709
Antibody to the HBsAg (anti-HBs), 709
Anticholinergic toxidrome, agents related to, 572–573
Anticoagulants
amylase activity inhibited by, 326–327
heparin. *see* Heparin
Antidepressant, 905. *see also* Tricyclic antidepressants
Antidiuretic hormone, 442t–444t, 655
characteristics of, 781, 782f, 784b
syndrome of inappropriate antidiuretic hormone, 683f, 684, 783–784, 784t
Antidotal treatment, 591
Antiepileptic drugs, 548–552
pharmacokinetic parameters of, 551t
Antifungal antibiotics, 554
Antigen, 237
excess, in light-scattering measurements, 143
Antigen-antibody binding, 238–239
binding forces, 238
cationic salts effect on, 239
chemical factors of, 238–239
Antigen-antibody binding *(Continued)*
ion species, 239
ionic strength effects, 239
polymer effect on, 239
reactions
mechanism of, 238
precipitin, 238, 238f
van der Waals-London dipole-dipole interactions, 238
Antihistamines, 573–574
Antimony (Sb), 593
Anti-Müllerian hormone, 838–839
Antimuscarinic agents, 573–574
Antineoplastic agents, 554–555
Antioxidant, copper and, 490
Antiparallel configuration, 917
α_1-antiprotease inhibitor (AAT), 736
Antipsychotics, 573, 573t
phenothiazines, 573t
Antiretroviral therapy (ART), 962
Antistreptolysin-O (ASO), reference intervals for, 989t–1025t
Anti-tissue necrosis factor (anti-TNF) drugs, for inflammatory bowel disease, 735
α_1-Antitrypsin, 301
deficiency, 298, 717
reference intervals for, 989t–1025t
synthesis, by liver, 704
Apoenzyme, 229, 317
Apolipoproteins, 402, 402t
Al, 989t–1025t
analysis of, 413–416
B-48, 392–393
B-100, 409, 989t–1025t
measurement of, 415–416
Apoptosis, 706
Applied pharmacokinetics, 542
absorption in, 543–544
basic concepts and definitions of, 542–543
biotransformation in, 545
change in the body, 543, 544b
distribution in, 544–545
excretion in, 545
fundamental principles of, 542–548
metabolism in, 545
Aprobarbital, 577t
Arachidonic acid, 658
Arginase, 494
Arginine, 290t–291t
reference intervals for, 989t–1025t
Arginine vasopressin, 781–782, 781f
Argininosuccinate lyase deficiency, 890
Argininosuccinate synthetase deficiency, 890
Arrhythmia, 630
Arsenazo III, 748–749
Arsenic (As), 593–594
reference intervals for, 989t–1025t
Arterial blood gases. *see also* Blood gases
testing, preanalytical aspects of, 46–47

Arterialized capillary blood, 434–435
Artifacts, in serum protein analysis, 176–177
AS trait, 505
Ascending loop of Henle, 657
Ascites, 717–718
Ascitic fluid, specimen collection, 85
Ascorbic acid, 989t–1025t. *see also* Vitamin C
Asparagine, 290t–291t, 989t–1025t
Aspartate, 290t–291t
Aspartate aminotransferase
  in liver disease, 317
  reference intervals for, 989t–1025t
Aspartic acid, 989t–1025t
Aspartyl proteases, 299
Aspirin. *see* Salicylates
Assay comparison, 16–26
  basic error model, 17–18, 18f
  data model, 18
  difference plot, 19, 19f
  regression analysis, 19–26, 20f
    application of, 25–26
    in cases of proportional random error, 23, 24f
    correlation coefficient, 23, 24f
    definition of, 16
    Deming, 20–22
    estimated regression line, evaluation of random error around, 22
    linearity testing for, 23–25, 24f
    nonparametric, 25
    ordinary least-squares, 20–22, 21f–22f
    systematic differences between methods obtained on, interpretation of, 25
Assay validation, for tumor markers, 338
Assays
  agglutination, 249
  chorionic gonadotropin, 876
  for creatine kinase, 316
  immunoassays. *see* Immunoassays
  selection of, 9, 10f, 10b
Assembly, 923–924
Assisted reproductive technologies, 876
Asymmetric polymerase chain reaction, 941
Atherosclerosis, 392, 633
Atherosclerotic plaque, 633
Atmospheric pressure chemical ionization, 205–206, 205f
Atmospheric pressure photoionization, 203
Atomic absorption spectrophotometry, 134–135, 134f
Atrial natriuretic peptide (ANP), 296, 442t–444t, 656
Atypical antipsychotics, 573t
Atypical gene, 328
Authors, responsibility of, 5
Autism, 599
Autocrine, 441
Autoimmune hepatitis (AIH), 703, 715–716, 716t
Autoimmune thyroid disease, 823
Automated analyzers
  centrifugal, 261
  closed, 263–264
  description of, 260
  discrete, 264
  interface to, 257
  open, 263–264
Automated systems, for electrophoresis, 173
Automatic pipettes, 112–113, 113f–114f
Automation/automated systems, 251–272
  of analytical processes, 260–266, 260b
  areas of, 266–271, 269b
    clinical chemistry, 266–267, 266f–267f
    coagulation, 268
    hematology, 268, 269f
    immunochemistry and immunology, 267, 267f
    microbiology, 269, 270f
    molecular diagnostics, 270–271
    transfusion medicine, 268–269, 269f
    urinalysis, 267–268, 268f
  basic concepts of, 260–261
  benefit of, 252
  conveyor belts, 257–259, 258f
  data handling in, 265
  definition of, 252
  instrument clusters in, 265, 265f–266f
  measurement approaches, 264
  microtiter plate systems, 265–266
  nephelometry, 267
  pipetting stations, 266
  polymerase chain reaction, 271
  postassay processing of, 264–265
  practical considerations for, 259–260, 259b
  problems of integration of, 260
  process control in, 265
  repeat testing of, 265
  requirements for, 259, 259b
  sequencing, 271
  signal processing in, 265
  technologies for, 253b
  total laboratory, 256–259, 256f, 256b
  turbidimetry, 267
Autosomal dominant acute porphyria, diagnosis of, 531
  presymptomatic, 535
Autosomal dominant diseases, 969–971
Autosomal dominant polycystic kidney disease (ADPKD), 669
Autosomal recessive disorders, 967–969, 973b
Autosomal recessive inheritance, 884, 884f
Autosome, 927
Autoverification, 265
Avian influenza (strain H5N1), 120
Avidity, 238
Avogadro's hypothesis, 430, 430t
Azathioprine (AZA), for inflammatory bowel disease, 735
Azotemia, 663

B

Bacterial overgrowth, intestinal, 732–733
Balances, types of, 116
Band-broadening, in chromatography, 184
Bandpass, 132
Bar code identification systems, in point-of-care testing devices, 275
Bar coding
  reading stations, 257
  reagent identification using, 263
  specimen identification using, 253
Barbiturates, 577
  analytical methods of, 577
Bartter syndrome, 669–670
Basal body temperature, 854–855
Basal metabolic rate, 813
Base, 688
Base pair, 923
Base peak, 202
Basic techniques
  principles of, 108–126
  procedures and, 111–118
Beer's law, 130, 130f
Bence-Jones proteins, 305, 668–669
  as tumor marker, 333
Benzethonium chloride, 360–361
Benzodiazepines, 577–578, 584
  analysis of, 577
  definition of, 578
  half-life of, 578t
Benzylpiperazines, 576t
Beriberi, 477
Beta-blockers, 542
Beta-thalassemias, 503–504, 504f
  intermedia, 504
  major, 503–504
  minor, 504
Between-subject variation (CV_G), 52
Bhutani nomogram, 519–520, 520f
Bias
  calibration, 17–18
  definition of, 13–14
  random, 17–18
Bicarbonate, 426–429, 688
  filtered, reclamation of, 691
  metabolic acidosis treated with, 692–695
  pH, 689
  reabsorption of, 656–657
Bile acids
  cholesterol, converted to, 393, 396f
  malabsorption, 733, 733b
  metabolism, by liver, 703
  for micelle formation, 392–393
  synthesis, 396f
Biliary atresia, 521
Biliary drainage, in liver, 701
Bilirubin, 499–523, 703
  analytical methods for, 521–522
  biochemistry of, 518–519, 519f
  chemistry of, 518, 518f
  clinical significance of, 519–521
  conjugated, 521

Bilirubin *(Continued)*
critical risk value, 1026t–1027t
definition of, 518
direct, 521–522
direct, reference intervals for, 989t–1025t
indirect, 521
total, 521–522
reference intervals for, 989t–1025t
unconjugated, 519
Binding process, 159
Bioavailability, 544
Bio-barcode immunoassay, 247, 248f
Biocatalytic reaction, 159
Biochemistry, 637–638
Bioinformatics, 928, 929f, 929b
Biological hazards, 121–123, 122f
Biological variation, 51–63, 62b
definition of, 51
generation of data on components of, 52–54
data analysis, 54
design of studies, 52
methods for analysis, 53–54
in health and disease, 55
nature of, 51–52
random, 52, 52t
Biomarkers. *see* Cardiac biomarkers
Biopsies, for *Helicobacter pylori*, 728
Biosensors, 159–165, 160b, 277
enzyme-based
with amperometric detection, 160–162, 160f–161f
with optical detection, 162–163
with potentiometric and conductometric detection, 162, 162f
Biotin, 243, 486–487, 486f, 989t–1025t
Biotin-responsive basal ganglia disease, 487
Biotransformation, 543, 545
Bitot spots, 473
Biuret method, for total protein measurement, 306
Biuret reaction, 306
Bladder cancer, tumor markers in, 343
relevant to, 343–344
Bladder tumor-associated antigens (BTAs), 343
Bland-Altman plot, 19, 19f
Blank reagent, 131
Blastocyst, 863
Blood. *see also* Plasma; Serum
acid-base abnormalities, 692–695, 692f
androgen transport in, 835
estrogen transport in, 840
ethanol, 570
fetal development of, 867–868
lead, 597
oxygen in, 431–432
oxygen saturation in, 432–433
porphyrins, analysis of, 536, 538f
pregnancy-related changes in, 865–866
steroid hormones and, 804
Blood *(Continued)*
testosterone transport in, 835
venipuncture for collection of. *see* Venipuncture
Blood collection tubes, 81
Blood gases, 437b
aqueous fluid control materials for, 436–437
behavior of, 429–431, 429b
blood-based and fluorocarbon-based control materials for, 436
conversion of, 429b
descriptors used in, 429b
instrumentation for, 435
measurements, 430t
analytical error in, 437
Henderson-Hasselbalch equation in, 431
instrumentation, 436f
physical principles in, 430t
prefixes for, 429b
monitoring for, 437–438
partial pressure of, 429
pH and, 429
quality assurance and quality control in, 435–436
Blood glucose
concentration, regulation of, 380
glycated hemoglobin and, 620–623
growth hormone effects on, 777, 777b
hormones regulating hormones, 606–609, 606f
monitoring of, 618–619
self-monitoring of, 617–618
Blood sampling, patient preparation for, 40
Blood specimen, 78–84
anticoagulants and preservatives added to, 82–83
acid citrate dextrose, 83
EDTA, 83
iodoacetate, 83
oxalates, 83
sodium citrate, 83
sodium fluoride, 83
centrifugation of, 87
collection of
arterial puncture for, 82
on filter paper, 82
site effects on, 83–84
skin puncture for, 81–82, 82f, 82t
composition of, site effects on, 83
description of, 78, 82
hemolysis of, 87
order of draw, 80, 81t–82t
stress effects on, 80
transport of, 87–88
Blood urea nitrogen (BUN), 369
Blotting
description of, 239–240, 239f
Southern. *see* Southern blotting
techniques, 171
Western. *see* Western blotting
BNP. *see* B-type natriuretic peptide
Body fluids, composition of, 680, 681t
Body water
extracellular compartment, 680–682, 681t
intracellular compartment, 680–682
total, 681f
Bohr effect, 433, 502
Boiling point, 427
Bonded phase gas chromatography, 187
Bone
in malignancies, 771
metabolism, hormones regulating, 756–765
calcitonin, 763–764
parathyroid hormone. *see* Parathyroid hormone
parathyroid hormone-related protein, 764–765, 765f
vitamin D. *see* Vitamin D
mineralization of, 745
remodeling, 746, 746f
resorption. *see* Bone resorption
Bone alkaline phosphatase, 766–767
Bone collagen, 768
Bone marrow
examination of, 518
tacrolimus in, 556t
Bone resorption
biochemical markers of, 767–769
collagen crosslinks in, 767–768
C-telopeptide CTX assay in, 768
C-telopeptide ICTP assay in, 768
deoxypyridinoline and pyridinoline in, 768
N-telopeptide in, 768
tartrate-resistant acid phosphatase in, 768–769
Bone turnover, biochemical markers of, 765–769
preanalytical and analytical variables of, 769
Bowman capsule, 654
Boyle's law, 430, 430t
BRAF, 336t
Brain natriuretic peptide, 296. *see also* B-type natriuretic peptide
Breast cancer
prognosis of, multigene profiling in, 344
therapy prediction for, 344
tumor markers for, 344
management of, 344
monitoring of, 345–357
prognosis of, 344
screening and diagnosis, 344
Breath ethanol, 570
Breath tests, 728
Bridge amplification, 957, 958f
Brown, Francis H., 2
Brush-border disaccharidases, 725, 731
Brush-border oligosaccharidases, 725, 726t

B-type natriuretic peptide (BNP), 442t–444t, 640–647, 643f
description of, 630–631
values and outcomes, 648f
Buccal cells specimen collection, 86
Buffer, 109
carbonic acid, 689
for electrophoresis, 171–172
hemoglobin, 689–690
phosphate, 689
plasma proteins, 689–690
Bullae, patients with, 534
Buprenorphine, 582f
Busulfan, 554–555
Butabarbital, 577t
Butalbital, 577t
1,4-Butanediol, 585
Butyrylcholinesterase, 328, 574

C

C cells, 812
^{13}C mixed-chain triglyceride breath test, for fat absorption, 740
Cadmium (Cd), 594–595
reference intervals for, 989t–1025t
Caffeine, 560
Calcitonin, 335t, 442t–444t, 763–764, 812
reference intervals for, 989t–1025t
Calcium, 746–752, 752b
biochemistry of, 746–747
chronic kidney disease-related disturbances of, 662–663
clinical significance of, 747–748
critical risk value, 1026t–1027t
distribution of, 749, 749b
estrogen effects on, 840
free (ionized), 749–751
anticoagulants in, 751
calibrators and quality controls in, 751
interferences in, 750
interpretation of, 752
pH, 750, 751f
preanalytical errors in measurement of, 751
reference intervals for, 751–752, 989t–1025t
specimen requirements in, 750–751
hypercalcemia, 748
hypocalcemia, 747–748
measurement of, 748–751
physiology of, 746–747, 747t
variation in, 752
regulation of, via parathyroid hormone, 747, 747f
total
adjusted or corrected, 749, 749b
interferences in, 749
interpretation of, 752
measurement of, 748–749
photometric methods for, 748–749
preanalytical errors in measurement of, 750b, 751
Calcium *(Continued)*
reference intervals for, 751, 989t–1025t
specimen requirements in, 749
spectrometry methods for, 749
urinary, 752
Calculated creatinine clearance (C_{Cr}), 660
Calibration, 102
definition of, 13–14, 14f
hierarchy of, 26f
Calibration bias, 17–18
Cancer. *see also specific cancer*
characterization, 930
diagnosis of, tumor markers and, 337
genetic predisposition to, 333
overview of, 331–333
screening for, tumor markers and, 337
of unknown primary, tumor markers for, 356
management of, 356–357
Cancer antigen 15-3, 335t, 344
reference intervals for, 989t–1025t
Cancer antigen 19-9, 335t, 339t–340t
in pancreatic cancer, 353
clinical applications of, 353–354
reference intervals for, 989t–1025t
Cancer antigen 125, 334t, 339t–340t
for ovarian cancer, 352
reference intervals for, 989t–1025t
response, 353
Cancer genomics, 928–930
trends in, 933, 933b
Cannabinoids. *see also* Marijuana.
definitions of, 578–579
Capillary columns, 188
Capillary electrophoresis (CE), 168, 173–174, 506, 622
detection of
direct, 174
improving limits of, 177
indirect, 174
general operations for, 173–174, 173f
sample injection for, 174
surface effects in, 176–177
technical considerations for, 175–177
temperature effects in, 175–176
Capillary gel electrophoresis (CGE), 174–175
Capillary isoelectric focusing electrophoresis, 506
Capillary zone electrophoresis (CZE), 173b, 174
Carbamate insecticide, 574
Carbamazepine, 549, 551t, 900
Carbamazepine-10,11-epoxide, 549
Carbohydrate(s), 377–390, 388b
analytical methodology in, 384–389
biochemistry and physiology of, 380–384
cellulose, 380
chemistry of, 378–380
description of, 378
disaccharides, 379–380, 379f
functions of, 378
glycogen, 380
Carbohydrate(s) *(Continued)*
glycoproteins, 380
monosaccharides, 378–379
polysaccharides, 380
starch, 380
stereoisomers, 378, 378f
three-carbon, 378f
Carbohydrate disorders
diabetes mellitus. *see* Diabetes mellitus
glycogen storage diseases, 383–384
hypoglycemia. *see* Hypoglycemia
metabolism, 895
Carbon dioxide
critical risk value, 1026t–1027t
dissolved, 688
partial pressure of
determination of, 433–437
electrodes for, 151–153, 151f–152f
reference intervals for, 437, 989t–1027t
schematic diagram of, 690f
temperature correction of measured, 437, 437t
total, 426–429
reference intervals for, 989t–1025t
Carbon monoxide (CO), 567–568
Carbonic acid buffer system, 689
Carbonic anhydrase, 691
Carboxyhemoglobin, 431, 501–502
γ-Carboxylation, of proteins, 298
Carcinoembryonic antigen (CEA), 334t, 339t–340t
for colorectal cancer, 345
reference intervals for, 989t–1025t
Carcinoid syndrome, 461, 487–488
Carcinoid tumors, 460–461, 461b
Cardiac biomarkers, 635–648, 635f, 637t
in acute coronary syndrome, 634, 640b
B-type natriuretic peptide, 643f
in congestive heart failure, 636t
definition of, 634
myoglobin, 648
with potential and limited utilization in clinical practice, 647–648
troponins I and T, 637–639, 637f, 643f
Cardiac cycle, 632–633, 632f
Cardiac disease, 633–634, 649b
Cardiac troponin T, immunostrip for, 277
Cardiolipins, 399
Cardiorespiratory failure, in DMD, 971
Cardiovascular disease, 629–650
acute coronary syndrome. *see* Acute coronary syndrome
acute myocardial infarction. *see* Acute myocardial infarction
chemistry test for, 630–631
chronic kidney disease and, 662
congestive heart failure, 630, 634–635
guidelines for, 407f
markers for. *see* Cardiac biomarkers
Carnitine, 891
β-Carotene HPLC, reference intervals for, 989t–1025t

Carrier ampholytes, 178
Carrier gas, 186–187, 187f
Carryover, 14, 262–263
Cartridge type devices, for point-of-care testing, 278–279, 279f
Catalysts
definition of, 221
enzymes as, 221, 221b
Catecholamines, 335t, 450–465. *see also specific catecholamines*
biochemical phenotypes, 459f
biosynthesis of, 452–453, 452f
central nervous system and, 454, 455f
chemistry of, 452, 452f
clinical applications of, 457
diet and, 462–463
drugs and, 462–463
metabolism of, 453–454, 453f
metabolites, 457
physiology of, 454–464, 454f
preanalytical and analytical considerations for, 461–462, 462f
reference intervals for, 464, 989t–1025t
release of, 453
samples for, 462
storage of, 453
structure of, 452f
uptake of, 453–454, 453f–454f
Catechol-*O*-methyltransferase (COMT), 453–454
Categorical variables, 13
Cathinone, 576t
Cation-exchange chromatography, 192
Cationic salts, 239
Cationic trypsinogen gene *(PRSS1)*, 736–737
Cations
calcium. *see* Calcium
magnesium. *see* Magnesium
phosphate. *see* Phosphate
Cation-selective ion-selective electrodes, for whole blood, plasma, and serum measurements, 148t
cDNA, 939
Cecum, 724
Celiac disease, 730–731, 731t
Cell culture methods, 965
Cell-free fetal DNA (cffDNA)
analytical aspects, 985–986
maternal
collection and processing, 985
plasma, 982
β-Cells, 611
loss of, 612
Cellular hypoxia, agents causing, 567–568
Cellular iron regulation, 509, 510f
Cellulose, 380
Cellulose acetate, for electrophoresis, 172
Central dogma, of molecular biology, 919–921
Central hyperthyroidism, 827–828
Central hypothyroidism, 825
Central limit effect, for distribution of errors, 12
Central location, 11
Central nervous system, copper and, 490
Centrifugal analyzers, 261
Centrifugation, 114–116
principles of, 115
Centrifuges, 115
automated, 257
Centromere, 916
Cepheid GeneXpert system, 281
Ceramide, 399
Cerebrospinal fluid, proteins in, 309–310
specific, assessment of, 309–310
total, 309, 310t
Cerebyx. *see* Fosphenytoin
Certified reference materials (CRMs), 27–28, 111
Ceruloplasmin, 301, 490
reference intervals for, 989t–1025t
synthesis, by liver, 704
Cervical cancer, tumor markers for, 345
management of, 345
^{14}C-glycocholate breath test, for bacterial overgrowth, 733
Channel electron multiplier, 211
Charge-coupled device, 278
Charles' law, 430, 430t
Chelation therapy, 597
Chemical fume hood, 119
Chemical hazards, 123–124
Chemical hygiene plan (CHP), 118–119
Chemical interference, in hemolysis, 42
Chemical ionization
atmospheric pressure, 205–206, 205f
description of, 204–205
negative ion electron capture, 205
Chemical sensors, 145–166
Chemicals, 109–111
Chemiluminescence, 141–142
immunoassay, 246–247, 246f
limitations of, 142
Chemosensors, 276
Child-Pugh System, for classifying severity of cirrhosis, 718t
Chiral separations, 193–194, 193f
Chiral stationary phase (CSP), 194
Chlamydia trachomatis, 964
Chloramphenicol, 552t, 553
Chlordiazepoxide, 578, 578t
Chloride, 424, 688, 688b
in body fluids, 424
coulometric-amperometric titration of, 425–426
critical risk value, 1026t–1027t
hyperchloremia, 688
hypochloremia, 688
ion-selective electrode methods, 424
reference intervals for, 424, 989t–1025t
specimens, 424
sweat, 424–426
Chloromycetin. *see* Chloramphenicol
1,25-$(OH)_2$ cholecalciferol, 442t–444t
Cholecystitis, acute, 325
Cholecystokinin, 739t
Cholestasis, 719
of pregnancy, 870
Cholestatic liver disease, 719–720
Cholesterol, 392–393
absorption of, 392–393
assays, 413–414
catabolism, 393
emulsification of, 392–393
esterification, 393
high-density lipoprotein. *see* High-density lipoprotein cholesterol
low-density lipoprotein. *see* Low-density lipoprotein cholesterol
maternal, 864
reference intervals for, 989t–1025t
structure of, 393f
synthesis of, 393, 393f–395f
Cholesterol ester transfer protein, 403–404
Cholesteryl esters, 393
Cholinergic toxidrome, 563
agents related to, 574
Cholinesterase, 328–329, 328f
biochemistry of, 328
clinical significance of, 328
liver function testing, 328
methods of analysis for, 328–329
reference intervals for, 989t–1025t
Choral hydrate, 584
Choriocarcinoma, tumor markers for, 349
Chorionic gonadotropin. *see also* Human chorionic gonadotropin
analytical methods for, 876–877
assays for, 876
biochemistry of, 877
concentration of, 868f
description of, 868f, 876
in ectopic pregnancy, 869
point-of-care testing for, 877
stimulation test, 853
Chorionic villus sampling, 867–868
description of, 86
specimen collection for, 86
Chromaffin cells, 452–453, 456
Chromaffin granules, 456
Chromatin, 919
Chromatogram, 182, 183f–184f
Chromatograph, 182
Chromatography, 180–200, 233
adsorption, 191, 191f
affinity, 193–194
analyte identification and quantification in, 185–186, 186f
anion-exchange, 192
band-broadening and efficiency in, 184
basic principles of, 182–186
bonded phase gas, 187
cation-exchange, 192
column. *see* Column chromatography

Chromatography *(Continued)*
definition of, 182
gas. *see* Gas chromatography
gas-liquid, 182, 187, 187f, 187t
gas-solid, 182, 186–187
gel filtration, 193
gel permeation, 193
general terms and components of, 182–183, 183f–184f
immobilized metal-ion affinity, 194
immunoaffinity, 194
ion-exchange, 192
ion-pair, 194
liquid. *see* Liquid chromatography
normal-phase, 192
paper, 198
partition, 191–192
planar, 182, 198–199
resolution of, 184–185, 185f, 185b
equation, 185
retention and selectivity in, 183–184
reversed-phase, 192
size-exclusion, 192–193, 193f
thin-layer, 198
total cortisol, 805
Chromium (Cr), 495, 595–596
reference intervals for, 989t–1025t
Chromogranin A, 335t, 739
Chromogranins, 453
Chromosome, 916
naming of, 927
Chronic diarrhea, 740–741, 741b
Chronic hepatitis, 712–717, 713t
Chronic hepatitis B, 713–714, 714t
Chronic hepatitis C, 715
Chronic kidney disease (CKD), 660
cardiovascular complications of, 662
classification of, 660–663, 660b–661b, 661t
disorders of bone and mineral in, 770
dyslipidemia in, 662–663
glomerular filtration rate in, 659
management of, 661
Chronic liver disease, 717b
Chronic obstructive pulmonary disease (COPD), 696
Chronic pancreatitis, 737
Chronic viral hepatitis, 709
Chylomicrons, 392–393, 398–399
Circadian rhythm, 444–445
Circulating cell-free fetal DNA analysis, 985
Circulating cell-free fetal nucleic acids, in maternal plasma, biology of, 982–985, 983f
clinical implementation, 985
Circulating nucleic acids, for prenatal diagnostics, 982
Circulating tumor cells, 977–979
analytical techniques, 978
and breast cancer, 978–979
ctDNA *versus*, 981–982, 982t
Circulating tumor DNA (ctDNA), 977–982
circulating tumor cells *versus*, 981–982, 982t
ctDNA integrity, 980–981
methylated ctDNA, 980
multimarker ctDNA assays, 981
tumor-specific ctDNA mutations, 979–980, 981f
types of, 979, 980f
viral ctDNA, 980–981
Cirrhosis, 717–719
cause and treatment of, 718t
Child-Pugh System for classifying severity of, 718t
tests of hepatic function and injury in, 718t
Citric acid, reference intervals for, 989t–1025t
Citrullinemia type II, 890
CK isoenzymes, reference intervals for, 989t–1025t
Class S weights, 116–117
Clearance, of kidneys, 373
Clinical and Laboratory Standards Institute (CLSI), 95, 109
reagent water specifications, 109t
Clinical chemist, functions of, 4b
Clinical chemistry, 3–4
definition of, 2
ethical issues in, 4b
molecular diagnostics and, 1–6, 6b
Clinical decision limits, 68
Clinical electrophoresis, 169–173
Clinical enzymology, 221
Clinical evaluation, of methods, 7–37, 35b
Clinical laboratory
accreditation of, 87
College of American Pathologists accreditation of, 87
function of, 2
plans for, 119–120
testing in, 2, 2b
Clinical Laboratory Improvement Amendments of 1988 (CLIA), 285
Clinical sensitivity, 73–75, 74f
Clinical specificity, 73–75, 74f
Clinical toxicology, 562–589
analytical considerations of, 564
basic information of, 564–565
clinical considerations of, 564
clinical evaluation of, 564
Clinicians, specifications derived from opinions of, 60
Clomipramine, 572f
Clonazepam, 551t, 578t
Cloned enzyme donor immunoassay (CEDIA), 244f, 245
Cloning, 956–957, 957f
Clopidogrel, application of, 907
Clorazepate, 578t
Closed-system analyzers, 263–264
Coagulation proteins, 704
Cobalt, 495–497
Cobalt (Co), 596–597
reference intervals for, 989t–1025t
Cobamide, 479–480
Cocaine, 579–580
metabolism of, 580f
Codeine, 581–583, 582f
application of, 905, 906f
Codes of conduct, 4
Codon, 916
Coefficient of variation (CV), 11
Coenzymes, 222, 229, 317
Cold agglutinins, 48
Colipase, 323
Collagen crosslinks, 767–768
Colligative properties, 427
Colloid, 812
Colon, 724
Colonization, of upper small bowel, 732
Colorectal cancer, tumor markers for, 345
Column chromatography, 182
gas chromatography. *see* Gas chromatography
in vitamin D metabolites, 762
Commercial coulometric titrators, 158
Committee on Publication Ethics, 5
Commutability, 26–27
Commutable samples, external quality assessment and, 101–102
Comparative genomic hybridization, 950–951, 951f
Compensation, creatinine and, 364
Competitive immunoassays, 241, 241f
Competitive inhibition, 227
Complement system, 302–303, 302f
inherited deficiencies of, 303t
Complete blood count (CBC), 503, 505
parameters that constitute, 506t
Compressed gases, hazards from, 124
Concentration quenching, in fluorescence measurements, 140
Conception, 863–864
Conductance, 157–158
Conductivity detector, in liquid chromatography, 197, 197t
Conductometric detection, enzyme-based biosensors with, 162, 162f
Conductometry, 157–158, 157b
Confidence intervals, 68
Confidentiality
of genetic information, 4
of patient medical information, 4
Confirmatory testing, 563
ConfirmMDx, 334t
Conflicts of interest, 5
Congenital adrenal hyperplasia, 801–802, 846–847
Congenital erythropoietic porphyria (CEP), 530t, 532
Congenital hypothyroidism, 831–832

Congestive heart failure
B-type natriuretic peptide. *see* B-type natriuretic peptide
description of, 630, 634–635
Conjugated bilirubin, 521
Conjugated hyperbilirubinemias, 520b, 521
Conn syndrome, 802
Connective tissue formation, copper and, 490
Constant neutral loss, 210
Continuous dynode electron multipliers, 211
Continuous-flow analyzers, 262
Contraction alkalosis, 695–696
Control charts
description of, 95, 95f
Levey-Jennings, 92, 92f, 99f
Control materials, 94
Conveyor belts, 257–259, 258f
Cooperativity, 502
Copper, 490–491
absorption of, 490
deficiency in, 490–491
description of, 593
excretion of, 490
functions of, 490
laboratory assessment for, 491
metabolism of, 490
reference intervals for, 989t–1025t
toxicity of, 491
transport of, 490
Coproporphyrin, 529
Coproporphyrinogen oxidase (CPOX), 528
Copy number variant (CNV), 926–927
Cori cycle, 382–383
Coronary arteries, 631–632
Coronary heart disease, 405–406
in children, 413f
prevention of, 410b
Corpus luteum, 839
Correlation coefficient, 23, 24f
Corrin, 479–480
Corticotropin hormone, 442t–444t
Corticotropin-releasing hormone, 442t–444t
description of, 776
stimulation test for, 793
Cortisol, 442t–444t, 609
in adrenal insufficiency, 793
blood, 806
deficiency of, 801
free
measurement of, 806
reference intervals for, 989t–1025t
measurement of, 805
metabolism, 789
release of, 790–791
synthesis, 791f
total, 805
reference intervals for, 989t–1025t
Cosyntropin stimulation test, 793b
Cosyntropin test, 793
Coulometric-amperometric titration, 425–426
Coulometry, 158
Coulter principle, 158, 268
Countercurrent multiplication mechanism, 657f, 658
Counterregulatory hormones
cortisol, 609
description of, 608–609
epinephrine, 608
glucagon, 608
growth hormone, 608
C-peptide, 607
measurement of, 610
C-reactive protein (CRP), 302
cardiac marker uses of, 648
Creatinase, 364–365, 365f
determination of, 368t
Creatine, synthesis of, 294
Creatine kinase (CK), 314–317, 314f, 634, 1026t–1027t
activity, 316
assays for, 316
biochemistry of, 315, 315t
clinical significance of, 315–316
elevation of, 315–316
isoenzymes, 316–317
laboratory tests of, 315
neuromuscular disorders and, 315
quantification of, 316–317
separation of, 316–317
subunits of, 315
Creatininase, 364, 365f
creatinase and, 364–365
Creatinine, 359–376, 367b
analytical methods for, 363–366
quality issues and preanalytical considerations with, 366
biochemistry and physiology of, 363, 363f
clearance of, 374
clinical significance of, 366
critical risk value, 1026t–1027t
cystatin C, 367
glomerular filtration rate and, 373–374, 374f
measurement
enzymatic, reference intervals for, 989t–1025t
Jaffe reaction, 989t–1025t
metabolism of, 374
reference intervals for, 366
sample collection for, 363
traceability of, 366
urinary excretion of, 366
Creatinine clearance, 989t–1025t
Creatinine deaminase, 365
o-Cresolphthalein complexone, 748
Cretinism, 817
Critical care analyzers, 279, 280f
features of, 280b
Critical difference, definition of, 52, 57
Crohn disease, 732
CRP. *see* C-reactive protein
Cryoglobulins, 48
Cryptorchidism, 839
C-telopeptide assay, 989t–1025t
C-telopeptide CTX assay, 768
C-telopeptide ICTP assay, 768
C-terminal residues, 296
C-type natriuretic peptide, 296
Cushing syndrome
biochemical evidence of, 800f
causes of, 795
classification of, 798f
clinical manifestations of, 795
differentiation of, 800b
Cutaneous porphyrias, 532, 533b
patients with, 534, 534f
Cuvets
in fluorescence measurements, 141
in spectrophotometers, 133
Cuvettes, 264
Cyanide, 568, 989t–1025t
Cyanmethemoglobin, 431
Cyclic adenosine monophosphate (cAMP), 446
Cyclobenzaprine, 572f
Cyclosporine, 555–557
CYFRA 21-1, 335t
CYP2D6, 581–582
Cystatin C, 367, 367b, 660, 989t–1025t
analytical methods for, 367
biochemistry and physiology of, 367
clinical significance of, 367
reference intervals, 367
representative glomerular filtration rate estimating equation, 368t
Cysteine, 290t–291t, 292
Cysteinyl proteases, 299
Cystic fibrosis (CF), 298, 426, 736, 911, 968–969
CFTR gene mutation testing, 969
screening in, 969
Cystic fibrosis transmembrane conductance regulator protein, 426, 968, 968b
Cystine, 989t–1025t
Cytochrome P450 (CYP), 898, 903
CYP2D6, 903–905, 904t
Cytochrome P450 2C family, 905–907
Cytochrome P450 3A family, 907
Cytogenetics, 974
Cytometry, 139

D

Dalton's law, 430, 430t
Data
generation of, on components of biological variation, 52–54
data analysis, 54
design of studies, 52
methods for analysis, 53–54
interpretation and use of, 56–57
consequences for population-based reference intervals, 56–57, 57t
individuality, 53f, 56
Data analysis, 54

Databases, 55–56, 55f
generation of, from patient populations, 55–56
merits and disadvantages of, 55
Deaminated gliadin peptide (DGP) antibodies, in celiac disease, 731
Decapper station, 257
Decontamination, 591
Deep venous thrombosis
accuracy of D-dimer test in diagnosis of, 29–30, 29f
added value of D-dimer testing in diagnosis of, 32, 34t
Defective glucose counterregulation, 382
Dehydroepiandrostenedione sulfate (DHEAS), 442t–444t
Dehydroepiandrosterone (DHEA), 442t–444t, 789
chemical structure of, 836f
description of, 835
measurement of, 858
unconjugated, 989t–1025t
Dehydroepiandrosterone sulfate, 989t–1025t
chemical structure of, 836f
description of, 835
measurement of, 858
Deionized water, 109
Deletion, 924–925
Delta-/beta-thalassemia, 504
Deming regression analysis, 20–21
computation procedures for, 21–22
weighted, 25–26, 25f, 26t
Denaturation, of proteins, 217, 297
Densitometry, 170, 170t
11-Deoxycortisol, 807
Deoxypyridinoline, 768
Deoxyribonucleic acid (DNA), 916
analysis, 507, 538
affinity sensors for, 164
intergenic, 924
methylation, 922–923
modification enzymes, 920
repair, 920
replication, 919–920, 919f
structure of, 916f, 917
transcription, 917
translation, 917, 921, 921b, 922f
Deoxyribonucleoside triphosphates (dNTPs), 938–939
Depakene. *see* Valproic acid
Depakote. *see* Valproic acid
Depletional hyponatremia, 683
Deproteinization, in vitamin D metabolites, 762
Derivatization, 212
Desipramine, 572f
Detection of drugs, screening procedures for, 565–567
Device integration, 260
Dexamethasone suppression test, 796–798, 798b–799b
Dextromethorphan (DXM), 584, 586
Diabetes
adult-onset, 605
categories of increased risk for, 605–606
diagnosis of, 627b
hemoglobin A_{1c} in, 621
juvenile-onset, 605
monitoring of, hemoglobin A_{1c} in, 621
specific types of, 605
type 1, 615–616
type 2, 616, 626b
Diabetes Control and Complications Trial (DCCT), 615
Diabetes insipidus (DI), 670, 684
causes of, 783t
hypothalamic, 782
nephrogenic, 782–783
Diabetes mellitus, 378, 603–628
chronic complications of, 615–616
classification of, 605, 605b
definition of, 604
diagnosis of, 613–614, 613b
impaired fasting tolerance, 605–606
impaired glucose tolerance, 605–606
oral glucose tolerance test, 613–614, 614b
environmental factors associated with, 612
gestational, 604–605
diagnosis of, 614–615
insulin-dependent, 605
non-insulin-dependent, 605
prevalence of, 604–605
role of clinical laboratory in, 616–617
role of laboratory in management of, 616t
type 1 (insulin-dependent), 605
genetics of, 611–612
pathogenesis of, 611–612
prenatal screening tests affected by, 876
type 2, 605
complications of, 616
environmental factors, 612–613
insulin resistance as cause of, 612
obesity and, 605
pathogenesis of, 612–613
Diabetes susceptibility genes, type 2, 613
Diabetic ketoacidosis, 617, 692
Diabetic nephropathy, 666
Diacylphosphoglyceride, 399
Diagnostic accuracy, 337
Diagnostic interpretation, 932, 933f
Diagnostic tests. *see also specific test*
impact of, 34–35, 35f
schematic representation of, 274f
Dialysis, 671–674, 672t
malnutrition and, 674
Diarrhea, 725
acidosis caused by, 694
chronic, 740–741, 741b
Diastole, 632–633
Diazepam, 578t, 585f
Diazo methods, for bilirubin, 521, 521f
Dideoxy-termination sequencing (Sanger sequencing), 939, 945–946, 945f–947f
Difference plot, 19, 19f
Differential solubility, 297
Differentiated thyroid cancer, 830
Diffraction gratings, 132
Digestion, 724
Digital hardware, in spectrophotometers, 133
Digital image analyzers, 170
Digital immunoassay, 248, 249f
Digital polymerase chain reaction (digital PCR or dPCR), 942
Digoxin, 558–559
Dihydropteridine reductase, 888–889, 888f
Dihydrotestosterone, 834–835, 836f, 989t–1025t
3,4-dihydroxyphenylalanine (L-dopa), 452–453
3,4-dihydroxyphenylglycol (DHPG), 453–454
1,25-dihydroxyvitamin D, 754t, 761
measurement of, 762
24,25 dihydroxyvitamin D, measurement of, 762
Dilantin. *see* Phenytoin
Dilutional hyponatremia, 683
Dilutions, 15, 117, 265
Dipeptide, 295f
Direct bilirubin, 521–522
Direct equilibrium dialysis, 821
Direct optical methods, for total protein measurement, 306
Disaccharidase deficiencies, 731–732, 732t, 732b
Disaccharides, 379–380, 379f
Discrete dynode electron multiplier, 211
Dispensers, 112–113, 113f–114f
Dispensing, 111–114
Dissociation interferences, 135
Dissolved carbon dioxide, 688
Distillation, 110
Distilled water, 109
Distribution
bimodal, 67
frequency, 9–10, 10f
Gaussian probability, 12, 12f, 67
inspection of, 67
polymodal, 67
probability, 10–11, 11f
student *t* probability, 12–13, 12f
Diuretics, 670
Dixon-Reed range test, 67
DMD gene (Xp21.2), 971
DNA microarray, 946
Docking station, 276
Dopamine, 451
reference intervals for, 989t–1025t
structure of, 452f
Dose-effect relationship, 488–489, 489f

Dose-response relationship, 548
Dot blotting, 240
Double-beam spectrophotometer, 131, 131f
Double-pan balance, 116
Down syndrome
  definition of, 874
  description of, 872
  prenatal screening for, description of, 874, 876f, 876b
Doxepin, 572f
2,3-DPG, 432
Dried blood spot, 884
Drift, 14
Dronabinol (Marinol), 579
Drug analysis, 567
Drug half-life, 546
Drug interactions, 548–549
Drug metabolism, 898–899, 905, 905b
Drug testing
  overview of, 574
  propoxyphene, 582f
  specimens, 574
Drug-facilitated crimes, 584–586
Drug-induced cholestasis, 720
Drug-induced liver injury (DILI), 717
Drug-metabolizing enzymes, 903–909, 903f
Drugs of abuse
  amphetamine. *see* Amphetamine
  barbiturates. *see* Barbiturates
  benzodiazepines. *see* Benzodiazepines
  cannabinoids. *see* Cannabinoids
  cocaine, 579–580
  description of, 574–580
  detection of, 586–588
  dextromethorphan, 584, 586
  ephedrine, 575–576
  ethanol. *see* Ethanol
  hair analysis to detect, 587–588
  ketamine, 583–584
  lysergic acid diethylamide, 563, 566–567, 581
  meconium analysis to detect, 586–587, 587t
  methylphenidate, 576–577
  opioids/opiates. *see* Opioids/opiates
  phencyclidine, 583–584
  pseudoephedrine, 575–576
  saliva analysis to detect, 587
  sweat analysis to detect, 588
  testing. *see* Drug testing
Drug's site of action, 542
Dry chemistry systems, 365
Dubin-Johnson syndrome, 520b
Duchenne muscular dystrophy, 971–972, 972f
Duodenum, 724
Dye-binding, for total protein measurement, 306
Dyes and probe, 952–953, 952b, 954f
Dysbetalipoproteinemia, 409
Dyshemoglobins, 431
Dyslipidemia
  in chronic kidney disease, 662–663
  treatment of, 414f

E

Ebola virus, 120
Eclampsia, 869
Ectopic pregnancy, 869
Editor-in-chief, responsibility of, 5
Editors, responsibility of, 5
EDTA, 83
Edwards syndrome. *see* Trisomy 18
Electrical charge, 297
Electrical hazards, 124
Electroblotting, 239
Electrocardiogram (ECG), 565
  acute myocardial infarction, 633f
Electrochemical cell, 146
Electrochemical detection, 943
Electrochemical detector, in liquid chromatography, 197, 197t
Electrochemical labels, affinity sensors for DNA analysis using, 164–165, 164f
Electrochemical sensors, for whole blood measurements, 153b
Electrochemiluminescence, 142
Electrochemistry, 145–166
Electrode potential, for redox couple, 147
Electrodes, 146
  glass, 148–149
  inert metal, 147
  ion-selective, 146, 148–153, 148t
  metal, 147–148
  for partial pressure of carbon dioxide, 151–153, 151f–152f
  polymer membrane, 149–151, 149f–150f
  redox, 146–147
  types of, 146–148
Electroendosmotic flow (EOF), 176
Electrolyte(s), 420, 427b, 680, 682
  chloride. *see* Chloride
  classification of, 420
  composition of body fluids compartments, 681–682, 681t
  homeostasis, 656–657, 657f
  magnesium. *see* Magnesium
  physiology and disorders of, 679–698
  potassium. *see* Potassium
  sodium. *see* Sodium
  specimens for, 420–421
  water homeostasis by, 420
Electrolyte exclusion effect, 423–424, 424f
Electromotive force (EMF), 146
Electron capture detector (ECD), 190
Electron ionization, 203, 205f
Electron multiplier, 211
Electronic balance, 116
Electropherogram, 168
Electrophoresis, 167–179, 168b, 233, 506, 506f, 939, 944–948, 944f, 944t
  agarose in, 172
  alternatives, 948, 948f
Electrophoresis *(Continued)*
  automated systems in, 173
  basic concepts of, 168
  blotting techniques for, 171
  buffers in, 171–172
  capillary, 168, 173–174
    direct detection for, 174
    general operations for, 173–174, 173f
    indirect detection of, 174
    sample injection for, 174
    surface effects in, 176–177
    technical considerations for, 175–177
    temperature effects in, 175–176
  capillary gel, 174–175
  capillary zone, 173b, 174
  cellulose acetate in, 172
  clinical, 169–173
  definition of, 168
  detection and quantification of, 169–170
  distorted, unusual, or atypical bands in, 175
  gel in, 175
  general operations for, 169–171
  high-resolution, 172
  immunofixation, 307, 309f
  instrumentation of, 171–173, 171f
  isoelectric focusing, 177–178, 177f
  micellar electrokinetic chromatography in, 178
  motility of ions in, 169b
  polyacrylamide gel in, 172–173
  power supplies in, 171
  protein analysis using, 307
  sample application, discontinuities in, 175
  sampling in, 175
  separation in, 169
  serum protein, 307, 307f–308f
  slab gel, 169
  staining in, 170t
  support media for, 172–173
  technical considerations for, 175–177
  techniques for, 177–178
  theory of, 168–169, 169f
  types of, 174–175
  unequal migration rates in, 175
  zone, 168
Electrophoretic mobility, 168
Electrospray ionization, 205, 206f
Electrospray mass spectroscopy, 507
Elemental Hg, 598
Elements, 590
ELISA. *see* Enzyme-linked immunosorbent assay
Embryo, 863–864
Emission, in optical methods, 129t
Emulsification, 392–393
Emulsion polymerase chain reaction, 957, 958f
Enantiomer, 563, 586
Encephalocele, 873
Endocardium, 631–632

Endocrine system, 441
Endogenous pathway, of metabolism of lipoproteins, 403–404, 404f
Endometrium, 863
Endoscopy, for *Helicobacter pylori*, 728
Endosmosis, 169
Endothelins, 296
Endpoint quantification, in amplification assays, 943
End-stage renal disease (ESRD), 660
  incidence rate of, 662f
Enteric nervous system, 457
Enzymatic assay, 622
Enzymatic methods, for creatinine, 364
Enzyme(s). *see also specific enzyme*
  activation of, 228–229
  active center of, 217
  aggregation of, 219
  allelic variants, 218
  as analytical reagent, 232–233
  basic concepts of, 314, 314b
  as catalysts, 221, 221b
  concentrations of, 222
  definition of, 216
  distribution of, 315t
  Enzyme Commission numbers for, 216t
  forms of
    nongenetic origins of, 218–220, 219f
    properties of, 220–221
  of heme biosynthesis, 527–528, 527t
  immobilized, analytical applications of, 233
  International Union of Biochemistry nomenclature system for, 216
  isoenzymes. *see* Isoenzymes
  isoforms, 218
  kinetics of, 221–222
  mass concentration measurements, 231–232
  measurements of, 538
  muscle, 314–317
  nomenclature of, 216
  oligomeric, 218
  origin of, 219f
  as proteins, 217
  rate analyses and, 215–235
  serum, 313–330
  specificity of, 217
  units for expressing activity of, 221
  variants
    genetic origins of, 218, 219f
    nongenetic origins of, 218–220, 219f
Enzyme immunoassays, 244–245
  description of, 233, 244
  detection systems for, 244
Enzyme induction, 545
Enzyme reactions
  activators of, 226–229
  antibody inhibition of, 228
  competitive inhibitor of, 227
  consecutive, 225
  continuous-monitoring measurement of, 229
Enzyme reactions *(Continued)*
  fixed-time measurement of, 229
  indicator, 222
  inhibitors effect on, 226–229
  optimization of, 231
  pH effects on, 225–226, 225f
  rate of
    factors that affect, 222–229, 229b
    measurement of, 229–231, 230f
  single-substrate, 222–224, 223f
  standardization of, 231, 232f
  substrate measurements, 231
  temperature effects on, 226, 226f
  two-substrate, 224–225
Enzyme-based biosensors, 159
  with amperometric detection, 160–162, 160f–161f
  with optical detection, 162–163
  with potentiometric and conductometric detection, 162, 162f
Enzyme-linked immunosorbent assay (ELISA), 244
  for HCV genome, 710
Enzyme-multiplied immunoassay technique (EMIT), 244–245, 244f
Enzyme-substrate complex, 221–222, 222f
Ephedrine, 575–576
Epicardium, 631–632
Epigenetics, 922–923, 923f
Epinephrine, 380–381, 442t–444t, 451
  in glucose counter-regulation, 608
  reference intervals for, 989t–1025t
  structure of, 452f
Epitope, 237
Epitope locations, on BNP and NT-proBNP, 643f
Epoc system, 278, 278f
Equilibrium, 240
  dialysis/ultrafiltration, 857
Equilibrium methods, 232
Erectile dysfunction, 839
Ergonomics, 118
  program, 120
Error model, 17–18, 18f
Eruptive xanthomas, 407–408
Erythroblastosis fetalis, 870
Erythrocytes, 249
  in conductometry, 157
Erythroid dysmaturation, 511t–512t
Erythroid maturation, disorders of, 516
Erythroid progenitor cells, defect in iron acquisition of, 511t–512t
Erythropoiesis, low iron availability for, 511t–512t
Erythropoietic protoporphyria (EPP), 530t, 532–533, 535
Erythropoietin (EPO), 442t–444t, 658
Eskalith. *see* Lithium
Essential amino acids, 292
Estazolam, 578t
Estimated regression line, evaluation of random error around, 22
Estradiol, 989t–1025t
  17β-, 840
  in blood, measurement of, 858
  chemical structure of, 840f
  metabolism of, 842f
Estrane, 840f
Estriol, 840, 840f
  characteristics of, 864
  free, 989t–1025t
  total, 989t–1025t
  unconjugated, 873f, 878
Estrogen, 442t–444t
  biosynthesis of, 840, 841f
  in blood
    measurement of, 858
    transport, 840
  calcium homeostasis affected by, 840
  chemistry of, 840, 840f
  functions of, 840–842
  metabolism of, 840–842, 842f
  during pregnancy, 867
  in pregnancy, 840
Estrogen receptor, 336t
Estrone, 840, 840f, 858, 989t–1025t
Eszopiclone, 585f
Ethanol, 989t–1025t
  analysis of, 570
  blood, 570
  breath, 570
  description of, 568–570
  metabolism of, 569f
  urine, 570
Ethics, definition of, 1
Ethosuximide, 549, 551t
Ethyl glucuronide, 571
Ethylene glycol, 563, 570, 694
Ethylenediaminetetraacetic acid. *see* EDTA
Ethylsulfate, 571
Eukaryote, 919
European Diagnostic Manufacturers Association, 5
Euthyroid, 814
Evaporation, 117
Event-based quality control sample measurement, 95
Everolimus, 557–558
Excitation interference, 135
Excretion, of drugs, 545
Exogenous base, 696
Exogenous pathway, of metabolism of lipoproteins, 403, 403f
Exome sequencing, 901–902
Exon, 916, 919–920
Exonuclease, 919
Expected date of confinement, 863
Exposure control plan, 119–120
Extension cords, in electrical hazards, 124
External quality assessment, 91, 101–106, 104b
  commutable samples, 101–102
  interpretation of, 102–106, 103f–105f, 106b

External quality assessment *(Continued)*
measurement procedure, 102
noncommutable samples, 102
External quality assurance (EQA), 284
External quality control, 91
Extracellular fluid (ECF), 681t
Extracted ion chromatogram, 202
Extraction, in vitamin D metabolites, 762
Extraction recovery, 14
Extrahepatic biliary atresia, 521

F

Facilitative glucose transporters (GLUTs), 607
Fallopian tubes, 839
False negative, 29f
False positive, 29f
False rejection, probability for, 96
Familial combined hyperlipidemia, 408
Familial hypercholesterolemia, 409
Familial hypertriglyceridemia, 408
Fanconi syndrome, 656
Faraday cup, 211
Fasting hypoglycemia, 609
Fasting plasma glucose, concentrations of, 613
FASTQ file, 929f
Fats, 726, 727f
absorption, 740
Fatty acids, 392, 394–396
catabolism of, 395, 397f
classification of, 394–395
essential, 399–400
in human tissue, 397t
ketone formation, 396
monounsaturated, 394–395, 397f
β-oxidation, 395
oxidation disorders, 892, 892b, 893f–894f
polyunsaturated, 394–395, 397f
saturated, 395, 397f, 399
trans, 395
Fatty acylation, of proteins, 298
Fatty liver of pregnancy, 870
Fecal material, proteins in, 310–311
Feces
porphyrins and porphyrin precursors in, 535–536
specimen of, 85
Felbamate, 550, 551t
Felbatol. *see* Felbamate
Female athletic triad, 851
Female infertility, 853b–854b, 854–856, 854f
Female pseudohermaphroditism, 846–847
Female reproductive system
abnormalities of
amenorrhea. *see* Amenorrhea
female pseudohermaphroditism, 846–847
hirsutism, 846, 850–851, 850b
precocious puberty, 847
anatomy of, 839
Female reproductive system *(Continued)*
biological functions of, 839–851, 851b
development of, 843–851
estrogens. *see* Estrogen
hypothalamic-pituitary-gonadal axis in, 839
menopause, 846, 846f
menstrual cycle. *see* Menstrual cycle
progesterone, 842–843
Fentanyl, 582f
Ferricyanide, 363
Ferritin, 509, 989t–1025t
iron deficiency and, 663
methods for, 517
Ferrochelatase (FECH), 528
Fetal alcohol spectrum disorders (FASD), 569
Fetal chromosomal aneuploidy screening, 983–985
Fetal DNA fraction, measurement, 985
Fetal fibronectin, 879
Fetal fraction, 982–983
Fetal hemoglobin, 568, 867
globin structure of, 501
hereditary persistence of, 504
Fetal lung maturity (FLM), 865
α-Fetoprotein, 334t, 339t–340t, 989t–1025t
amniotic fluid, 873, 877
biochemistry of, 877
chemistry of, 877
hepatocellular carcinoma and, 350
diagnosis of, 350
posttreatment monitoring, 350
prognosis for, 350
screening for, 350
maternal screening for fetal defects, 873
neural tube defect screening with, 873
during pregnancy, 877–878
specimen, 877
synthesis, by liver, 704
Fetus
assessments of, 868–869
blood development in, 867–868
definition of, 863–864
disorders of
Down syndrome. *see* Down syndrome
maternal serum screening for, 874–876, 875t
neural tube defects. *see* Neural tube defects
trisomy 18, 874
female reproductive development in, 843
growth and development of, 867–868
lung development in
description of, 867
tests for assessing, 880
preterm delivery of, 871
pulmonary surfactant in, 867
Fiber optics, in spectrophotometers, 132–133
Fibroblast growth factor 23 (FGF23), 753, 763, 763b
Fibrosis, biomarkers of, 718
Filter paper, blood specimen collection on, 82
Filter photometers, 130–131
Filters, 117–118
in spectrophotometers, 132
Filtration, 117–118
Fire extinguishers, 124–125, 125t
Fire hazards, 124–125, 125t
First-order kinetics, 233
First-pass effect, 544
FK506. *see* Tacrolimus
Flame ionization detector (FID), 188–189, 189f
Flame photometry, 423
Flameless AA techniques, 134
Flatbed scanners, 170
Flow control, in gas chromatography, 187, 187f
Flow cytometer, 139–140, 140f
Flow cytometry, 139, 268, 946
Fluidics, 276
Flunitrazepam, 578t
Fluorescence, 135, 943
labels, 943
measurements, limitations of, 140–141
Fluorescence decay process, 136f
Fluorescence detector, in liquid chromatography, 197, 197t
Fluorescence emission, time relationships of, 135–136, 136f
Fluorescence in situ hybridization (FISH), 950–951, 974
Fluorescence intensity, concentration and, relationship of, 136
Fluorescence polarization, 136–137
Fluorescence polarization immunoassay (FPIA), 245
Fluorescent labels, affinity sensors for DNA analysis using, 164
Fluoride, 989t–1025t
Fluoroimmunoassay, 245, 245f, 245t
Fluorometers, 138–140
Fluorometry, 135–141
components of, 137, 137f–138f
concepts of, 135–137, 135f
fluorescence emission, time relationships of, 135–136, 136f
fluorescence intensity, concentration and, relationship of, 136
fluorescence polarization in, 136–137
instrumentation of, 137–140
performance verification in, 137
Fluorophores, 135, 245
Flurazepam, 578t
Folic acid/folate, 482–485, 989t–1025t
absorption of, 482–483
deficiency in, 294, 481f, 483–485, 873
excretion of, 482–483
functions of, 483, 484f
laboratory assessment for, 485
metabolism of, 482–483

Folic acid/folate *(Continued)*
structure of, 483f
toxicity of, 485
transport of, 482–483
Follicle, 839
Follicle-stimulating hormone, 442t–444t, 834, 845, 989t–1025t
biochemistry of, 780, 781b
characteristics of, 780
definition of, 774
physiological action of, 780
Follistatin, 442t–444t
Fosphenytoin, 548
Fractional oxyhemoglobin, 432
Fragility, patients with, 534
Fragment ions, 202
Free (ionized) calcium, 749–751
anticoagulants in, 751
calibrators and quality controls in, 751
interferences in, 750
interpretation of, 752
pH, 750, 751f
preanalytical errors in measurement of, 751
reference intervals for, 751–752, 989t–1025t
specimen requirements in, 750–751
Free cortisol
measurement of, 806
reference intervals for, 989t–1025t
Free (ionized) magnesium
measurement of, 755
reference intervals for, 756, 989t–1025t
specimen requirements for, 755–756
Freeze drying, 117
Freezing point, 427
Freezing point depression osmometer, 428, 428f
Frequency distribution, 9–10, 10f
Fructosamine, 623–624, 989t–1025t
determination of, 623–624
reference intervals for, 624
Full-term infants, reference intervals for, 989t–1025t
Fulminant hepatic failure, 705–706
Fulminant hepatitis, 710
Functional hypoglycemia, 382
Functioning adrenocortical tumors, 802
Furosemide stimulation test, 794
Fusion, 927
Fusion transcript, 961

G

Gabapentin, 551t
Galactorrhea, 839, 854
Galactosemia, 895
Gallstones, 720
Gamma-aminobutyric acid (GABA) transaminase, in valproic acid, 549
Gamma-butyrolactone, 585
Gamma-hydroxybutyrate (GHB), 563, 585
Gangliosides, 399
Gas chromatography (GC), 182, 186–190
based on the stationary phase, 187b
carrier gas sources and flow control in, 187, 187f
columns and supports in, 188
data acquisition and system control in, 190
detectors of, 188–190, 189t
drug screening uses of, 567
injection systems and sample derivatization in, 187–188
instrumentation of, 187–190
temperature control in, 188
total cortisol evaluations, 805
types of, 186–187
Gas chromatography-mass spectrometry (GC-MS), 190
applications of, 211–212
drug screening uses of, 211
limitations of, 212
Gases, exchange of, in lungs, 690–691
Gas-liquid chromatography (GLC), 182, 187, 187f, 187t
Gas-solid chromatography (GSC), 182, 186–187
Gastric acid secretion, 729–730
Gastric cancer, tumor markers for, 346
management of, 346–347
Gastric function, 723–743
Gastric hypochlorhydria, 732
Gastrin, 442t–444t, 724, 729–730, 739t
Gastrin-17, 739t
Gastrinomas, 729–730
Gastritis, 728
Gastroduodenoscopy, for *Helicobacter pylori*, 728
Gastroenteropancreatic neuroendocrine tumors (GEP-NET), 460–461, 461b
Gastrointestinal disorders
acute pancreatitis, 741–742, 742b
bacterial overgrowth, intestinal, 732–733
bile acid malabsorption, 733, 733b
celiac disease, 730–731, 731t
chronic diarrhea, 740–741, 741b
disaccharidase deficiencies, 731–732, 732t, 732b
gastrinomas, 729–730
inflammatory bowel disease and irritable bowel syndrome, 733–736
malabsorption. *see* Malabsorption
neuroendocrine tumors, 738–739
Gastrointestinal losses, acidosis and, 694
Gastrointestinal neuroendocrine tumors, 738–739
Gastrointestinal regulatory peptides, 738, 739t
Gastrointestinal stromal tumors, tumor markers for, 346–347
management of, 347
Gastrointestinal tract. *see also specific anatomy*
absorption
of carbohydrates, 725–726
of lipids, 726–727
Gastrointestinal tract *(Continued)*
processes of, 725
of proteins, 727–728
anatomy of, 723–725
large intestine, 724
pancreas, 724, 724f
small intestine, 724
stomach, 724, 724f
digestion
of carbohydrates, 725–726
gastric phases of, 725
intestinal phases of, 725
of lipids, 726–727
neurogenic phase of, 724
phases of, 724
processes of, 725
of proteins, 727–728
Gaussian probability distribution, 12, 12f, 69
Gel, in electrophoresis, 175
Gel filtration chromatography, 193
Gel permeation chromatography, 193
Gene
naming of, 927
structure of, 920, 921f
Gene duplication, 898
Gene expression, 930
copper and, 490
General gas equation, 430
Genetic code, 919
Genetics, 967–972
autosomal dominant diseases, 969–971
autosomal recessive disorders, 967–969
of hematopoietic malignancies, 974
X-linked diseases, 971–972
Genomes, 916
Genome-wide association studies (GWAS), 925
Genotype, 898, 902, 916
Genotyping, 898, 901
Gentamicin, 552t
Germ cell tumors, tumor markers for, 347
childhood, 351
diagnosis of, 347
long-term surveillance of, 348–349, 348f
management of, 347
monitoring of, 348, 348f
prognosis of, 347
screening for, 347
Gestation, 863
Gestational diabetes mellitus, 604–605
diagnosis of, 614–615
screening for and diagnosis of, 615t
Gestational trophoblastic disease
human chorionic gonadotrophin and
diagnosis of, 349
monitoring of, 349–350
prognosis of, 349
in screening for, 349
tumor markers in, 348–349
Gigantism, 778
Gilbert syndrome, 520b

Gitelman syndrome, 669–670
Glass electrode, 148–149
Glomerular diseases, 666–668
primary, 667
Glomerular filtration, 659
pregnancy-related changes in, 866
Glomerular filtration rate (GFR), 373–375, 656–657, 659, 989t–1025t
creatinine and, 373–374, 374f
clearance of, 374
filtration rate of, 373–374
estimating, 368t, 374–375
future considerations for, 375
kidney failure and, 659
markers for
endogenous, 373
exogenous, 373
reference intervals for, 375
Glomerular permeability, 659
Glomerulonephritis, 661
Glomerulus
filtration by, 654
permeability of, 658
Glucagon, 378, 442t–444t, 989t–1025t
glucose metabolism affected by, 611
Glucagon-like peptide-1 (GLP-1), 608
Glucagonomas, 611
Glucocorticoids
deficiency of, 795–803, 796f
description of, 789
Gluconeogenesis, 380
Glucose, 626b
blood
concentration, regulation of, 380
glycated hemoglobin and, 620–623
growth hormone effects on, 777, 777b
hormones regulating hormones, 606–609, 606f
monitoring of, 618–619
self-monitoring of, 617–618
critical risk value, 1026t–1027t
fasting plasma, 613
formula for, 379, 379f
hormones that influence
glucagon, 611
somatostatin, 609
thyroxine, 609
impaired fasting tolerance, 605–606
impaired glucose tolerance, 605–606
measurement of, 384–386
in blood, 384–386
point- of- care testing and, 277
reference intervals for, 386, 386t
specimen collection and storage in, 384
in urine, 386
reference intervals, 989t–1025t
Glucose dehydrogenase methods, 386
Glucose meters, 618
analytical goals for, 617–618
performance of, 618
Glucose oxidase, 277
Glucose oxidase methods, 385–386
Glucose strips, 277, 277f
Glucose tolerance test, oral, 613–614, 614b
Glucose transport, 607
facilitative human, 607t
Glucose-6-phosphate dehydrogenase (G6PD), 907, 908t
deficiency of, 895
Glucose-dependent insulinotropic peptide (GIP), 739t
Glutamate, 290t–291t, 294
Glutamic acid, 298–299, 989t–1025t
Glutamine, 290t–291t, 989t–1025t
γ-Glutamyltransferase, 321–322, 322f, 989t–1025t
biochemistry of, 322
clinical significance of, 322
methods of analysis for, 322
reference intervals for, 67f, 69t, 71t
Glutaric acidemia type I, 887–888, 891
Glutathione, 295, 297
Glutathione peroxidase, 491
Glycated hemoglobin (A_{1c}), 626b, 989t–1025t
assays for, standardization of, 622
blood glucose concentrations and, 620–623
formation of, 620f
high-performance liquid chromatography measurement of, 621
measurement of, 620
methods for determination of, 621–622
performance goals, 623
reference intervals for, 623
specimen collection and storage, 623
units for reporting of, 622–623
Glycated serum proteins, 623
Glyceraldehyde-3-phosphate, 317
Glycerol esters, 396–399, 398f
Glycine, 290t–291t, 294, 989t–1025t
Glycogen, 380, 606
Glycogen storage diseases, 383–384, 895
Glycogenesis, 380
Glycogenolysis, 380
Glycolysis, 380
Glycopeptides, 552t
Glycoproteins, 380
Glycosyl transferase, 494
Glycosylation, of proteins, 299
Glycosylphosphatidylinositol anchor, of proteins, 298
"Goblin," 596
Goiter, 827
Gold, 147
Gonadotropin-releasing hormone, 442t–444t, 834, 845
Gonadotropins, 444–445
Goodness-of-fit test, 69
Gout, 371–372
G-protein-coupled receptors (GPCRs), 445
signal transduction, 446
Gratings, in spectrophotometers, 132
Graves, Robert James, 2
Graves disease, 827, 829–830
neonatal, 870
Gravimetry, 116–117
Growth hormone, 442t–444t, 989t–1025t
biochemistry of, 776–777
blood glucose levels affected by, 777, 777b
clinical significance of, 778
deficiency of, 778–779
description of, 608
excess of, 778
measurement of, 779
physiological actions of, 777–778
pituitary tumors that secrete, 778
resistance, 779
secretion of, 777–779, 779b
synthesis of, 776
Growth hormone-inhibiting hormone. *see* Somatostatin
Growth hormone-releasing hormone, 442t–444t, 776
Growth retardation, 778–779
Gut hormones, 335t
Gynecomastia, 839

H

Hair
analysis
arsenic poisoning detected by, 594
drug abuse detected using, 587–588
specimen, 86
steroid hormones and, 804
Hairpin probes, 953, 954f
Half-life, 441, 567, 577t
Handheld integrated cartridge technology, 278
Haplotype, 898, 902
Hapten, 237–238
Haptoglobin, 302, 989t–1025t
Harmonization, 101–102
Hartnup disease, 487–488
Hashimoto thyroiditis, 823–825
Haworth formula, for sugars, 379, 379f
Hazards, 120–125
biological, 121–123, 122f
chemical, 123–124
from compressed gases, 124
electrical, 124
fire, 124–125
identification of, 120–121, 121f
from volatiles, 123–124
Hb Spanish Town, 505
HCV RNA measurement, 710
Headspace analysis, 188
Health-associated reference values, 66t
Healthcare resources, allocation of, 4
Heart
anatomy and physiology of, 631–635, 631f
tacrolimus in, 556t

Heart failure. *see* Congestive heart failure
Heat-resistant (nonasbestos) gloves, 118–119
Height equivalent of a theoretical plate (HETP), 184
Helena SPIFE 4000 (Helena Laboratories, Beaumont, TX), 173
Helicobacter pylori, in stomach, 728–729, 728b–729b
HELLP syndrome, 869
Hemagglutination, 249
Hematocrit (Hct), 506t
Hematofluorometer, 140
Hematologic and immunologic analyzers, 279–280
Hematology, 47–48
Hematopoietic malignancies
  gene mutations in, 976–977
  genetics of, 974
  massively parallel in, 977
Hematuria, 660
Heme, 501
  biosynthesis of, 527–528
    enzymes of, 527–528, 527t
    regulation of, 528–529
  function of, 528–529
  precursors, solubility and excretion of, 529, 529t
Hemochromatosis, 515–516, 704, 716
Hemodiafiltration, 671
Hemodialysis, 673
  assessment of adequacy of, 673
Hemodialyzer setup, 673f
Hemoglobin (Hb), 499–523, 506t
  A, 431, 867
  Bart's, 503
  biochemistry of, 501–502, 501f
  clinical significance of, 502–505
  definition of, 501
  fetal, 568, 867
  globin structure of, 501
  glycated (A_{1c}), 626b, 989t–1025t
    assays for, standardization of, 622
    blood glucose concentrations and, 620–623
    formation of, 620f
    high-performance liquid chromatography measurement of, 621
    measurement of, 620
    methods for determination of, 621–622
    performance goals, 623
    reference intervals for, 623
    specimen collection and storage, 623
    units for reporting of, 622–623
  H
    determining, 507–508
    disease, 503
  modified, 501–502, 501f
  oxygen saturation, 432–433
  physiological role of, 502, 502f
Hemoglobin *(Continued)*
  S, 505, 505f
    solubility tests, 507–508, 509f
  unstable, tests for, 508
  variants of, classification of, 505
Hemoglobinopathies, 502–503, 505
  analytical methods for, 505–508
  DNA analysis for, 507
  electrospray mass spectroscopy for, 507
  nomenclature of, 505
  specific tests for, 507–508
  techniques for, 505–507
Hemoglobin-oxygen dissociation, 433
Hemolysis, 41–43, 44b, 84
  background of, 41
  in blood specimens, 87
  definition of, 41, 84
  detection of, 43, 44f
  mechanisms of interference caused by, 41–42
    chemical interference as, 42
    lipemia and. *see* Lipemia
    release of cell components into sample as, 41–42, 41t
    sample dilution as, 42
    spectrophotometric interference as, 41
  in vitro, 41
  in vivo, 41
Hemolytic anemia, 327
Hemolytic disease of the newborn, 870–871
Hemorrhagic disease of the newborn, 476
HemoSense device, 278
Hemosiderin, 509
Hemostasis testing, 47
Henderson-Hasselbalch equation, 420, 688–689, 689f
Henry's law, 430, 430t
Heparin, blood specimen addition of, 83
Hepatic extraction rate, 544
Hepatic glycogenesis, 719
Hepatic tumor, 720
Hepatic viral infection, 707–711, 712b
  hepatitis A virus, 707
  hepatitis B virus, 707–710
  hepatitis C virus, 709–710, 710f
  hepatitis D virus (delta agent), 710
  hepatitis E virus, 711
  hepatitis G virus, 711
  types of, 708t
Hepatitis
  acute, 711–712, 711t
  alcoholic, 703
  autoimmune, 715–716, 716t
  chronic, 712–717, 713t
  ischemic, 712
  during pregnancy, 870
  toxic, 712, 712f
  viral, 713b
Hepatitis A virus, 707
Hepatitis B e antigen (HBeAg), 708
  mutants, 708–709
Hepatitis B surface antigen (HBsAg), 708
  mutants, 708
Hepatitis B virus, 707–710
  acute, course of, 709b
  chronic, 713–714, 714t
  diagnostic tests for, 708–709
Hepatitis C core antigen (HCV Ag), 710
Hepatitis C genotype, 710
Hepatitis C virus, 709–710, 710f
  chronic, 715
  diagnostic tests for, 710
Hepatitis D virus (delta agent), 710
Hepatitis E virus, 711
Hepatitis G virus, 711
Hepatobiliary disease, alkaline phosphatase in, 320, 320t
Hepatoblastoma, tumor markers for, 351
Hepatocellular carcinoma (HCC), 704, 720
  α-fetoprotein and
    diagnosis of, 350
    posttreatment monitoring, 350
    prognosis of, 350
    screening for, 350
  tumor markers for, 349–350
Hepatocytes, 702
  ultrastructure of, 702, 704f
Hepcidin, 296, 510–511, 716
  methods for, 517
Hepcidin-ferroportin axis, disorders of, 515–516
Hereditary coproporphyria (HCP), 530t
Hereditary hemochromatosis (HH), 508–509
Hereditary persistence of fetal hemoglobin (HPFH), 504
Hermaphroditism, 846–847
Heroin, 582f, 583t
Heteroduplexes, 948
Heterogeneous immunoassays, 242
Heterophilic antibodies, 341–342
Heteroplasmy, 926
Heteroscedastic, 243
Heteroscedasticity, 52
Hexokinase (HK) methods, 384–385, 384f–385f
High performance liquid chromatography (HPLC) detectors, 132
High-density lipoprotein cholesterol (HDL-C), 410t, 415, 989t–1025t
High-density lipoproteins, 401, 401t
High-dose "hooking," of tumor markers, 341, 341f
Highly quenched probes, 953, 954f, 957f
High-performance liquid chromatography (HPLC), 190, 507, 508f
  column sizes used in, 196t
  drug screening uses of, 567
  glycated hemoglobins and, 621
High-pressure liquid chromatography, 191
High-resolution electrophoresis, 172
High-resolution melting analysis (HRMA), 954, 956f

High-sensitivity C-reactive protein, 416–417, 417f
High-vacuum pumps, 207
Hirsutism, 846, 850–851, 850b
Histamine (H_2)-receptor antagonists, for *Helicobacter pylori*, 728
Histidine, 290t–291t, 989t–1025t
Histogram, 70f
Histones, 923
Holoenzyme, 229, 317
Homeostasis
  energy, 445
  water, 420
Homeostatic model, 58
Homocysteine, total, 989t–1025t
Homocystinuria, 889
Homoduplexes, 948
Homogeneity, 54, 54f
Homogeneous immunoassays, 242
Homovanillic acid, 989t–1025t
Hook effect, 242
Hormone(s), 440–449, 442t–444t, 448b. *see also specific hormone*
  action of, 441–445
    clinical disorders of, 447
  endocrine control of reproduction, 445
  energy homeostasis and integrated metabolism, 445
  growth, development, and maturation, 444–445
  measurement of, 447–448
  menstrual cycle, 844–846
  pituitary. *see* Pituitary hormones
  placental, 864
  steroid. *see* Steroid hormones
  synthesis and release of, 441–444
Hormone receptors
  cell signaling mechanisms of, 445
  cell surface, 445, 446f
Horseradish peroxidase immunoassays, 244, 244f
Human anti-mouse antibodies, interference from, of tumor markers, 341–342
Human chorionic gonadotropin, 334t, 339t–340t. *see also* Chorionic gonadotropin.
  analytical methods for, 876–877
  assays for, 876
  biochemistry of, 877
  concentration of, 868f
  description of, 868f, 876
  in ectopic pregnancy, 869
  in germ cell tumor, 347
  gestational trophoblastic disease and diagnosis of, 349
    screening for, 349
  point-of-care testing for, 877
  stimulation test, 853
Human chromosomal DNA, structure of, 918f, 923
Human epidermal growth factor receptor-2, 336t
Human epididymal protein 4 (HE4), 335t, 352
Human genome, variation in, 926–927
Human Genome Variation Society (HGVS) nomenclature, 911
Human glucose transporters, facilitative, 607t
Human immunodeficiency virus, HIV-1 antibody, 239, 239f
Human immunodeficiency virus 1, 962–963, 963b
Human leukocyte antigen complex, 911, 912t
Human microbiome, 967
Human placental lactogen, 864
Human reference genome, 923–925, 924t
Humoral hypercalcemia of malignancy (HHM), 756
Huntington disease, 298, 969–971, 970f, 970b
Hybrid isoenzymes, 218
Hybridization, 939, 948, 949f, 949b
Hybridization probes, 952–953, 954f
Hydatidiform moles, tumor markers for, 349
Hydramnios, 864–865
Hydrocodone, 582f, 583t
Hydrogen
  breath tests, for disaccharidase deficiencies, 732
  as H_2PO_4-, 691
Hydrolysis probes, 952, 954f
Hydromorphone, 582f, 583t
12-l-Hydroperoxy-5,8,10,14 eicosatetraenoic acid (HPETE), 400
Hydrophilic interaction liquid chromatography (HILIC), 194
Hydrops fetalis, 871
Hydroxide, concentration of, 364
β-Hydroxybutyrate, 619
  determination of, 388, 388f
5-Hydroxyindoleacetic acid, 463–464
11ß-Hydroxylase, deficiency of, 847
21-Hydroxylase
  congenital adrenal hyperplasia caused by, 801
  deficiency of, 801, 847
Hydroxylation, of proteins, 299
Hydroxymethylbilane synthase (HMBS), 527
17-Hydroxyprogesterone, 791, 807, 989t–1025t
3ß-Hydroxysteroid dehydrogenase deficiency, 847
25-Hydroxyvitamin D, measurement of, 762
Hyperaldosteronism, 802
Hyperamylasemia, 325, 325t
Hyperbilirubinemia, 519
  conjugated, 520b, 521
  unconjugated, 519–520, 520b
Hypercalcemia, 748
Hypercalcemia-associated malignancy (HCM), 748
Hyperchloremia, 688
Hyperchylomicronemia, 407–408
Hypercortisolism, 800b
Hyperemesis gravidarum, 870
Hyperglycemia, 604, 612
Hypergonadotropic hypogonadism
  causes of, 838b
  evaluation of, 855, 855f
  in males, 838
Hyperkalemia, 687, 687f
Hyperkalemic normal anion gap acidosis, 694
Hyperlipidemia
  causes of, 408t
  familial combined, 408
Hyperlipoproteinemia
  type III, 409
  type V, 408
Hypermagnesemia, 755
Hypermineralocorticoidism, 803t
Hypernatremia, 684–685, 685f
Hyperosmotic hyponatremia, 684
Hyperphosphatemia, 753
Hyperprolactinemia, 780, 780t, 853
Hyperreninemic hypoaldosteronism, 797t
Hypertension, 666
Hyperthyroidism, 828–829, 828b
Hypertrichosis, 850
Hypertriglyceridemia, familial, 408
Hyperuricemia, 371
  causes of, 372b
Hypervolemia, 684
Hypervolemic hypernatremia, 685
Hypoaldosteronism, 794–795, 797t
Hypoalphalipoproteinemia, 409
Hypocalcemia, 747–748
Hypochloremia, 688
Hypoglycemia, 380–382, 608
  in diabetes mellitus, 382
  fasting, 609
  fasting, in adults, 381–382
  in neonates and infants, 381
  postprandial, 382
  unawareness, 382
Hypogonadotropic hypogonadism
  causes of, 838b
  evaluation of, 855–856
  in males, 838
Hypokalemia, 686–687, 686f
Hypomagnesemia, 755
Hyponatremia
  definition of, 682–684
  differential diagnosis of, 683f
  dilutional, 683
  hyperosmotic, 684
  hypoosmotic, 683–684
  isosmotic, 684
  syndrome of inappropriate antidiuretic hormone as cause of, 783
Hypoosmotic hyponatremia, 683–684

Hypoparathyroidism, 747
Hypophosphatemia, 753
Hyporeninemic hypoaldosteronism, 797t
Hypothalamic diabetes insipidus, 782
Hypothalamic-pituitary portal system, 775f
Hypothalamic-pituitary-gonadal axis
in female reproductive biology, 839
in male reproductive biology, 835–839
Hypothalamus, pituitary hormone secretion regulated by, 776–784, 776t, 777f
Hypothyroidism, 823–826, 824b–825b, 826t, 829, 831–832
Hypouricemia, 372
Hypovolemia, 683–684, 790
Hypovolemic hypernatremia, 684
Hypoxia
cellular, agents causing, 567–568
tissue, lactic acidosis caused by, 694

I

Icterus, 43, 44b, 519
detection of, 43, 44f
mechanisms of interference caused by, 43
Identification, specimen, 86, 88b
704-M Identification System, 121
Idiopathic postprandial syndrome, 382
Idiopathic reactive hypoglycemia, 382
IgA nephropathy, 667
IGF-1, 442t–444t
IGF-2, 442t–444t
Ileum, 724
Illicit drugs. *see* Drugs of abuse
Imerslund-Grasbeck syndrome, 481–482
Imipramine, 572f
Immobilized metal-ion affinity chromatography, 194
Immunoaffinity chromatography, 194
Immunoassay systems, separation in, 264
Immunoassays, 237
analytical and functional sensitivity of, 243
calibration of, 243
chemiluminescence, 246–247, 246f
clinical analysis uses of
amphetamine, 577
estrogen, 858, 859b
fetal fibronectin, 879
lysergic acid diethylamide, 563, 566–567, 581
testosterone, 857, 859b
cloned enzyme donor, 244f, 245
competitive, 241, 241f, 243
designs, 241f
electrochemiluminescence, 246f, 247–248
enzyme, 244–245, 244f
enzyme-multiplied technique, 244–245
fluoroimmunoassay, 245–246, 245f, 245t
glycated hemoglobins and, 621–622
heterogeneous, 242
homogeneous, 242
horseradish peroxidase immunoassays, 244, 244f
Immunoassays *(Continued)*
interferences in, 248–249
labeled, example of, 243
multianalyte, 247–248
noncompetitive, 241–242
nonisotopic, 241t
particle, 247, 247f
phosphor, 246
radioimmunoassay, 243
sandwich, 246–247
simplified, 247
simultaneous multianalyte, 247–248
surface plasmon resonance-based, 240
thyroid-stimulating hormone, 819
total thyroxine measurement using, 820
triiodothyronine measurement using, 821
Immunochemical assays, 234
labeled, 240–249
Immunochemical techniques, 236–250
basic concepts and definitions of, 237–238
cell-based and tissue-based, 249
immunoassays. *see* Immunoassays
immunofixation, 239, 239f
protein microarrays, 247–248
qualitative methods, 239–240, 239f
quantitative, 240–249, 241b–242b
Immunocytochemistry, 249
Immunofixation, 239, 239f
Immunofixation electrophoresis, 307, 309f
Immunogen, 237–238
Immunoglobulins, 303, 303f
A, 304, 989t–1025t
in celiac disease, 731
nephropathy. *see* IgA nephropathy
D, 304, 989t–1025t
deficiency in, 304–305, 304f
E, 304, 989t–1025t
G, 303–304, 989t–1025t
in celiac disease, 731
characteristics of, 864
maternal, 864
transient physiologic deficiency of, 304–305
M, 304, 989t–1025t
measurement of, clinical utility of, 304–305
monoclonal, 305, 305t
schematic diagram of, 237f
synthesis, by liver, 703–704
Immunophilin, 557
Immuno-polymerase chain reaction (Immuno-PCR), 247
Immunostrips, 276–277
Immunosuppressant, 555
in kidney transplantation patients, 676t
Immunosuppression, 675–677
Immunosuppressive drugs, 555–558
Impaired glucose tolerance, 605–606
Impaired hepcidin-ferroportin axis, 511t–512t
Implanted sensors, 619
In vitro fertilization, 876
In vitro hemolysis, 41
In vivo hemolysis, 41
Inborn errors of metabolism, 883
definition of, 883
diagnosis of, 895–896
newborn screening for, 882–897, 884b
Incandescent lamps, 131
Indels, 927
Index of individuality (II), 56
Indicator reactions, 222, 225
Indirect bilirubin, 521
Inductively coupled plasma, 206
Inductively coupled plasma-mass spectrometry (ICP-MS), 134, 213
Inert metal electrodes, 147
Infarction. *see* Acute myocardial infarction
Infectious disease, 962–963
Infertility
definition of, 851–856
female, 853b–854b, 854–856, 854f
male, 851–853, 851b–852b, 852f, 852t
Inflammation
trace elements and, 467
vitamins and, 467
Inflammatory bowel disease and irritable bowel syndrome, 733–736, 736b
calprotectin, 734, 735b
functional small bowel length in, assessment of, 735–736
protein-losing enteropathy, 736
therapeutic drug monitoring in, 735
Infliximab (IFX), for inflammatory bowel disease, 735
Influences, in urine testing, 44–45
Influencing factors, 39, 40b
modifiable, 39–40
unmodifiable, 40, 40t
Influenza A (H1N1), 120
Influenza viruses, pandemic plan for, 120
Inherited diseases, 372
Inherited liver disease, 716–717
Inherited tubulopathies, 669–670
Inhibin, 834–835
Inhibin A, 335t, 873f, 878–879, 989t–1025t
Inhibin B, 335t, 442t–444t
Inhibitor, 226
Injection systems
in gas chromatography, 187–188
in liquid chromatography, 195–196
Inner filter effect, 136
in fluorescence measurements, 140
Inorganic acidosis, 694–695
Inorganic lead, 597f
Inorganic phosphate, 753
Insecticides, organophosphate, 574
Insert, 956
Insertion, 925
Institute for Reference Materials and Measurements (IRMM), 111, 133
Instrument clusters, 265, 265f–266f

Instrumentation, 130–133
Insulin, 378, 442t–444t, 606–607, 989t–1025t
measurement of, 609–610
Insulin autoantibodies (IAAs), 611
Insulin resistance, 612
Insulin-dependent diabetes, 876
Insulin-like growth factor
characteristics of, 776–777
description of, 607–608
measurement of, 779
Insulin-like growth factor binding protein, 777–778
Insulin-like growth factor-1, 989t–1025t
Insulin-like growth factor-II, 989t–1025t
Insulinoma-associated antigens, 611
Intact parathyroid hormone methods, 757–758, 758f
Interference factors, 40, 40b
Interference filters, 132
Interferences
endogenous, 40–48
exogenous, 40, 43–44
in urine testing, 44–45
Interlab Microgel system (Interlab Srl, Rome, Italy), 173
Interleukin-1, 776
Interleukin-6, 776
Intermediate precision, 15
Intermediate-density lipoproteins, 401, 401t
Internal quality control, 91, 91b, 101b
Internal standard, 186, 186f
stable isotope-labeled, 203
International Agency for Research on Cancer (IARC), 600
International Committee of Medical Journal Editors, 5
International normalized ratio (INR), 898, 900, 900b
measurement, point-of-care testing and, 278
International Union of Biochemistry enzyme nomenclature system, 216
International unit, 221
Interpercentile interval, 68
Interstitial fluid, 681–682
Interstitial nephritis, 668
Intestinal disorders, 730–736. *see also* Large intestine; Small intestine.
bacterial overgrowth, 732–733
bile acid malabsorption, 733, 733b
celiac disease, 730–731, 731t
disaccharidase deficiencies, 731–732, 732t, 732b
inflammatory bowel disease and irritable bowel syndrome, 733–736
Intestinal function, 723–743
Intoxication, 563
salicylate, 571–572, 694
stages of, 569t
treatment for, 564t
Intracellular cholesterol transport pathway, 404, 405f
Intracellular fluid (ICF), 682
Intracellular nuclear receptors, signal transduction and, 446–447
Intron, 916, 920
Invasive moles, tumor markers for, 349
Invasive tests, of exocrine pancreatic function, 737
Iodine, 494, 989t–1025t
biochemistry of, 817
containing products and goitrogens, 817
deficiency of, 817
dietary, 814
disorders, 817
excess, 817
supply, geography and, 817
urinary, measurement of, 822
Iodoacetate, 83
Iodothyronine deiodinase, 491
Ion exchange, in reagent grade water, 110
Ion trap, 207
Ion-exchange chromatography (IEC), 192
Ionization constants, 290t–291t
Ionization interference, 135
Ion-pair chromatography (IPC), 194
Ions, 202
motility of, in electrophoretic system, 169b
Ion-selective electrode (ISE), 146, 148–153, 148t, 420–421
for determination
of chloride, 424
of sodium and potassium, 422–423
direct potentiometry by, 152–153, 153t
Ion-selective membrane electrode-based potentiometric cell, 146f
Iontophoresis, pilocarpine, 424
Iron, 499–523, 989t–1025t
acquisition of, disorder, 516
analytical methods for, 516–518
balance, 511
body, distribution of, 509–511, 510t
clinical relevance of, 517
clinical significance of, 511–516
deficiency, 511–512, 515f
description of, 591–592
gold standard assessment of, 518
homeostasis of, 510–511
laboratory tests for, 513t–514t
metabolic acidosis associated with, 694
metabolism of, 509–511
copper and, 490
disorders, 511t–512t
overload, 515–516
reference intervals of, 517
regulation of, systemic and cellular, 509
serum, methods for, 516–517
tissue, 509
transport of, 510
disorder, 511t–512t, 516
Irreversible inhibition, 228
Irritable bowel syndrome, 733–736
Ischemia
metabolic acidosis associated with, 694
myocardial, 630
Ischemic hepatitis, 712
Islet cell cytoplasmic antibodies, 611
Isocratic elution, 194–195, 195f
Isoelectric focusing (IEF), 176, 297
electrophoresis, 177–178, 177f, 506, 507f
Isoelectric point, 168, 291–292
Isoenzymes, 218–221, 314
alkaline phosphatase, 320–321
definition of, 218
distribution of, 220
embryogenesis-related changes in, 220
hybrid, 218
measurement of, 233–234
Isoform of glutamic acid decarboxylase (GAD65), antibodies to, 611
Isoforms, 218, 314
definition of, 218
measurement of, 233–234
Isoimmunization. *see* Hemolytic disease of the newborn
Isoleucine, 290t–291t, 989t–1025t
Isoniazid, 694
Isopeptide bonds, 295
Isopropanol, 570
Isosmotic hyponatremia, 684
Isothermal elution, 188
Isotope dilution mass spectrometry (IDMS), 203
Isotopes, 203, 204f
Isovaleric acidemia, 887–888
i-STAT analyzer, 278
Ivacaftor, 911

J

Jaffe reaction, 363–364, 989t–1025t
Jaundice, 519, 703
physiologic classification of, 520b
Jejunum, 724
Juvenile-onset diabetes, 605
Juxtaglomerular apparatus (JGA), 655, 655f, 789–790, 790f

K

Kallmann syndrome, 838
Karyotype, 974
Kashin-Beck disease, 492
Katals, 221
Keppra. *see* Levetiracetam
Keratomalacia, 473
Kernicterus, 519
Keshan disease, 492
Ketamine, 583–584
Ketoacidosis, 694
α-Ketoacids, 293
Ketone, 378, 378f
formation of, 396
Ketone bodies, 396, 398f, 619–620, 619f
in urine, 620

Ketonemia, 620
Ketones, clinical significance of, 620
Ketonuria, 620
Ketoses, 378, 378f
Ketosis, 396
17-Ketosteroids
  description of, 835
  in urine, 858
Ketostix, 388
Kidney(s)
  anatomy of, 653, 653f
  ascending loop of Henle, 657
  blood supply to, 655–656
  countercurrent multiplication mechanism, 657f, 658
  disease of, 666–671
  distal tubule of, 657
  electrolyte homeostasis by, 656–657
  endocrine function of, 658
  epithelial cell membranes, 656
  excretion by, 655
  fetal development of, 867
  function of, 656–658
  glomerular filtration rate. *see* Glomerular filtration rate
  glomerulus. *see* Glomerulus
  hypertension, 666
  metabolic acidosis compensatory mechanisms, 695
  nephron, 653–655, 654f
  physiology of, 658–659
  pregnancy-related changes in, 866
  proximal convoluted tubule of, 654
  regulatory function of, 656–658
  sodium reabsorption, 655
  tacrolimus in, 556t
  transplantation of, 674–677, 675f, 675b
  water homeostasis by, 657–658
Kidney diseases, 372, 651–678
  chronic. *see* Chronic kidney disease
  diabetes insipidus, 670
  diabetic nephropathy, 666
  failure. *see* Kidney failure
  hematuria associated with, 660
  IgA nephropathy, 667
  inherited tubulopathies, 669–670
  interstitial nephritis, 668
  monoclonal light chains and, 668–669
  nephrotic syndrome, 667
  nonsteroidal antiinflammatory drugs in, 670
  obstructive uropathy, 669
  overview of, 660
  pathophysiology of, 659–666
  polycystic, 669
  prostaglandins in, 670
  proteinuria and, 659
  rapidly progressive glomerulonephritis, 668
  renal calculi, 670
  renal tubular acidoses, 669
  terminology of, 660
Kidney diseases *(Continued)*
  toxic nephropathy, 670–671, 671t–672t
  tubular disease, 669–670
  uremic syndrome, 663–664
Kidney failure
  description of, 660
  uremia of, 694
Kidney function tests, 359–376
Kidney stones, 372, 670
Kinetic methods, 232–233
Kisspeptin, 442t–444t
KIT protein, 347
Kjeldahl method, for total protein measurement, 305–306
Klinefelter syndrome, 838
KRAS mutation, 336t
Krebs (citric acid) cycle, 380, 395
Kringle domain, 401
Kwashiorkor, 292–293

L

"Lab Right to Know Standard," 118
Labels, definition of, 240
Labile iron pool, 509–510
Laboratory errors, 39
Laboratory information systems (LISs), 253, 281
Laboratory medicine, definition of, 1
Laboratory safety, principles of, 108–126
Laboratory testing, definition of, 1
Laboratory tests. *see also specific test*
  diagnostic accuracy of, 28–35
    in clinical context, 32–35
    combinations of, 33–34, 34f
    comparison of two, 32–35, 32f, 33b
  results of, unavoidable influences on, 40t
Lactase deficiency, 732
  prevalence of, 732t
  tests for, 732, 732b
Lactase-phlorizin hydrolase complex, 726
Lactate, 382–383
  in whole blood, measurement of, 387, 387f, 387t
Lactate dehydrogenase (LDH), 319, 327, 327f, 334t
  biochemistry of, 327
  clinical significance of, 327
  critical risk value, 1026t–1027t
  hemolytic anemia effects on, 327
  liver disease effects on, 327
  macro-, 327
  malignant disease effects on, 327
  methods of analysis for, 327
  molecular weight of, 327
  reference intervals, 989t–1025t
Lactate plasma, 1026t–1027t
Lactic acidosis, 383, 694
Lactose intolerance, 731
Lactulose, 733
Lagging strand, 919
Lamellar bodies, 880
Lamictal. *see* Lamotrigine
Lamotrigine, 550, 551t
Lanoxin. *see* Digoxin
Large intestine, 724. *see also* Intestinal disorders.
Large-scale discovery efforts, 930
Laser, 131
Latex allergy, 121–122
Laxative abuse, 741
L-dopa, 452–453
Lead (Pb), 597–598, 597f, 989t–1025t
Lecithin, 880
Lecithin cholesterol acyltransferase, 393
Lecithin/sphingomyelin ratio, 880
Leptin, 442t–444t
Leucine, 290t–291t, 989t–1025t
Levetiracetam, 550, 551t
Levey-Jennings chart, 92, 92f, 99f
Levorphanol, 586
Leydig cells, 838
Liddle syndrome, 669–670
Li-Fraumeni syndrome, 977
Ligandins, 518–519
Light, 129
  in spectrophotometer, 131
Light emitting diodes (LEDs), 131
Light pipes, 132–133
Light scattering, 142
  assays, 143b
  in fluorescence measurements, 141
  limitations of, 143–144
Light-scattering immunoassay, 279, 279f
Light-scattering methods, for total protein measurement, 306
Likelihood ratio, 31
Limit of detection (LOD), 16
Limits of quantification, 15–16
Linear ion trap, 209
Linearity, 15
  testing for, 23–25, 24f
Lineweaver-Burk plot, 223, 223f
Linoleic acid, 399–400
Lipase, 323–325, 323f
  biochemistry of, 323
  clinical significance of, 323–324, 324f
  critical risk value, 1026t–1027t
  definition of, 323
  methods of analysis for, 324–325, 324f–325f
  reference intervals, 989t–1025t
Lipemia, 42, 44b
  detection of, 43, 44f
  mechanism of interference caused by, 42
    partitioning of sample as, 42
    physico-chemical mechanisms as, 42
    spectrophotometric interference as, 42
    volume depletion effect as, 42
  removal of lipids from sample and, 43
Lipid(s), 392
  acylglycerols, 396–399
  analysis of, 413–416, 414t
  basic, 392–400
  cholesterol. *see* Cholesterol

Lipid(s) *(Continued)*
clinical significance of, 405–410
coronary heart disease and, 405–406
fatty acids. *see* Fatty acids
prostaglandins, 399–400, 399t–400t
in pulmonary surfactant, 867
sphingolipids, 399
synthesis, by liver, 704
terpenes, 400
Lipid panel, 410
Lipoprotein(s), 392, 400–401, 415–416
a, 409–410
analysis of, 413–416, 414t
apolipoproteins, 402
chemical composition of, 402t
chylomicrons, 401, 401t
classification of, 401, 401t
high-density, 401, 401t
intermediate-density, 401, 401t
low-density, 401, 401t
classification, 410t
steroid hormone production and, 789
metabolism of, 403–405
endogenous pathway, 403–404, 404f
exogenous pathway, 403, 403f
genetic disorders of, 406–409
intracellular cholesterol transport pathway, 404, 405f
reverse cholesterol transport pathway, 405, 406f
structure of, 401f–402f
synthesis, by liver, 704
very low-density, 401t
X, 409–410
Lipoprotein disorders
dysbetalipoproteinemia, 409
familial combined hyperlipidemia, 408
familial defective apolipoprotein B-100, 409
familial hypercholesterolemia, 409
familial hypertriglyceridemia, 408
hypoalphalipoproteinemia, 409
management of
adult, 410–413
in pediatrics, 412–413
Tangier disease, 409
type V hyperlipoproteinemia, 408
Lipoprotein lipase, 403
activity, deficiency in, 407–408
Liquid biopsy, 977, 978f
Liquid calibration systems, 279
Liquid chromatography (LC), 190–198
columns and supports in, 196, 196t
data acquisition and system control in, 198
detectors of, 196–198, 197t
high-performance. *see* High-performance liquid chromatography
high-pressure, 191
hydrophilic interaction, 194
injection systems and sample derivatization in, 195–196
instrumentation of, 194–198
Liquid chromatography *(Continued)*
mobile phase reservoirs and delivery systems of, 194–195, 195f
normal-phase, 192
reversed-phase, 192
temperature control in, 196
types of, 191–194
ultra-high-performance, 191
Liquid chromatography-electrochemical detection (LC-EC), 197
Liquid chromatography-mass spectrometry (LC-MS), 197–198, 197t
applications of, 210f, 212
metabolic disorders, screening for, 212
total cortisol evaluations, 805
Liquid chromatography-tandem mass spectrometry (LC-MS/MS), 295
Liquid controls, 97
Liquid-level sensing, 261–262
Liquid-solid chromatography (LSC), 191
Lithane. *see* Lithium
Lithium, 559, 817–818
Lithonate. *see* Lithium
Lithotripsy, 670
Liver
acinus of, 702, 703f
albumin synthesis by, 703
ammonia metabolism by, 705, 705f
anatomy of, 700–702
gross, 700–701, 701f
microscopic, 701–702, 702f
α_1-antitrypsin synthesis by, 704
bile acids metabolism by, 703
biliary drainage in, 701
bilirubin, 703
biochemical functions of, 702–706
blood supply to, 700–701
carbohydrate metabolism by, 705–706
ceruloplasmin synthesis by, 704
excretory function of, 702–703
fetal development of, 867
α-fetoprotein synthesis by, 704
immunoglobulins synthesis by, 703–704
lipid and lipoprotein synthesis by, 704
metabolic function of, 705–706
nutritional and metabolic abnormalities in, 706
protein synthesis by, 703–705
storage function of, 706
synthetic function of, 703–705
tacrolimus in, 556t
transplantation, 720
transthyretin synthesis by, 703
urea synthesis by, 704–705
Liver disease, 699–722
alcoholic, 717–718
aspartate aminotransferase affected by, 317
bilirubin metabolism, disorders of, 707
cholestatic, 719–720
chronic, 717b
cirrhosis. *see* Cirrhosis
disordered hemostasis in, 706
Liver disease *(Continued)*
drug-induced, 717
enzyme released from tissue of, 706
gallstones, 720
hepatic tumor, 720
hepatic viral infection, 707–711, 712b
hepatitis A virus, 707
hepatitis B virus, 707–710
hepatitis C virus, 709–710, 710f
hepatitis D virus (delta agent), 710
hepatitis E virus, 711
hepatitis G virus, 711
types of, 708t
hepatitis. *see* Hepatitis
inherited, 716–717
lactate dehydrogenase levels in, 327
mechanisms and patterns of, 706–707, 707b
natural history of, 706f
nonalcoholic fatty, 712–713, 715
pregnancy-related, 869–870, 872b
Liver enzymes, 317–323, 317f
alkaline phosphatase. *see* Alkaline phosphatase
aminotransferase. *see* Aminotransferases
cholinesterase. *see* Cholinesterase
γ-glutamyltransferase. *see* γ-Glutamyltransferase
5'-nucleotidase, 323
Liver function tests, 707b
L-lactate, 989t–1025t
Long-chain 3-hydroxyacyl-CoA dehydrogenase deficiency, 892
Loop of Henle, 654
Loop-mediated amplification methods, 942
Loop-mediated isothermal amplification (LAMP), 942
Lorazepam, 578t
Lot-analyzed reagent, 110
Low molecular weight proteins, 659
Low-density lipoprotein cholesterol (LDL-C), 410b, 411t, 989t–1025t
Low-density lipoproteins, 401, 401t
classification, 410t
steroid hormone production and, 789
Lower limit of quantification (LLOQ), 15–16
Lowry (Folin-Ciocalteu) method, for total protein measurement, 306
LSD. *see* Lysergic acid diethylamide
Luminometry, 141–142
Lung cancer, tumor markers for, 350
management of, 350
monitoring, 351
prognosis of, 351
screening and as aids to diagnosis of, 350–351
therapy prediction, 351
Lungs
fetal development of
description of, 867
tests for assessing, 880
tacrolimus in, 556t

Luteinizing hormone (LH), 442t–444t, 834, 845, 989t–1025t
characteristics of, 780, 781b
physiological action of, 780
Lyophilization, 117
Lysergic acid diethylamide, 566–567, 581
Lysine, 290t–291t, 989t–1025t
Lysosomal storage disorders, 892–894
Lysosomes, 702

M

M protein, 668–669
Macroamylasemia, 326f
Macrocomplex, 219
Macroduct Sweat-Chek (Wescor), 426
α2-Macroglobulin, 989t–1025t
Macroprolactin, 779
Macula densa, 654, 789–790
Magnesium, 754–756
biochemistry and physiology of, 754–755, 754t
clinical significance of, 755
critical risk value, 1026t–1027t
free (ionized)
measurement of, 755
reference intervals for, 756, 989t–1025t
specimen requirements for, 755–756
total
atomic absorption spectrometry for, 755
enzymatic methods for, 755
measurement of, 755
photometric methods for, 755
reference intervals for, 756, 989t–1025t
specimen requirements for, 755–756
Magnesium deficiency, 755
Magnetic sector mass spectrometers, 207
Malabsorption, 725, 740, 740t
Malate dehydrogenase (MD), 319
MALDI. *see* Matrix-assisted laser desorption/ionization
Maldigestion, 725, 740
Male infertility, 851–853, 851b–852b, 852f, 852t
Male reproductive system
androgens. *see* Androgen(s)
biological functions of, 834–835, 839b
erectile dysfunction, 839
hypergonadotropic hypogonadism
causes of, 838b
evaluation of, 855, 855f
in males, 838
hypogonadotropic hypogonadism
causes of, 838b
evaluation of, 855–856
in males, 838
hypothalamic-pituitary-gonadal axis in, 835–839
Manganese, 494–495, 989t–1025t
Maple syrup urine disease, 889
Marasmus, 292–293
Marfan syndrome, 889
Marijuana, 579
Mass analysis, 202
Mass spectrometers
beam-type designs, 207–208
clinical applications of, 211–213
components of, 204f
definition of, 202
inductively coupled plasma, 206, 213
instrumentation of, 203–211
analyzers, 207–211
data acquisition, 211
data analysis software, 211
detectors, 211
ion source, 203–206
vacuum system, 207
ion cyclotron resonance, 209
ionization methods, 203, 206b
atmospheric pressure chemical, 205–206, 205f–206f
atmospheric pressure photoionization, 203
chemical, 204–205, 205f
electron, 203, 205f
electrospray, 205, 206f
matrix-assisted laser desorption/, 206, 207f
linear ion trap, 209
magnetic sector, 207
quadrupole, 207–208, 208f
quadrupole ion trap, 209
tandem, 209–211, 210f
time-of-flight, 208, 208f
trapping type of, 209
types of, 207–209, 211b
Mass spectrometry (MS), 201–214, 295, 307
basic concepts of, 202–203
clinical applications of, 211–213
definitions of, 202–203
gas chromatography-mass spectrometry. *see* Gas chromatography-mass spectrometry
isotope dilution, 203
liquid chromatography-. *see* Liquid chromatography-mass spectrometry
proteomics and metabolomics for, 213
spectrometers for. *see* Mass spectrometers
tandem, 202
tumor markers and, 333
Mass spectrum, 202, 203f
Massively parallel maternal plasma DNA sequencing, 985–986
Massively parallel sequencing (MPS), 956–959, 956t, 972–973, 973t
in hematopoietic malignancies, 977
of solid tumor samples, 931–933
Mass-to-charge ratio (m/z), 202
Material Safety Data Sheet (MSDS), 119
Maternal plasma, cell-free fetal DNA in, 982
Matrix, 13–14
Matrix effects, in light-scattering measurements, 143–144
Matrix-assisted laser desorption/ionization (MALDI), description of, 206, 207f
Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry, applications of, 212–213, 213f
Maximum impurities reagents, 110
Maximum likelihood (ML), 54
MCAD deficiency, 892
Mean cell Hb concentration, 506t
Mean cell volume (MCV), 506t
Mean platelet volume, 506t
Measured value, 17
Measurement uncertainty, specifications for, 60
Measuring pipettes, 112, 112f
Mechanical bile duct obstruction, 719
Meconium, 865
drug abuse detected using, 586–587, 587t
Median, 11
Medium-density arrays, 950
Medullary thyroid cancer, 356, 830–831
Melanin synthesis, copper and, 490
α-Melanocyte-stimulating hormone, 442t–444t
Melanoma, tumor markers for, 351
management of, 351
therapy prediction, 351
Melatonin, 442t–444t
biosynthesis of, 452f
Melting analysis, 953f, 954, 955f
Melting temperature (Tm), 948
Membrane filters, 117–118
Menadione, 474
Menaquinones, 474
Menarche, 844
Meningomyelocele, 873
Menkes syndrome, 490–491
Menopause, 846, 846f
Menses, 847–851
Menstrual cycle
endometrial changes during, 844f
hormones involved in, 844–846
ovarian changes during, 844, 844f
phases of, 845
Meperidine, 582f
Mephobarbital, 577t
Mercaptopurine (MP), for inflammatory bowel disease, 735
Mercury (Hg), 598–599, 989t–1025t
Metabolic acidosis, 663
anion gap, 693–694, 693f, 693b
bicarbonate for, 692–695
characteristics of, 692t
compensatory mechanisms in, 695
Metabolic alkalosis, 695–696, 695b
characteristics of, 692t
chloride-resistant, 695b, 696
conditions that cause, 693b
Metabolic bone disease, 769–771
biochemical measurements in, 746–752
Metabolic pathway, 884f

Metabolism
acetaminophen, 571
amino acid, disorders, 885–891, 892b
androgens, 788
bone. *see* Bone; metabolism; Lipoprotein(s); metabolism of
carbohydrate
disorders, 895
inborn errors, 383
by liver, 705–706
cortisol, 789
definition of, 221
of drugs, 545
estradiol, 842f
estrogen, 842f
inborn errors of. *see* Inborn errors of metabolism
progesterone, 843, 843f
steroid hormones, 788–789, 807f
testosterone, 835, 837f
thyroid hormone control of, 815
thyroxine, 815
triiodothyronine, 815
Metabolites, 898, 900f
measurement of, 232–233
Metabolizer classification, 898
Metabolomics, 213
Metagenomics, 967
Metal(s). *see also* Toxic elements.
aluminum, 592–593
antimony, 593
arsenic, 593–594
cadmium, 594–595
chromium, 595–596
cobalt, 596–597
copper. *see* Copper
iron. *see* Iron
lead, 597–598, 597f
mercury, 598–599
nickel, 599–600
thallium, 600–601
Metal electrodes, in redox reactions, 147–148
Metalloproteases, 299
Metanephrines, 454, 458
Methadone, 582f
Methamphetamine, 575, 577t
Methanol
description of, 570
metabolic acidosis caused by, 693
Methemoglobin (MetHb), 431, 502, 989t–1025t
Methemoglobinemia, 568t
acquired (toxic), 568
Methemoglobin-forming agents, 568
Methionine, 290t–291t, 292, 989t–1025t
Method comparison, study planning of, 18–19
Methodological approaches, 983–985, 984f
Methohexital, 577t
Methotrexate, 554
3-methoxy-4-hydroxyphenylglycol (MHPG), 453–454
Methylated ctDNA, 980
Methylmalonic acidemia, 887–888
Methylmercury poisoning, 599
Methylphenidate (Ritalin), 576–577
Methylresorufin, 324
Metrologic transferability, 111
Metyrapone stimulation test, 794, 794b
Micellar electrokinetic chromatography (MEKC), 178
Michaelis-Menten plots, 222–223, 223f
Micro RNA (miRNA), 923
Microalbuminuria, 362, 625
Microarrays, 950–951, 950f
protein, 247–248
Microbiome, alterations of, 967, 967t
Microcephaly, 888
Microchannel plate electron multipliers, 211
Microdeletion, 962
Microduplication, 962
β_2-Microglobulin, 302, 335t, 989t–1025t
Micropipettes, 112–113
Microsomal transfer protein, 403
Microtiter plate systems, 265–266
Micturition, 656
Midazolam, 578t
Migration rates, in electrophoresis, 175
Mineral metabolism, 744–772
Mineralocorticoids, 789
Minimally invasive glucose monitoring, 619
Minimum variance quadratic unbiased estimator (MIVQUE), 54
Miosis, 564–565
Missense mutation, 926
Mitochondria, hepatocyte, 702
Mitochondrial DNA, 926
Mitochondrial electron transport chain, 380
Mivacurium, 328
Mixed-bed resin exchanger, 110
Mixed-mode methods, 194
Mnemonics, 566
Mobile phase, 182, 191, 191b
reservoirs, 194–195, 195f
Mobile robots, for specimen delivery, 254
Model for End-Stage Liver Disease (MELD) Score, 720
Modified hemoglobins, 501–502, 501f
Modified target value, 17
Molar absorptivity, 130, 130t
Molar pregnancies, 869
Molecular applications, 961–987
Molecular biology
central dogma of, 919–921
essentials, 916, 916f–917f
Molecular detection, of respiratory viruses, 965
Molecular diagnostics, 270–271
clinical chemistry and, 1–6, 6b
definition of, 1
ethical issues in, 4b
Molecular diagnostics specialist, functions of, 4b
Molecular ion, 202
Molecular mass, of proteins, 297
Molecular principles, 914–935
Molecular sieving, 172
Molecular size, of proteins, 297
Molecular techniques, 936–960, 938f
detection techniques, 943
discrimination techniques, 943–959
Molecular testing, for respiratory viruses, 966
Molybdenum, 496–497, 989t–1025t
Monitoring, definition of, 57
Monoamine transporter, 453
Monoamines, 453
Monocarboxylate transporter 8 mutation, 815
Monoclonal gammopathy of undetermined significance (MGUS), 305
Monoclonal immunoglobulins, 305, 305t
Monoclonal light chains, 668–669
Monosaccharides, 378–379
Monounsaturated fatty acids, 394–395, 397f
Morphine, 582f
Morula, 863
Mother
pregnancy-related adaptations, 865–867
prenatal screening. *see* Prenatal screening
race of, 876
weight of, 876
Mucopolysaccharidosis type 1, 894
Müllerian duct agenesis, 847
Multianalyte desktop analyzers, 279
Multigene profiling, 344
Multiple endocrine neoplasia type 1 (MEN1), 730
Multiple of the median, 872, 873f
Multiple reaction monitoring (MRM), 211b
Multiple-channel analysis, 260–261
Multiplex analysis, 895
Multiplex ligation-dependent probe amplification (MLPA), 948
Multiplex polymerase chain reaction, 931–932, 941
Multiplexed immunoassays, 247–248
Multiplexed RT-PCR, 965
Multiplexing, 275
Muscle enzymes, 314–317
aldolase, 317
creatine kinase, 314–317, 314f
Mutation, 920
Mycophenolate mofetil (MMF), 558
Mycophenolic acid (MPA), 558
Myocardial infarction. *see* Acute myocardial infarction
Myocardial injury, 630
Myocardial ischemia, 631–632
Myocardium, 630
Myoglobin, 509
Myostatin, 442t–444t
Mysoline. *see* Primidone
Myxedema, 823
m/z ratio, 202

N

N-acetylcysteine (NAC), 571
N-acetyltransferase (NAT) polymorphism, 907
Na^+-H^+ exchange, 691
Nails specimen, 86
Naloxone, 582f
Naltrexone, 582f
Nanopore sequencing, 959
National Cholesterol Education Program (NCEP) guidelines, 416
Natriuretic peptides, 296
Natural opium alkaloids, 581
N-benzoyl-l-tyrosine-*p*-aminobenzoic acid, in exocrine pancreatic function, 738
Necrosis (cell death), 630, 706
Negative ion electron capture chemical ionization, 205
Neisseria gonorrhoeae, 964
Neonatal tumors, tumor markers for, 351
  management of, 351
Neonate. *see also* Newborn
  jaundice in, 519–521
Neoral. *see* Cyclosporine
Nephelometric assays, 240
Nephelometry, 143, 143f
  description of, 240, 267
  protein analysis using, 307
Nephritic syndromes, 667
Nephrogenic diabetes insipidus, 782–783
Nephrolithiasis, 670
Nephron, 653–655, 654f
Nephropathy
  analgesic, 670
  diabetic, 666
  hypertension, 666
  IgA, 667
  toxic, 670–671, 671t–672t
Nephrotic syndrome, 667
  management of, 667b
Nernst equation, 147
Nested polymerase chain reaction, 941
Neural tube defects
  anencephaly, 864–865, 873
  definition of, 873–874
  encephalocele, 873
  meningomyelocele, 873
  prenatal screening for, 874, 874f
  prevalence of, 873
  screening for, 868
Neuroblastoma, 459–460, 459b–460b
  tumor markers for, 351
Neuroblastoma Phox, 459
Neuroendocrine tumors
  gastroenteropancreatic, 460–461, 461b
  gastrointestinal, 738–739
Neuroleptics, 573
Neuron-specific enolase, 335t
Neuropathy, 591
  copper and, 491
Neuropeptide Y, 442t–444t
New York Heart Association (NYHA) classification system, 635, 635t
Newborn
  feces specimen collection from, 85
  hemolytic disease of the, 870–871
Newborn screening
  amino acid metabolism disorders, 885–891, 892b
  aminoacidopathies, 885–887, 891
  carbohydrate metabolism disorders, 895
  criteria for, 884–885
  fatty acid oxidation disorders, 892, 892b, 893f–894f
  galactosemia, 895
  glutaric acidemia type I, 887–888, 891
  homocystinuria, 889
  inborn errors of metabolism, 882–897
  lysosomal storage disorders, 892–894
  maple syrup urine disease, 889
  MCAD deficiency, 892
  mucopolysaccharidosis type 1, 894
  peroxisomal disorders, 894–895
  phenylketonuria, 887–889
  Pompe disease, 893–894
  recommendations for, 886t–887t
  second-tier testing, 885
  steps involved in, 885
  tandem mass spectrometry for, 895
  tyrosinemia, 889
  urea cycle disorders, 889–890, 891f
  X-linked adrenoleukodystrophy, 895
Niacin, 487–488, 487f, 989t–1025t
Nickel (Ni), 599–600, 989t–1025t
Nicotinamide adenine dinucleotide (NADH), 224
Nicotinamide-adenine dinucleotide phosphate (NADPH), 224
Ninhydrin approach, 294–295
NIST SRM 934, 117
Nitrogen-phosphorus detector (NPD), 189–190
Nitrosylation, of proteins, 299
Nonacute porphyrias, 532–533
Nonalcoholic fatty liver disease (NAFLD), 712–713, 715
Nonalcoholic steatohepatitis (NASH), 715
Nonbenzodiazepine sedative-hypnotics, 585, 585f
Noncommutable samples, external quality assessment and, 102
Noncompetitive immunoassays, 241–242
Noncompetitive inhibitors, 227–228, 227f
Nonfunctioning adrenocortical tumors, 802
Nonhuman genome, 925–926, 926f
Noninvasive glucose monitoring, 619
Noninvasive tests, of exocrine pancreatic function, 738
Nonparametric approach
  interpercentile intervals determined using, 68
  reference limits determined using, 69, 69t
Nonparametric regression analysis, 25
Nonsense mutation, 926
Nonspectral interferences, in atomic absorption spectrophotometry, 135
Non-ST segment elevation myocardial infarction (NSTEMI), 633
Nonthyroidal illness, 818, 819f, 819t
Nordoxepin, 572f
Norepinephrine, 442t–444t
  reference intervals for, 989t–1025t
  structure of, 452f
Normal distribution, 53–54
Normal-phase chromatography, 192
Normetanephrine, 454
Normovolemic hypernatremia, 684–685
Northern blotting, 171
Nortriptyline, 573
N-telopeptide, 768, 989t–1025t
N-terminal-truncated parathyroid hormone, 757–758
NT-proBNP-derived peptide products, 645f, 647f
  values and outcomes, 648f
Nuchal translucency (NT), 874
Nucleases, 920
Nucleic acid(s). *see also* Deoxyribonucleic acid (DNA); Ribonucleic acid (RNA)
  amplification techniques, 962
  preparation, 937–938
  structure of, 917–919
Nucleic acid amplification test (NAAT), 962
  advantages of, 965
Nucleosome, 916, 923, 924f
5'-nucleotidase, 323
Nucleotide, 917
Number of theoretical plates. *see* Plate number
Numerical chromosomal abnormalities, 974

O

αO_2, 430–431
Obesity, 605
Observed value, 71–72
Obstructive uropathy, 669
Occult blood, 85
Occupational Safety and Health Administration (OSHA), 118, 595
Odds ratio, 31
Off-label indications, challenge of, 933
O-glycosidic bond, 379
Ohm's law plot, 175–176
O'Leary method, 363
Oligohydramnios, 864–865
Oligomenorrhea, 847
Oligonucleotide, 938–939
Oligonucleotide ligation assay (OLA), 946, 947f
Oliguria, 656
Online sample concentration, in capillary electrophoresis, 177

Open-tubular columns, 188
Opioids/opiates
antagonists, 583
codeine, 581–583
definition of, 581–583
heroin, 582f
hydromorphone, 582f, 583t
mixed agonists/antagonists, 583
morphine, 581, 582f
oxymorphone, 582f
from poppy seed ingestion, 581
propoxyphene, 582f
types of, 581–583
Optical chemical sensors, 158–159, 158f
applications of, 159
basic concepts of, 158–159
Optical detection, enzyme-based biosensors with, 162–163
Optical methods, scope of, 129t
Optical path length, in capillary electrophoresis, 177
Optical techniques, 127–144
Optode, 158
Oral glucose tolerance test, 613–614
Ordinary least-squares regression (OLR) analysis, 20–21, 21f–22f
computation procedures for, 21–22
Organic acidemia, 885–887, 891
Organification, 814
Organophosphate insecticides, 574
Orotic acid, 989t–1025t
OSHA. *see* Occupational Safety and Health Administration
Osmol gap, 566
Osmolal gap, 428, 566
Osmolality, 1026t–1027t
plasma, 428
urine, 428
Osmolar gap, 1026t–1027t
Osmometry, 427
Osmosis, 427
Osmotic pressure, 427, 429b
Osteoblasts, 745, 746f
Osteocalcin (OC), 442t–444t, 761, 767, 989t–1025t
Osteoclasts, 745
Osteogenesis imperfecta (OI), 771
Osteomalacia, 592, 770
Osteonectin, 442t–444t
Osteopenia, 769
Osteopontin, 442t–444t
Osteoporosis, 769–770
Osteoprotegerin (OPG), 765
Ostwald-Folin pipettes, 112, 112f
Outliers, 53, 53f
assessment of, 22, 23f
handling of, 67–68
identification of, 67–68
Ovarian cancer
diagnosis of, 352
monitoring of, 353
prognosis of, 352–353
Ovarian cancer *(Continued)*
screening for, 352
tumor markers for, 352
management of, 352
Ovaries, 839
Ovulation, 854, 863
Oxalic acid, 989t–1025t
Oxazepam, 578t
β-Oxidation, 395
Oxycodone, 582f, 583t
Oxygen
in blood, 431–432
partial pressure of, 989t–1027t
determination of, 433–437
reference intervals for, 437
schematic diagram of, 690f
temperature correction of measured, 437, 437t
Oxygen dissociation curve, 434f
Oxygen saturation (SO_2), 432
Oxyhemoglobin, 431
Oxymorphone, 582f, 583t
Oxytocin, 442t–444t, 774, 989t–1025t

P

P_{50}, 433
PAGE-IEF, 178
Paget disease, 320
of bone, 770–771
Pain management, 586
Pancreas
anatomy of, 724, 724f
disease of, 736–742
enzymes, 323–327
amylase, 325–327, 326b
lipase, 323–325, 323f
exocrine function of, 736–742
adult disorders of, 737, 737b
pediatric disorders of, 736–737, 736b
tests of, 737–738
function of, 723–743
Pancreatic cancer
CA19-9 in, 353–354
monitoring of, CA19-9 in, 353
prognosis and therapy prediction for, CA19-9 in, 353
screening and diagnosis of, CA19-9 in, 353
tumor markers for, 353
management of, 353
Pancreatic chymotrypsin, in exocrine pancreatic function, 738
Pancreatic elastase-1, in exocrine pancreatic function, 738
Pancreatitis
acute, 323, 324f, 741–742, 742b
chronic, 737
Pancreolauryl test, in exocrine pancreatic function, 738
Pandemic plan, 120
Pantothenic acid, 488
Paper chromatography, 198
Paracetamol (pyroglutamic acid), 694
Paracrine, 441
Parafollicular cells, 812
Paraganglioma, 457–459, 457b–458b
Paraldehyde, 693b
Parallel analysis, 261
Parameters, 11
Parametric approach
interpercentile intervals determined using, 68
reference limits determined using, 69–71
Paraproteins, 305, 305t, 334t, 339t–340t
Parathyroid hormone (PTH), 442t–444t, 759b, 989t–1025t
biochemistry and physiology of, 756–757
biological actions of, 757
in bone and mineral metabolism, 756–759
calcium regulation via, 747f
circulating heterogeneity of, 757
clinical significance of, 757
definition of, 756
in end-stage renal disease patients, 759
measurement of, 757–758
methods, 757–758
pregnancy-related changes, 866–867
reference intervals for, 758
results, interpretation of, 758–759, 759f
synthesis and secretion of, 756–757
Parathyroid hormone-related peptide (PTHrP), 748
Parathyroid hormone-related protein, 764–765, 765f
Partial pressure, 429
Partition chromatography, 191–192
Partitioning
of reference group, 66t
of reference values, 67
of sample, in lipemia, 42
Passing-Bablok procedure, 25
Passive transport, 682
Path length, 130t
Pediatric tumors, tumor markers for, 351
management of, 351
Pellagra, 487–488
Pentobarbital, 577t
Peptide(s), 288–312, 311b. *see also specific peptide*
clinically relevant, 295–297
heterogeneity and analysis of, 295
Peptide bond, 295, 295f, 306
Perchloric acid, hazards in, 123
Percutaneous coronary intervention (PCI), 633
Pericardial fluid, specimen collection, 85
Pericardium, 631–632
Perimenopausal, 846
Peripheral dopaminergic system, 456–457, 456f
Peripheral tissues, 690–691
Peritoneal dialysis (PD), 673–674, 674f
Pernicious anemia, 481

Peroxisomal disorders, 894–895
Peroxisomes, 702
Personal protective equipment, 79, 121–122
pH, 989t–1027t
  blood gases and, 429
  buffer systems involved in regulation, 689–690
  definition of, 688
  enzyme-catalyzed reactions affected by, 225–226, 225f
  plasma, 689
Pharmacodynamics, 542–543, 543f, 898–899
Pharmacogene, 898–899
  associations, 903–911
    with pharmacodynamics, 909–911
    with pharmacokinetics, 903–909
Pharmacogenetics, 898–913
  analytical strategies, 901–902
  clinical interpretation, 901t, 902–903, 903t
  drug response, 899–900
  gene dose, 902
  haplotyping, 902
  implementation and logistics of, 901–903, 901t
  principles, 899–903
  reporting, 902
  specimens, 901
Pharmacogenomics, 899
Pharmacokinetics, 543, 543f, 899
Pharmacology, 567
Phenobarbital, 551, 551t, 577t
Phenotype, 899, 900b, 901–902, 928
Phenylalanine, 290t–291t, 888f, 989t–1025t
Phenylethylamine, 575, 576t
Phenylketonuria, 887–889
Phenylpiperazine, 576t
Phenytoin, 548–549, 548f, 551t
  free, 548
Pheochromocytoma, 457–459, 457b–458b
Phosphatase, 989t–1025t
Phosphate, 752–754, 754b
  biochemistry of, 752–753
  clinical significance of, 753
  critical risk value, 1026t–1027t
  interferences in, 754
  measurement of, 753–754
  physiology of, 752–753
  reference intervals, 754, 989t–1025t
  regulation of, 753
  specimen requirements in, 754
Phosphate interference, 135
Phosphatidylglycerol, 880
Phosphoglycerides, 399, 400f
Phospholipids, 392
Phosphor immunoassay, 246
Phosphorescence intensity, 159
Phosphorylation, of proteins, 298
Photodecomposition, in fluorescence measurements, 141
Photodegradation, of specimen, 261
Photodetectors, in spectrophotometers, 133
Photodiode array, 133
Photodiodes, 133
Photometers, filter, 130–131
Photometry, 128–130
  basic concepts in, 129–130, 129t
  reflectance, 134
Photomultiplier tube (PMT), 133
Photons, 129
Phred score, 929b
Phylloquinone, 474
Physiological jaundice of the newborn, 519–520
Piccolo chemistry analyzer, 275
Picrate, concentration of, 364
Pilocarpine iontophoresis, 424
PIMA flow cytometer device, 280
Pipettes, 111–113, 262
  automatic, 112–113, 113f–114f
  measuring, 112, 112f
  micropipettes, 112–113
  Ostwald-Folin, 112, 112f
  semiautomatic, 112–113, 113f–114f
  technique for, 112
  transfer, 112, 112f
    volumetric, 112, 112f
Pipetting stations, automated, 266
Pituitary gigantism, 778
Pituitary gland, 445, 834
  adenohypophysis. *see* Adenohypophysis
  anatomy of, 774–776, 774f
  definition of, 774, 775b
  disorders of, 773–786
  hypothalamic regulation, 776–784, 776t
  neurohypophysis, 774
Pituitary hormones
  adenohypophysis, 774–776
  adrenocorticotropic hormone. *see* Adrenocorticotropic hormone
  description of, 775t, 775b
  growth hormone. *see* Growth hormone
  overproduction of, 784t
  prolactin, 774, 779–780, 780b
  underproduction of, 784t
pK, 688
Placenta, 839–840
  definition of, 864
  estrogen production by, 864
  function of, 864, 865b
  hormones produced by, 864
Placental alkaline phosphatase, 335t
Placental alpha microglobulin-1, 879
Placental hormones, 864
Placental site trophoblastic tumors, tumor markers for, 349
Planar chromatography, 182, 198–199, 198f
Plaque, 633
  atherosclerotic, 633
Plasma, 681, 681t. *see also* Blood; Serum
  amino acids in, 292
  composition of, 79t
  definition of, 82–83
Plasma *(Continued)*
  description of, 681–682
  ethanol, 570
  gastrin, measurement of, 730
  metanephrines and, 463
  pH, 689
  porphyrins, analysis of, 536
  during pregnancy, 866t
  separation of, 87
  total carbon dioxide, 426, 426f
Plasma proteins, 299–300, 703–704. *see also specific protein*
Plasmid, 925
Plate height, 184
Plate number, 184
Platelet count, 506t
Platelet distribution width, 506t
Plateletcrit, 506t
Platinum, 147
Pleural fluid, specimen collection, 85
Pneumatic tube systems, for specimen delivery, 254
Pneumoconiosis, 592
Point of care, 567
Point-of-care testing, 273–287
  accountability for, 283
  accreditation and regulation of, 285
  advantages and disadvantages of, 274b
  audit for, 284–285
  certification for, 284
  chorionic gonadotropin, 877
  Clinical Laboratory Improvement Amendments of 1988 (CLIA), 285
  competency for, 284
  coordinating committee for, 283
  core outcome measures for, 285, 286t
  cost-effectiveness of, 285–286
  definition of, 273
  description of, 97, 104f
  devices for
    bar code identification systems in, 275
    classification of types of, 275t
    control and communications systems for, 276
    data management and storage in, 276
    design of, 275–276
    for glucose measurement, 277
    immunostrips as, 276–277
    larger cartridge type and bench-top, 278–280, 279f–280f
    manufacturing of, advances in, 273
    operator interface in, 275
    reaction cell in, 276
    required features of, 274–275
    sample types and sample delivery and, 275–276
    sensors in, 276
    single use, small handheld, 276–277
  documentation for, 285
  effectiveness of, 286b
    clinical, 285
    evidence of, 285–286

Point-of-care testing *(Continued)*
environments for, 274b
equipment for, procurement and evaluation of, 283–284
informatics and, 281–282
inventory control for, 285
maintenance of, 285
need for
assessing for, 283b
establishment of, 283
network for, 282, 282f
policy for, 283, 283b
quality assurance for, 284–285
quality control for, 284–285
quality management in, 283–285, 285b
technology for, 274–282
newer and emerging, 280–281, 281f
points to remember in, 282
types of existing, 276–280
training for, 284, 284b
Poisoning
lead, 597
mercury, 599
Polarization, in optical methods, 129t
Polyacrylamide gel, 172–173, 172f
Polyclonal antiserum, 238
Polyclonal hyperimmunoglobulinemia, 305
Polycystic kidney disease, 669
Polycystic ovary syndrome, 850, 850t
Polydipsia, 782
Polyglutamine disease, 298
Polyhydramnios, 864–865
Polymer membrane electrodes, 149–151, 149f–150f
Polymerase, 916
Polymerase chain reaction (PCR), 270
asymmetric, 941
automation, 271
contamination, inhibition, and control, 941
detection limits of, 941–942
digital, 942
for heat denaturation, 942–943
heparin contraindications to specimens for, 83
kinetics, 939–940, 940f
multiplex, 941
nested, 941
optimization, 940–941
process of, 938–939, 939f
product length, 944–945
selective amplification of sequence variants by, 941
-target amplification, 938–942
Polymerase mutants, 708
Polymodal distribution, 67
Polymorphic population, 218
Polypeptide hormones, 441
Polysaccharides, 380
Polyunsaturated fatty acids, 394–395, 397f
Polyuria, 656, 782
Pompe disease, 893–894
Poppy seed ingestion, 581
Population
definition of, 10
healthy reference, for determining 99th percentile, 639–640, 641t, 642f
probability distribution of, 11
Population frequency distribution, 10–11
Population mean, 11
Population standard deviation (SD), 11
Population variance, 11
Population-based reference values, 65–66, 72f
Porphobilinogen (PBG), 525, 535–536
Porphyria cutanea tarda (PCT), 529–532, 530t
Porphyrias, 524–540, 530t
acute neurovisceral attack of, 531t
asymptomatic relatives of patients with, 534–535
definition of, 525
external quality assessment for, 535
laboratory analysis of, 535–538
laboratory diagnosis of, 533–535
patients with symptoms of, 533–534
preanalytical aspects of, 535
quality management of, 535
specimen collection and stability of, 535
types of, 530t
Porphyrinogen, 525
Porphyrins, 524–540, 989t–1025t
biosynthetic pathway of, 526f
chelation of metals in, 525
chemistry of, 525–527
definition of, 525
diagnostics in, 536b
metabolism of, abnormalities of, 533, 533t
nomenclature of, 525
precursors, 525
spectral properties of, 527
structure of, 525, 525f
in urine and feces
analysis of, 536, 536f–537f
methods for, 535–536
Portal hypertension, 706
Porter-Silber chromogens, 807–808
Posterior probabilities, 31
Postpartum, thyroid disorders in, 828–829
Potassium, 421
concentrations, specimens for, 422
critical risk value, 1026t–1027t
deficit, 686–687
determination of, 422–423
dietary intake of, 421
diffusion of, 421
hemoglobin and, 422
homeostatic disorders of, 685–687
hyperkalemia, 687
hypokalemia, 686–687, 686f
intracellular distribution of, 686
ion-selective electrodes, 422–423
plasma, 420–421
Potassium *(Continued)*
redistribution of, 687
reference intervals for, 422, 989t–1025t
retention, 687
skeletal muscle activity and, 422
specimens for, 422
spectrophotometric methods for, 423, 423f
Potential difference, 146
Potentiometric detection, enzyme-based biosensors with, 162, 162f
Potentiometry, 146–148, 157b
Power supplies, for electrophoresis, 171
Prealbumin, 300–301
Preanalytical phase, 39–40
Preanalytical variation, 38–50
Preassigned values, quality control and, 96
Precipitin reaction, 238, 238f
Precision
description of, 14–15, 15f–16f
profile of, 15
Precocious puberty, 847
Precursor ions, 210
Predictive values, 31, 73–75, 74f–75f
Preeclampsia, 372, 869, 869b
Pregnancy, 861–881. *see also* Fetus; Mother; Placenta
alkaline phosphatase increases during, 320
amniotic fluid. *see* Amniotic fluid
assessments during, 868–869
assisted reproductive technologies for, 876
chorionic villus sampling, 867–868
complications of, 872b
cholestasis, 870
eclampsia, 869
ectopic pregnancy, 869
fatty liver, 870
HELLP syndrome, 869
hemolytic disease of the newborn, 870–871
hyperemesis gravidarum, 870
liver disease, 869–870
threatened abortion, 869
trophoblastic disease, 871
dating of, 868–869
diagnosis of, 868–869
ectopic, 869
estrogen levels during, 867
folic acid intake during, 868
laboratory test during
chorionic gonadotropin. *see* Chorionic gonadotropin; Human chorionic gonadotropin
description of, 868, 868b
fetal fibronectin, 879
α-fetoprotein, 877–878
inhibin A, 873f, 878–879
placental alpha microglobulin-1, 879
pregnancy-associated plasma protein A, 879, 879f
unconjugated estriol, 873f, 878

Pregnancy *(Continued)*
maternal adaptations during, 865–867
prenatal screening during. *see* Prenatal screening
thyroid disorders in, 828–829
Pregnancy-associated plasma protein A, 879, 879f
Premature rupture of membranes, 871
Prenatal screening, 861–881
adjustments for factors that affect, 876
chorionic villus sampling, 867–868
description of, 872–873, 876b
Down syndrome, 874
first trimester combined test, 874–875
maternal race effects on, 876
maternal weight effects on, 876
multiple of the median, 872, 873f
neural tube defects, 874, 874f
risk calculations, 872–873
second trimester, 875–876, 876f
trisomy 18, 874
Prenylation, of proteins, 298
Preproinsulin, 607
Preterm delivery, 871
Prevalence, 31
definition of, 75
Primary amenorrhea, 847
Primary biliary cholangitis, 719
Primary biliary cirrhosis (PBC), 703–704
Primary hyperaldosteronism, 802–803
Primary hypothyroidism, 823
Primary ovarian insufficiency, 848
Primary prostaglandins, 399
Primary reference materials, 111
Primary sclerosing cholangitis (PSC), 718–720
Primary structure, of proteins, 297
Primer, 938–939
Primidone, 551–552, 551t
Prisms, in spectrophotometers, 132
Probability, 10–11
for false rejection, 96
Probability distribution
description of, 10–11, 11f
of population, 11
Probe, 939, 949
ProBNP-derived peptide products, 644f
Procollagen type I carboxy-terminal propeptide (PICP), 765–766
Procollagen type I N-terminal propeptide (PINP), 765–766
Prodrug, 899, 900f
Product inhibition, 227
Product ions, 202, 210
Proficiency testing programs, 101–106, 106b
Progastrin-releasing peptide, 335t
Progesterone, 442t–444t
biosynthesis of, 842
chemistry of, 842–843, 842f
description of, 842
measurement of, 858–859
Progesterone *(Continued)*
metabolism of, 843, 843f
transport of, 843
Progesterone receptor, 336t
Proglucagon, 608
Prognostic interpretation, 932
Prograf. *see* Tacrolimus
Proinsulin, 607, 609–611, 989t–1025t
measurement of, 610
Prolactin, 334t, 442t–444t, 774, 779–780, 780b, 989t–1025t
Prolactin-releasing peptide, 442t–444t
Proline, 289, 290t–291t, 989t–1025t
Promoter, 916
Pro-opiomelanocortin (POMC), 295–296, 780
Propionic acidemia, 887–888
Proportional random error, regression analysis in cases of, 23
Propoxyphene, 582f
Prorenin, 808–809
Prostacyclin (PGI_2), 400, 658
Prostaglandins, 398f, 399–400, 399t
Prostanoic acid, 399
Prostate cancer
age-related reference intervals and, 354–356
management of, tumor markers for, 354
monitoring of, prostate-specific antigen in, 355–356
screening for, prostate-specific antigen in, 354–355
therapy prediction of, prostate-specific antigen in, 355
tumor markers for, 353–354
Prostate cancer gene 3 protein, 335t
Prostate marker algorithms, 335t
Prostate-specific antigen, 334t–335t, 339t–340t, 989t–1025t
density of, 354
percent-free, 354
for prostate cancer, 354
monitoring of, 355–356
screening for, 354–355
therapy prediction of, 355
velocity of, 354
Prosthetic groups, 229, 317
Proteases, 299, 300t
Protein hormones, 441
Protein microarrays, 247–248
Proteins, 288–312, 311b
analysis of, methods for, 305–307
γ-carboxylation of, 298–299
catabolism of, 299
in cerebrospinal fluid, 309–310
specific, assessment of, 309–310
critical risk value, 1026t–1027t
denaturation of, 217, 297
differential solubility of, 297
enzymes as, 217
fatty acylation of, 298
in fecal material, 310–311
Proteins *(Continued)*
folding of, 298
formation of, 298
glycated, 620–625
glycosylation of, 299
glycosylphosphatidylinositol anchor in, 298
in human serum and plasma, 299–300, 301t
hydroxylation of, 299
on iron homeostasis, 511
low molecular weight, 659
molecular mass, 297
molecular size of, 297
nitrosylation of, 299
in other body fluids, 309–311
in peritoneal and pleural fluids, 310
phosphorylation of, 298
physical properties of, 297–298
plasma. *see* Plasma proteins
posttranslational modifications of, 298–299
prenylation of, 298
primary structure of, 217, 297
quaternary structure of, 217, 297
reference intervals, 989t–1025t
secondary structure of, 217, 297
sulfation of, 299
synthesis, by liver, 703–705
targeting of, 298
tertiary structure of, 217, 297
total
albumin and, 360–363
in cerebrospinal fluid, 309, 310t
determination of, 305–307
measured concentrations of, variables affecting, 306–307
urinary. *see* Urinary proteins
Proteinuria, 360–363
albumin and, 362–363
clinical significance of, 362–363
consequences of, 659
definitions of, 362
detection of, 659
reference intervals for, 362
Proteome, 289
circulating, 299–300
abundant components of, 300–304
Proteomics, 213
Prothrombin time, 476
Proton pump inhibitors (PPIs) therapy, for *Helicobacter pylori*, 728
Protoporphyrin, 525
Protoporphyrinogen oxidase (PPOX), 528
Proximal convoluted tubule, 654
Pseudocholinesterase, 328
Pseudoephedrine, 575–576
Pseudoequilibrium, 240
Pseudogenes, 899, 901–902, 925
Pseudohermaphroditism, 846–847
Pseudohypoaldosteronism, 803
Pseudohypoaldosteronism type 1, 669–670
Pseudoporphyria, 533

Psychogenic polydipsia, 782
Puberty
in females, 844
in males, 835
precocious, 847
Public health program, for iodine disorders, 817
Publication of documents, issues with, 5
Publishers, responsibility of, 5
Pulmonary surfactant, 867, 871–872
Pure gonadal dysgenesis, 847
Purified water, quality, use, and storage of, 110, 110b
Purine, 916, 916f
Pyelonephritis, 668
Pyranose, 379
Pyridinium crosslinks, 768
Pyridinoline, 768
Pyridoxal-5'-phosphate, 317
Pyrimidine, 916, 916f
Pyrophosphorylase, 477
Pyrosequencing, 943–944, 957
Pyrrolidinophenone, 576t
Pyruvate, 382–383
in whole blood, measurement of, 387–388, 387f
Pyruvate carboxylase, 494
Pyruvic acid, 989t–1025t

Q

Quadrupole ion trap mass spectrometer, 209
Quadrupole ion traps, 209, 209f
Quadrupole mass filters, 207, 208f
Quadrupole mass spectrometers, 207
Quality assessment
external, 91, 101–106, 104b
commutable samples, 101–102
interpretation of, 102–106, 103f–105f, 106b
measurement procedure, 102
noncommutable samples, 102
proficiency testing programs, 101–106
Quality Assessment Tool for Diagnostic Accuracy Studies (QUADAS-2), 32
Quality control
internal, 91, 91b, 101b
materials for
frequency, 94–95
limitations of, 94, 94f
selection of, 94
measurement for, 91–93, 92f, 94b
analytical bias and imprecision, 91–93
performance of, 93, 93f
problem, 97–99
target value, 95–96
procedures, 94–101
results, 98f
review of, 100–101
rules, 96–97, 96t
specifying, 97
verifying, 99–100, 100f–101f
Quality management, 90–107
Quantitative assays, 626
Quantitative index tests, selection of cutoff value in case of, 30–31, 30f
Quantitative polymerase chain reaction (qPCR), 962
Quantum dot, 245–246
Quaternary structure, of proteins, 217, 297
Quazepam, 578t

R

Radiant energy, 129b
in analytical measurements, 142b
Radioimmunoassays, description of, 243
Radiometer AQT, 279
Raman scattering, 141
Random bias, 17–18
Random error, definition of, 9
Random sample, 11–12
Random-access analysis, 260
Rapamycin. *see* Sirolimus
Rapid antigen-based EIAs, 965
Rapid plasma reagin, 868
Rapidly progressive glomerulonephritis (RPGN), 668
Rasburicase, 908, 908t
Ratio-referencing spectrofluorometer, 138–139, 139f
Rayleigh scattering, 141
Reactants, mixing, 264
Reaction cell, 276
Reaction vessels, 264
Reactive hypoglycemia, 382
Readout devices, in spectrophotometers, 133
Reagent grade water, 109–110
preparation of, 110
Reagent-grade chemicals, 109–110
Reagents, 109
bar coding of, 263
chemical reaction phase of, 264
configuration of, 263
delivery of, 264
identification of, 263
management of, 263
mixing of, 264
storage of, 263
vendor-supplied *versus* user-defined, 263–264
Real-time PCR, 270, 939, 939f, 953–954, 955f
accuracy and precision, 954
detection, 953–954
quantification, 954, 955f, 955t
Recapper station, 257
Receiver operating characteristic (ROC) curves, 643, 646f
description of, 30, 30f, 32f
Receptor activator for nuclear factor kappa B (RANK), 765
Receptors, 441
Recognition element in, 276
Recombination, 916
Recorders, in spectrophotometers, 133
Recovery, 14
Red blood cell (RBC)
count, 506t
indices for, 518
Red cell distribution width, 506t
Redox couple, 146
Redox electrodes, 146–147
Reducing sugars, 379
5α-Reductase deficiency, 839
Reference change value, 57–60
definition of, 52
for more than two serial analyses, 58
Reference group partitioning, 66t
Reference individuals, 66, 72f
Reference information, for clinical laboratory, 988, 989t–1027t
Reference intervals, 989t–1025t
adrenocorticotropic hormone, 989t–1025t
alanine aminotransferase, 989t–1025t
albumin, 989t–1025t
aldosterone, 989t–1025t
antidiuretic hormone, 989t–1025t
α_1-antitrypsin, 989t–1025t
aspartate aminotransferase, 989t–1025t
bicarbonate, 426, 989t–1025t
bilirubin, 989t–1025t
calcitonin, 989t–1025t
calcium, 989t–1025t
catecholamines, 989t–1025t
ceruloplasmin, 989t–1025t
chloride, 424–426, 989t–1025t
cholinesterase, 989t–1025t
chorionic gonadotropin, 989t–1025t
copper, 989t–1025t
definition of, 68, 68b
dehydroepiandrostenedione sulfate, 989t–1025t
dehydroepiandrosterone, 989t–1025t
estradiol, 989t–1025t
ethanol, 989t–1025t
α-fetoprotein, 989t–1025t
fructosamine, 624, 989t–1025t
glomerular filtration rate, 989t–1025t
glycated hemoglobin, 624, 989t–1025t
growth hormone, 989t–1025t
high-density lipoprotein cholesterol, 989t–1025t
homocysteine, 989t–1025t
iron, 989t–1025t
lactate dehydrogenase, 989t–1025t
low-density lipoprotein cholesterol, 989t–1025t
magnesium, 989t–1025t
nonparametric determination of, 70t
partial pressure of oxygen, 989t–1025t
pH, 989t–1025t
phosphate, 989t–1025t
porphyrins, 529t, 989t–1025t
potassium, 422, 989t–1025t

Reference intervals *(Continued)*
prostate-specific antigen, 989t–1025t
protein, 989t–1025t
proteinuria and albuminuria, 362
pyridoxine, 989t–1025t
retinol-binding protein, 989t–1025t
riboflavin, 989t–1025t
serotonin, 989t–1025t
sodium, 421, 989t–1025t
sweat chloride, 425, 989t–1025t
testosterone, 989t–1025t
thyroxine, 989t–1025t
thyroxine-binding globulin, 989t–1025t
triiodothyronine, 989t–1025t
troponins, 989t–1025t
uric acid, 989t–1025t
vitamin A, 989t–1025t
vitamin B_{12}, 989t–1025t
vitamin E, 989t–1025t
vitamin K, 989t–1025t
Reference limits, 68–71, 73b
nonparametric, 69t
parametric, 70
Reference materials, 111
certified, 27–28
Reference measurement procedure, 26–27
Reference range, 68
Reference values, 64–76
analytical procedures, 66–67
background, 65–66
definition of, 65
establishment of, 65–71
health-associated, 66t
observed value and, 71–72
partitioning of, 67
population-based, 65–66, 72f
quality control, 66–67
specimen collection, 66
statistical treatment of, 67–71
subject-based, 65–66, 72, 72f
transferability of, 72–73
Reflectance photometry, 134
Refractive index detector, in liquid chromatography, 197, 197t
Refractometry, for total protein measurement, 306
Regression analysis, 19–26, 20f
application of, 25–26
in cases of proportional random error, 23, 24f
correlation coefficient, 23, 24f
definition of, 16
Deming, 20–22
computation procedures for, 21–22
estimated regression line, evaluation of random error around, 22
linearity testing for, 23–25, 24f
nonparametric, 25
ordinary least-squares, 20–22, 21f–22f
computation procedures for, 21–22
systematic differences between methods obtained on, interpretation of, 25
Relative centrifugal force (RCF), 115
Relaxin, 442t–444t
Reliability coefficient, calculation of, 60
Renal calculi, 670
Renal diseases. *see* Kidney diseases
Renal failure
description of, 660
uremia of, 694
Renal osteodystrophy, disorders of bone and mineral in, 770
Renal replacement therapy (RRT), 659, 671–677
kidney transplantation, 674
Renal tubular acidoses, 669, 695
Renin, 296, 658, 789–790
measurement of, 808–809
Renin activity, 808
Renin-angiotensin-aldosterone system, 442t–444t
Repeatability, 14
Replication, 919–920, 919f
Reproducibility, 15
Reproduction-related disorders, 834–860
Reproductive system. *see* Female reproductive system; Male reproductive system
Resistin, 442t–444t
Resolution, in chromatography, 184–185, 185f, 185b
equation, 185
Respiratory acidosis, 692t, 696, 696b–697b
compensatory mechanisms in, 696
Respiratory alkalosis, 692t, 696–697, 697b
characteristics of, 692t
compensatory mechanisms in, 695b, 696–697
in salicylate overdosage, 571–572
Respiratory distress syndrome, 871
Respiratory tract infections, 964–966
Respiratory virus panels, parameters of, 965, 966t
Restricted maximum likelihood (REML), 54
Restriction fragment length polymorphism (RFLP), 945, 945f
Retardation factor, 198
Retention factor, 183
Retention time, 182
Retention volume, 182
Retinol-binding protein (RBP), 989t–1025t
Reverse cholesterol transport pathway, 405, 406f
Reverse immunoblot assays (RIBAs), for HCV genome, 710
Reverse osmosis, 110
Reverse triiodothyronine, 821
Reversed-phase chromatography, 192
Reversed-phase liquid chromatography (RPLC), 192
Reversible fluorescent terminators, 957–958
Reversible inhibition, 227–228, 227f
Reviewers, responsibility of, 5
Reye syndrome, 705–706
Rheumatoid factor (RF), 989t–1025t
Riboflavin (vitamin B_2), 989t–1025t
Ribonucleic acid (RNA), 916
composition and structure of, 918–919
protein productions and, 919
sequencing of solid tumor isolates, 932
transcription and splicing, 920–921
Ribosome, 916
Ribozyme, 919
Rickets, 770
Risk stratification, of CHF patient, 647
Rolling circle amplification (RCA), 943, 959
Rumack-Matthew nomogram, 571, 571f

S

S-100 proteins, 335t
Safety, 118–125, 125b
equipment, 118–119
ergonomics program for, 120
inspections of, 119
program, 118
Salicylates, 571–572
Saliva
drug abuse detected using, 587
proteins in, 309
steroid hormones and, 804
Saliva specimen, 45–46
advantages and challenges of, 46b
patient preparation for sampling of, 46
stimulated, 45–46
types of, 45–46
unstimulated whole, 46
Salivary sex steroids, measurement of, 859
Sample, 10
carryover of, 262–263
continuous-flow analyzers, 262
discrete processing systems for, 262
internal transport, 262–263
introduction, 262–263
pretreatment, 263
Sample matrix effects, in fluorescence measurements, 141
Sample-related random bias, 17
Sandimmune. *see* Cyclosporine
Sandwich immunoassays, 246–247
Sanger sequencing, 939, 945–946, 945f–947f
Saturated fatty acids, 395, 397f, 399
Scarring, patients with, 534
Scattering, in optical methods, 129t
Screening tests
anion gap, 565–566
description of, 563
electrocardiogram, 565
immunoassay, 566–567
newborn. *see* Newborn screening
osmol gap, 566
prenatal. *see* Prenatal screening
spot tests, 565
Scurvy, 486
Secobarbital, 577t
Second messenger, 441

Secondary amenorrhea, 847–849
Secondary hyperaldosteronism, 802
Secondary reference materials, 111
Secondary structure, of proteins, 217, 297
Secretin, 442t–444t, 739t
Secretin-cholecystokinin test, of exocrine pancreatic function, 737, 737f
Sedative-hypnotic drug, 578
Selected ion mode, 207
Selected ion monitoring (SIM), 202–203
Selective glucose unresponsiveness, 612
Selenium, 491–492, 989t–1025t
Selenophosphate synthetase, 491
Self-monitoring of blood glucose, 617–618
Self-probing primers, 953, 954f
Semiautomatic pipettes, 112–113, 113f–114f
Semiconductor sequencing, 957
Semiconservative replication, 919
Semiquantitative assays, 626
Semisynthetic opiates, 583t
Sensitivity, 28–32, 29t
  clinical, 73–75, 74f
Sensors
  biosensors. *see* Biosensors
  optical chemical. *see* Optical chemical sensors
Separation factor, 184
Sequencing automation, 271
Sequential analysis, 260–261
Serial monitoring, 337–338
  results, 26
Serine, 290t–291t
Serine protease inhibitor Kazal type 1 *(SPINK1)* gene, 736–737
Serine proteases, 217, 299
Serotonin (5-hydroxytryptamine), 451
  biosynthesis of, 452f
  physiology of, 454–464
  structure of, 452f
Sertoli cells, 834–835
Serum. *see also* Blood; Plasma
  chorionic gonadotropin, measurements, 869
  composition of, 79t
  definition of, 82–83
  ethanol, 570
  α-fetoprotein levels in, 878
  25-hydroxyvitamin D concentrations, 761t
  during pregnancy, 866, 866t
  separation of, 87
  total carbon dioxide, 426, 426f
Serum analysis, 768
Serum cholinesterase, 328
Serum protein analysis, in capillary electrophoresis, 176, 176f
Serum protein electrophoresis, 307, 307f–308f
Severe combined immune deficiency (SCID), 304
Sex hormone-binding globulin (SHBG), 791, 835, 989t–1025t
Sexually transmitted infections, 964
Shock liver, 712
Short tandem repeat (STR), 924–925
Sick euthyroid syndrome, 818
Sickle cell anemia, 505
Sickling solubility tests, 507–508, 509f
Sideroblastic anemia, congenital, 511t–512t
Siderophilin. *see* Transferrin
Sigma metric, 93
Signal amplification, 938, 943
Signal transduction pathways, 976–977
Simple sequence repeat (SSR), 924–925
Simultaneous multianalyte immunoassays, 247–248
Simultaneous pancreas-kidney transplantation, 677
Simvastatin, 909
Single nucleotide extension, 946
Single nucleotide polymorphism (SNP), 923–924
Single nucleotide variation (SNV), 926–927
Single-beam spectrophotometer, 131, 131f
Single-bed deionizer, 110
Single-channel analysis, 260–261
Single-copy visualization, 951, 952f
Single-molecule sequencing, 958–959
Single-pan balance, 116
Sirolimus, 557
Sitosterolemia, 409
Six Sigma, 93, 93f
Size-exclusion chromatography (SEC), 192–193, 193f
Size-selection, 962, 985
Skeletal muscle activity, K^+ and, 422
Skin puncture
  blood specimen collection using, 81–82, 82f, 82t
  sites for, 81–82, 82f
Slab gel electrophoresis, 169
Small bowel tissue transglutaminase, 730
Small intestine. *see also* Intestinal disorders
  anatomy of, 724
  bacterial breath tests, 733
  bacterial overgrowth, 733
Small sequence variants, nomenclature for naming, 927b
Smooth endoplasmic reticulum, hepatocyte, 702
Sodium, 421–422
  critical risk value, 1026t–1027t
  determination of, 422–423
  homeostatic disorders of
    description of, 682–685
    hypernatremia, 684–685
    hyponatremia. *see* Hyponatremia
  ion-selective electrodes for, 422–423
  reference intervals for, 989t–1025t
Sodium *(Continued)*
  specimen of, 421
  spectrophotometric methods for, 423, 423f
  urinary, 421
    loss, 803
Sodium citrate, 83
Sodium fluoride, 83
Sodium iodide symporter, 817
Software computers, in spectrophotometers, 133
Solarization, 132–133
Solid phase adsorption, 242
Solid tissue specimen, 86
Solid tumor genomics, sampling and preservation methods for, 929–930
Solid-phase hybridization, 949–951
Soluble serum transferrin receptor (sTfR), 512
  method for, 517–518
Solute carrier organic anion transporter family member 1B1, 909
Solute volatilization interference, 135
Solution
  colligative properties of, 427
  procedures for processing, 117–118
  vapor pressure of, 427
Solution-phase hybridization, 951–953, 953f
Solvent effects, in fluorescence measurements, 141
Solvent programming, 194–195, 195f
Somatostatin, 609
Somatotropin, 442t–444t
Sorter, 257
Southern blotting, 171, 270
Specificity, 28–32, 29t
  clinical, 73–75, 74f
Specimen
  for amino acid analysis, 294–295
  amniotic fluid, 86
  ascitic fluid, 85
  automation applications, 252–259
    delivery, 254
    identification of, 252–254, 253b
    infectious disease transmission concerns, 262
    introduction, 262–263
    loading, 261
    preparation, 254
    processing, 254–256
    retrieval, 259
    sampling, 261–262
    storage, 257, 259
    transport, 262–263
  bar coding of, 253
  bicarbonate, 426
  blood. *see* Blood specimen
  buccal cells, 86
  cerebrospinal fluid, 85
  chloride, 424
  chorionic villus sampling, 86

Specimen *(Continued)*
closed-container sampling systems for, 262
collection of, 77–89
description of, 66
degradation of, 261
delivery of, 254
electrolyte, 420–421
feces, 85
hair, 86
handling of, 86
for hormone measurement, 448
identification of
automation of, 252–254
bar coding of, 253
description of, 86, 88b
errors in, 253–254
input area, 257
integrity, 257
labeling of, 253
loading of, 261
nails, 86
pericardial fluid, 85
photodegradation of, 261
pleural fluid, 85
potassium, 422
preparation of, 254
preservation of, 87
processing of, automated, 254–256
retrieval of, 259
separation of, 87
sodium, 421
solid tissue, 86
stool, 85
storage of, 87, 259
synovial fluid, 85
trace elements, 489
transport of, 87–88
to conveyor belt, 257–259
urine, 45b, 84–85
catheter-acquired, 84
from children, 85, 85f
clean timed, 84–85
collection of, 584
drug testing, 584
first morning, 45
first void, 45
midstream, 45
preservatives added to, 85
refrigeration of, 85
timed, 45, 84–85
transport and storage of, 45
types of, 44–45, 45t
Specimen validity testing, 584
Spectral bandwidth, 132
Spectral interferences, in atomic absorption spectrophotometry, 134–135
Spectrofluorometers, 138–140
Spectrophotometers, 128–131
atomic absorption, 134–135
components of, 131–133
cuvets in, 133
digital hardware in, 133
Spectrophotometers *(Continued)*
double-beam, 131, 131f
fiber optics in, 132–133
filters in, 132
gratings in, 132
light sources in, 131
performance parameters in, 133
photodetectors in, 133
prisms in, 132
readout devices in, 133
recorders in, 133
reflectance photometry in, 134
single-beam, 131, 131f
software computers in, 133
spectral isolation in, 132
wavelength isolation device in, 132
Spectrophotometric interference
of hemolysis, 41
of lipemia, 42
Spectrophotometry, 128–130
atomic absorption. *see* Atomic absorption spectrophotometry
basic concepts in, 129–130, 129t
instrumentation of, 130–133
nomenclature, 130t
Sphingolipids, 399, 399f
Sphingomyelin, 867
Sphingosine, 399
Spina bifida, 864–865
Spliceosome, 921
Spline function, 243
Spot tests, 565
Squamous cell carcinoma antigen, 335t
cervical carcinoma and, 345
ST segment elevation myocardial infarction (STEMI), 633
Standard deviation, 9, 11
intervals, 92
quality control and, 95–96, 95f
Standard Gaussian variable, 12
Standard Precautions, 121
Standard reference materials (SRMs), 111
Standard uncertainty, 27
Standards for Reporting of Diagnostic Accuracy Studies (STARD), 32, 33b
Star (*)-allele nomenclature, 899, 902, 904t
Starch, 380
Stationary phase, 182
Statistics, 9–13
categorical variables, 13
definition of, 11
frequency distribution, 9–10, 10f
Gaussian probability distribution, 12, 12f
nonparametric, 13
parameters, 11
population, 10
probability, 10–11
probability distributions, 10–11, 11f
random sampling, 11–12
sample, 10–11
student *t* probability distribution, 12–13, 12f
Steady-state voltammetry, 154
Steatorrhea, 733
Stereoisomers, 378, 378f
Steroid hormones, 441
androgens. *see* Androgen(s)
chemical structure of, 788
circulating forms of, 789
glucocorticoids. *see* Glucocorticoids
mineralocorticoids. *see* Mineralocorticoids
placental production of, 864
Steroidogenic acute regulatory protein, 802
Stimulation tests
adrenocorticotropic hormone, 793, 794b
cosyntropin, 793b
metyrapone, 794, 794b
Stokes shift, 135
Stomach
anatomy of, 724, 724f
diseases, 728–730
gastric acid secretion and gastrinomas, 729–730
Helicobacter pylori in, 728–729, 728b–729b
laboratory and, 728–730
lavatory investigations in, 728–730
Stoppers, for evacuated blood tubes, 80, 88
Strand displacement amplification (SDA), 942
Stress, 80
Strip tests, 276
Strips
glucose, 277, 277f
immunostrips as, 276–277
Strong acids, hazards in, 123
Strong mobile phase, 191
Structural variations, 926
Student *t* probability distribution, 12–13, 12f
Subacute combined degeneration of the spinal cord, 482
Subclinical hyperthyroidism, 828
Subclinical hypothyroidism, 817
Subject-based reference values, 65–66, 72, 72f
"Submarine" techniques, 172
Substrate, 216
concentration of, 222–225, 223f
Succinyldicholine, 328
Sulfation, of proteins, 299
Sulfhemoglobin, 431, 502
Sulfur amino acid metabolism, 890f
Superoxide dismutase, 494
Support media, for electrophoresis, 172–173
Suppression tests, dexamethasone, 796–798, 798b–799b
Suprapubic aspiration, 45
Surface adsorption, 297–298
Surface enzyme-coupled receptors, 446–447
Surface plasmon resonance-based immunoassays, 240
Surfactant, 867, 871–872

Sweat
conductivity screening, 426
testing, for drug abuse detection, 588
Sweat chloride, 424–426
collection of, 424
quality of, 425
sufficient sample, 424–425
falsely elevated, 425
quantitative analysis of, 425–426
Sympathetic nervous system, 455–456, 455f
Sympathomimetic drugs, 575
Syndrome of inappropriate antidiuretic hormone (SIADH), 684, 783–784, 784t
Synovial fluid specimen collection, 85
Systematic error, 13–14
Systemic inflammatory response syndrome, 467
Systemic iron regulation, 509–511, 510f
Systemic lupus erythematosus (SLE), 668
Systole, 632–633

T

T_3. *see* Triiodothyronine
T_4. *see* Thyroxine
Tacrolimus, 555, 556t, 907
Tamm-Horsfall glycoprotein, 659
Tandem mass spectrometers, 202, 209–211
Tandem mass spectrometry
definition of, 884
newborn screening uses of, 895
Tangier disease, 409
Target amplification, 938
polymerase chain reaction-, 938–942
Target value, 17
quality control and, 95
Taring, 116
Tartrate-resistant 5b isoform, 329
Tartrate-resistant acid phosphatase, 768–769
Task-targeted automation systems, 252, 254
Tegretol. *see* Carbamazepine
Telomere, 916
Telopeptides, 768
Temazepam, 578t
Temperature
in capillary electrophoresis, 175–176
creatinine and, 364
in fluorescence measurements, 141
in gas chromatography, 188
in liquid chromatography, 196
programming, 188
Terpenes, 400
Tertiary structure, of proteins, 217, 297
Testes
endocrine control of, 836f
function of, 834–835
Testicular cancer, tumor markers for, 356
Testicular feminization syndrome, 838–839
Testosterone, 442t–444t, 989t–1025t
biosynthesis of, 835
blood transport of, 835
Testosterone *(Continued)*
chemical structure of, 837f
concentrations, 835–838, 837f
free, 838, 857–858
measurement of, 856
metabolism of, 835
total, 838, 856–857
Tetrahydrocannabinol. *see* THC
α-Thalassemia minor, 503
Thalassemias, 502–504
alpha, 503, 503t
analytical methods for, 505–508
beta, 503–504, 504f
definition of, 503
major, 503
minor, 503
techniques for, 505–507
Thallium (Tl), 600–601, 989t–1025t
THC, psychoactive effects of, 579
Theophylline, 559–560
Therapeutic drug monitoring (TDM), 541–561, 675–677
analytical factors on, 547–548, 548b
clinical use of
general considerations for, 546–547, 546f
in specific drugs, 548–560
preanalytical factors on, 546–547, 547b
Therapeutic range, concept of, 547–548
Thermal regulation, of automated analyzers, 264
Thermometry, 117
Thiamylal, 577t
Thin-layer chromatography, 198
Thiopental, 577t
Thiopurine S-methyltransferase (TPMT), 908
Thiopurines, 908, 909t
analogs, for inflammatory bowel disease, 735
Thioredoxin reductases, 491
Thirst, 684–685
Threatened abortion, 869
Three-carbon carbohydrates, 378f
Threonine, 290t–291t, 989t–1025t
Thrombolysis, 633
Thromboxanes, 399, 400f
description of, 658
Throughput, definition of, 271
Thyrocalcitonin, 442t–444t
Thyroglobulin, 335t, 989t–1025t
analytical and reporting requirements for, 356
definition of, 812
measurement of, 822
thyroid cancer and, 356
Thyroid autoantibodies, 822
Thyroid cancer, tumor markers for, 356
management of, 356
monitoring of, 356
screening, diagnosis, and prognosis of, 356
Thyroid disorders, 811–833
in children, 831–832
hyperthyroidism, 828–829, 828b
hypothyroidism, 823–826, 824b–825b, 826t, 829, 831–832
in pregnancy and postpartum, 828–829
Thyroid dysgenesis, 812–813
Thyroid dysmorphogenesis, 812–813
Thyroid follicle, 812
Thyroid function
evaluation of, 819–820, 820f
extrathyroidal factors affect, 817–818, 818t
testing in pregnancy, 830
Thyroid gland
anatomy of, 812–813, 813f
description of, 812
drugs that affect, 818t
Thyroid hormone-binding protein, 816b
Thyroid hormones. *see also specific hormone*
biochemistry of, 813–815
biological function of, 813
formation of, 815f
free, measurement of, 821–822
genetics, 832
metabolism of, 815, 816f
mutations, 832
regulation of, 815–816
resistance, 828, 832
synthesis of, 814f
tests for, 813t
thyroxine. *see* Thyroxine
triiodothyronine. *see* Triiodothyronine
Thyroid morphogenesis, 812–813
Thyroid neoplasia, 830–831
Thyroid peroxidase, 822
Thyroid storm, 827
Thyroiditis, 822–823, 823t, 829–830
Thyroid-stimulating hormone (TSH)
characteristics of, 781–784
definition of, 815
genetics, 832
immunoassays for, 819
measurement of, 819–820
mutations, 832
receptor, 827
resistance to, 832
specimen collection and storage of, 819
Thyroperoxidase, 816
Thyrotoxicosis, 826–829, 827b
Thyrotrophs, 816
Thyrotropin, 815, 989t–1025t
Thyrotropin-releasing hormone (TRH), 442t–444t, 781
biochemistry of, 816
function of, 816
structure of, 816f
Thyroxine (T_4), 442t–444t, 989t–1027t
chemical structure of, 814f
definition of, 813
glucose metabolism affected by, 609
metabolism of, 815

Thyroxine (T_4) *(Continued)*
synthesis of, 815f
total, measurement of, 820
Thyroxine-binding globulin, 989t–1025t
definition of, 815
measurement of, 822
Thyroxine-binding prealbumin, 815
Time-of-flight mass spectrometers, 208, 208f
Time-resolved fluorometers, 136, 139
Tissue iron, 509
Tissue necrosis factor-α, 442t–444t
Tissue polypeptide antigen, 335t
Tobramycin, 552t
Topamax. *see* Topiramate
Topiramate, 550, 551t
Torr, 207, 430
Total calcium
adjusted or corrected, 749, 749b
interferences in, 749
interpretation of, 752
measurement of, 748–749
photometric methods for, 748–749
preanalytical errors in measurement of, 750b, 751
reference intervals for, 751, 989t–1025t
specimen requirements in, 749
spectrometry methods for, 749
Total cortisol, 805
reference intervals for, 989t–1025t
Total ion chromatogram (TIC), 202, 204f
Total iron-binding capacity, methods for, 516–517
Total magnesium
atomic absorption spectrometry for, 755
enzymatic methods for, 755
measurement of, 755
photometric methods for, 755
reference intervals for, 756, 989t–1025t
specimen requirements for, 755–756
Total parenteral nutrition, 467
Total proteins
albumin and, 360–363
in cerebrospinal fluid, 309, 310t
determination of, 305–307
measured concentrations of, variables affecting, 306–307
Total reducing substances, qualitative method for measurement of, 386
Total thyroxine, measurement of, 820
Tourniquet, 80
Toxic agents, 567
Toxic elements, 590–602, 591b. *see also* Metal(s)
content, 590
diagnosing, 591, 592b
periodic table with emphasis on, 591f
prevalence of, 590
specific elements of, 592–601
treatment of, 591–592
Toxic hepatitis, 712, 712f
Toxic multinodular goiter, 827
Toxic nephropathy, 670–671, 671t–672t
Toxic syndromes, 564–565
Toxicity, diagnosing, 591
Toxicokinetics, 592–593
Toxicologic interest
alcohols of, 568–571
ethanol, 568–570
isopropanol, 570
methanol, 570
Toxicology, 542, 563–564
Toxidrome, 564–565, 565t
TP53, 977
Trace elements, 466–498
analytical methods for, 489–497
copper, 490–491
inflammation and, 468t
intakes for adults, 468t
laboratory assessment for, 489
reference intervals for, 489t
Traceability
definition of, 9
description of, 26–28, 28b
Tramadol, 582f
Trans fatty acid, 395
Transcription, 917
Transcription-mediated amplification (TMA), 942, 942f
Transfer pipettes, 112, 112f
Transferrin, 302, 508–509, 989t–1025t
methods for, 516–517
saturation of, 512, 516–517
Transfusion iron overload, 516
Transitional cell carcinomas (TCCs), 343
Transketolase, 477
Translation, 917, 921, 921b, 922f
Translocations, recurrent, 974–976, 974f–975f
Transmittance (T), 129–130, 129f, 130t
Transplantation, liver, 720
Transport system, 257
Transthyretin, 300–301, 815, 989t–1025t. *see also* Prealbumin
synthesis, by liver, 703
Triazolam, 578t
Tricyclic antidepressants (TCAs), 905, 906t
pharmacological response of, 572–573
structure of, 572f
Triglycerides, 392–393, 398, 414–415, 989t–1025t
from plant sources, 398
Triiodothyronine (T_3), 442t–444t, 989t–1025t
chemical structure of, 814f
definition of, 813
description of, 813
measurement of, 821
metabolism of, 815
reverse, 821
synthesis of, 814, 815f
Trimesters, 863
Trimipramine, 572f
Trisomy 18, 874
Trophoblastic disease, 871
Trophoblasts, 863
Troponins, cardiac
I, 637–639, 637f, 643f
T, 637–639, 637f, 643f
True negatives, 29
True positives, 29, 29f
True value, 17
Trueness, 14, 14t
Trypsin inhibitors, 228
Tryptophan, 290t–291t, 294, 989t–1025t
Tryptophan hydroxylase, 452–453
Tuberculosis control plan, 120
Tuberoeruptive xanthomas, 409
Tumor(s)
adrenal, 802
functioning adrenocortical, 802
germ cell. *see* Germ cell tumors
hepatic, 720
neuroendocrine
gastroenteropancreatic, 460–461, 461b
gastrointestinal, 738–739
nonfunctioning adrenocortical, 802
Tumor cells, staining and selection of, 930
Tumor markers, 331–358, 342b, 357b
analytical requirements for, 338–342
assay validation for, 338
for bladder cancer, 343
for breast cancer, 344
for cancers of unknown primary, 356
carcinoembryonic antigen. *see* Carcinoembryonic antigen
for cervical cancer, 345
for childhood germ cell tumors, 351
for choriocarcinoma, 349
clinical applications of, 332t–333t, 335t, 337–338, 339t
assessing prognosis, 337
diagnosis of cancer, 337
monitoring disease post treatment to detect progression or recurrence, 337–338
monitoring response during and/or shortly after treatment, 337
prediction of treatment response, 337
screening for cancer, 337
clinical presentations and, 339t–340t
for colorectal cancer, 345
concentration of, nonmalignant conditions increasing, 340t–341t
definition of, 331
α-fetoprotein. *see* α-Fetoprotein
for gastric cancer, 346
for gastrointestinal neuroendocrine tumors, 738–739
for gastrointestinal stromal tumors, 346–347
for germ cell tumors, 347
for gestational trophoblastic disease, 348–349
half-lives of, 342
for hepatoblastoma, 351

Tumor markers *(Continued)*
for hepatocellular carcinoma, 349–350
high-dose "hooking" of, 341, 341f
historical background of, 333
for hydatidiform moles, 349
ideal, 336
interferences in, clinically relevant, 341–342
for invasive moles, 349
for lung cancer, 350
for melanoma, 351
for neonatal and pediatric tumors, 351
for neuroblastoma, 351
for ovarian cancer, 352
for pancreatic cancer, 353
for placental site trophoblastic tumors, 349
postanalytical requirements for, 342, 343f
preanalytical requirements for, 338
pre-preanalytical requirements for, 338
proficiency testing/external quality assessment in, 341
for prostate cancer, 353–354
quality control for
internal, 338–341
material for, 341
quality requirements for, 338–344
specimen for
carry-over of, 341
handling of, 338
timing of, 338
standardization of, 341
for testicular cancer, 356
for thyroid cancer, 356
use of
general principles of, 333–338
in specific malignancies, 343
validation of, 336
Tumor-specific ctDNA mutations, 979–980, 981f
Turbidimetric assays, 240
Turbidimetry, 142
description of, 240, 267
protein analysis using, 307
Turnaround time (TAT), 638
Turner syndrome, 847
Type I procollagen, 766f
Type III hyperlipoproteinemia, 409
Type V hyperlipoproteinemia, 408
Tyrosine, 290t–291t, 989t–1025t
Tyrosinemia, 889

U

Ubiquitin-proteasome system, 299
Ulcerative colitis (UC), 733–734
Ultrafiltration, 821–822
Ultra-high-performance liquid chromatography (UPLC), 191
Ultrapure reagents, 111, 111b
Ultraviolet (UV) light, absorbance of, 306
Ultraviolet oxidation, in reagent grade water, 110
Uncertainty, 27–28
direct assessment of, 27–28
indirect evaluation of, 28
standard, 27
Uncompetitive inhibition, 228
Unconjugated bilirubin, 519
Unconjugated estriol, 873f, 878
Unconjugated hyperbilirubinemia, 519–520, 520b
Unit operations, 260, 260b
United States Pharmacopeia (USP) standards, in therapeutic drug monitoring, 546
Unsaturated iron-binding capacity (UIBC), 516
Unstable angina, 630
Upregulation, 531
Urea, 359–376
analytical methods for, 369
back-diffusion of, 368–369
biochemistry of, 368–369
breath test, for *Helicobacter pylori*, 729
clinical significance of, 369
critical risk value, 1026t–1027t
cycle pathway of, 368f
reference intervals for, 369, 989t–1025t
Urea cycle disorders, 889–890, 891f
Urea nitrogen, 1026t–1027t
Urea synthesis, by liver, 704–705
Uremia, 663
of renal failure, 694
Uremic syndrome, 663–664
biochemical characteristics of, 664b
Uric acid (urate), 359–376, 373b
biochemistry and physiology of, 369–371, 370f
cardiometabolic outcomes for, 372–373
clinical significance of, 371
critical risk value, 1026t–1027t
reference intervals for, 371, 989t–1025t
renal handling of, 371
sample collection for, 371
Urinalysis, 660
automated analyzers for, 267–268
Urinary calcium, 752
Urinary Hg, 599
Urinary proteins, 363b
analytical methods and traceability of, 360–361, 361f
clinically important, 362t
sample collection for, 360
Urination, 656
Urine
analysis, 768
glucose in
measurement of, 386
strip test, 386
homovanillic acid, 463
ketone bodies in, 620
determination of, 388–389
17-ketosteroids in, 858
metanephrines and, 463
Urine *(Continued)*
osmolality, 428
porphyrins and porphyrin precursors in, 535–536
protein loss, 659
steroid hormones and, 804
vanillylmandelic acid, 463
Urine specimen, 45b, 84–85
collection of, 584
from children, 85, 85f
drug testing, 584
preservatives added to, 85
refrigeration of, 85
transport and storage of, 45
types of, 44–45, 45t
catheter-acquired, 84
first morning, 45
first void, 45
midstream, 45
suprapubic aspiration, 45
timed, 45, 84–85
Urine testing, influences and interferences in, 44–45
Urobilinogens, 519
Uroporphyrin, 529
Uroporphyrinogen decarboxylase, 528
Uroporphyrinogen-III synthase, 527–528

V

Valine, 290t–291t, 989t–1025t
Valinomycin, 422
Valproic acid, 549, 551t
van Deemter equation, 184
Van der Waals-London dipole-dipole interactions, 238
Vancomycin, 552t, 553–554
Vanillylmandelic acid (VMA), 452–454, 989t–1025t
Vapor pressure, 427
Vapor pressure osmometer, 428
Vaporization, hazards and, 124
Variant call format (VCF), 928
Variant databases, 927–928, 928b
Variants, naming of, 927
Varices, 715
Variegate porphyria (VP), 530t, 531
Vasoactive intestinal polypeptide (VIP), 738, 739t
Vasopressin, 297, 781–782, 781f
reference intervals for, 989t–1025t
Venipuncture, 77–89
illustration of, 79f
order of draw, 80, 81t–82t
patient identification verification before, 78
phlebotomist, 79
system for, 80
technique for, 79f, 81
tourniquets for, 80
Ventricles, 631–632
Very low-density lipoproteins, 401t–402t, 403–404

Viral ctDNA, 980–981
Viral hepatitis, 713b
during pregnancy, 870
Viral load testing, 963
Virilization, 850–851
Virologic failure, 963
Vitamers, 474
Vitamin, 467, 496b–497b
inflammation and, 467, 472t
Vitamin A, 468–473, 989t–1025t
absorption of, 468–472
deficiency of, 472–473
excretion of, 468–472
functions of, 472, 473f
laboratory assessment of, 473
metabolism of, 468–472
required by human, 469t–471t
structure of, 472f
toxicity of, 473
transport of, 468–472
Vitamin B_1, 469t–471t, 476–477, 989t–1025t
absorption of, 476
deficiency in, 477
excretion of, 476
functions of, 476–477
laboratory assessment for, 477
metabolism of, 476
structure of, 476f
toxicity of, 477
transport of, 476
Vitamin B_2, 469t–471t, 477–478, 477f, 989t–1025t
absorption of, 478
deficiency in, 478
excretion of, 478
functions of, 478
laboratory assessment of, 478
metabolism of, 478
toxicity of, 478
transport of, 478
Vitamin B_3, 469t–471t
Vitamin B_5, 469t–471t, 488
absorption of, 488
deficiency in, 488
excretion of, 488
functions of, 488
laboratory assessment of status in, 488
metabolism of, 488
toxicity in, 488
transport of, 488
Vitamin B_6, 469t–471t, 478–479, 479f
absorption of, 479, 479f
deficiency in, 479
excretion of, 479
functions of, 479
laboratory assessment for, 479
metabolism of, 479
toxicity of, 479
transport of, 479
Vitamin B_7, 469t–471t
Vitamin B_9, 469t–471t
Vitamin B_{12}, 469t–471t, 479–482, 480f, 989t–1025t
absorption of, 480
deficiency in, 481–482
excretion of, 480
functions of, 480–481
laboratory assessment for, 482
metabolism of, 480
toxicity of, 482
transport of, 480
Vitamin C, 469t–471t, 485–486, 485f, 989t–1025t
absorption of, 485
deficiency in, 486
excretion of, 485
functions of, 485–486
laboratory assessment for, 486
metabolism of, 485
toxicity of, 486
transport of, 485
Vitamin D, 473, 763b, 989t–1025t
assays for, 762
biochemistry and physiology of, 760–761
cholecalciferol, 760
clinical significance of, 761–762, 761t
deficiency of, 320, 770
description of, 747–748
1,25-dihydroxyvitamin D, 754t, 761
metabolism of, 760f, 761
metabolites of, 754t, 760–763
measurement of, 762
reference intervals, 762–763
regulation of, 761
sources, 762
transport of, 761
Vitamin D binding protein, 761
1,25(OH_2) Vitamin D_3, 658
Vitamin E, 469t–471t, 474, 474f, 989t–1025t
absorption of, 474
deficiency of, 474
excretion of, 474
functions of, 474, 475f
laboratory assessment for, 474
metabolism of, 474
toxicity of, 474
transport of, 474
Vitamin K, 469t–471t, 474–476, 989t–1025t
absorption of, 474–475
deficiency in, 475–476
excretion of, 474–475
forms of, 475f
functions of, 475
laboratory assessment for, 476
metabolism of, 474–475
toxicity of, 476
transport of, 474–475
Vitamin K epoxide reductase complex (VKOR), 909–910, 910f
Vitros analyzer, 387
Void time, 182
Void volume, 182
Volatile alcohols, analysis of, 571
Volatiles, hazards from, 123–124
Voltammetry, 153–157
applications of, 155–157, 156f–157f
basic concepts of, 153–155, 154f
Voltammogram, 153
Volume depletion effect, in lipemia, 42
Volumetric flasks, 113–114, 114f
Volumetric sampling, 111–114
Volumetric transfer pipette, 112, 112f
von Pettenkofer, Max Josef, 2

W

Warfarin, 900b, 910–911
Water, 680
body. *see* Body water
homeostasis of, 420
physiology and disorders of, 679–698
purity, testing for, 110
Waterhouse-Friderichsen syndrome, 805
Watery diarrhea hypokalemic achlorhydria syndrome (WDHA syndrome), 738–739
Wavelength, 129
creatinine and, 364
isolation device, in spectrophotometers, 132
unit, 130t
WBC differential count, 506t
Weak mobile phase, 191
Weighing, principles of, 116
Weighted analysis of means (WAM), 54
Werner-Morrison syndrome, 738–739
Wernicke encephalopathy, 477
Wernicke-Korsakoff syndrome, 477
Western blotting, 171, 307
HIV-1 antibody detection using, 239
White blood cell count (WBC), 506t
Whole blood, for analysis, 254
Whole genome sequencing, 931
Whole-exome sequencing, 973
Whole-genome amplification, 943
Wick flow, 169
Wide bandpass filters, 132
Wilson disease, 491, 591, 716–717
Within-subject variation (CV_I), 52
Workstations, 252
multifunction, 254–256, 255f, 256b
single-function, 254, 254b, 255f

X

Xenobiotics, 545
Xerophthalmia, 473
X-linked adrenoleukodystrophy, 895
X-linked diseases, 971–972
X-linked erythropoietic protoporphyria (XLEPP), 532–533

Z

Zaleplon, 585f
Zarontin. *see* Ethosuximide

Zero-order kinetics, 224
Zinc, 492–494, 989t–1025t
absorption of, 492
deficiency in, 493
excretion of, 492
functions of, 493
laboratory assessment for, 493–494
metabolism of, 492, 493f
Zinc *(Continued)*
toxicity of, 493
transport of, 492
Zinc protoporphyrin (ZPP), 528, 598
Zinc transporter ZnT8, 611
Zinc-induced hypocupremia, 490
Zollinger-Ellison syndrome, 730
Zolpidem, 585f
Zona fasciculata, 788–789
Zona glomerulosa, 788–789, 792f
Zona reticularis, 788–789
Zone electrophoresis, 168
Zonisamide, 551t, 552
Z-score, 103–104
Zwitterions, 291, 292f
Zygote, 863

This page intentionally left blank

IUPAC Periodic Table of the Elements

Key:

atomic number
Symbol
name
conventional atomic weight
standard atomic weight

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1 **H** hydrogen 1.008 [1.0078, 1.0082]																	2 **He** helium 4.0026
3 **Li** lithium 6.94 [6.938, 6.997]	4 **Be** beryllium 9.0122											5 **B** boron 10.81 [10.806, 10.821]	6 **C** carbon 12.011 [12.009, 12.012]	7 **N** nitrogen 14.007 [14.006, 14.008]	8 **O** oxygen 15.999 [15.999, 16.000]	9 **F** fluorine 18.998	10 **Ne** neon 20.180
11 **Na** sodium 22.990	12 **Mg** magnesium 24.305 [24.304, 24.307]											13 **Al** aluminium 26.982	14 **Si** silicon 28.085 [28.084, 28.086]	15 **P** phosphorus 30.974	16 **S** sulfur 32.06 [32.059, 32.076]	17 **Cl** chlorine 35.45 [35.446, 35.457]	18 **Ar** argon 39.948
19 **K** potassium 39.098	20 **Ca** calcium 40.078(4)	21 **Sc** scandium 44.956	22 **Ti** titanium 47.867	23 **V** vanadium 50.942	24 **Cr** chromium 51.996	25 **Mn** manganese 54.938	26 **Fe** iron 55.845(2)	27 **Co** cobalt 58.933	28 **Ni** nickel 58.693	29 **Cu** copper 63.546(3)	30 **Zn** zinc 65.38(2)	31 **Ga** gallium 69.723	32 **Ge** germanium 72.630(8)	33 **As** arsenic 74.922	34 **Se** selenium 78.971(8)	35 **Br** bromine 79.904 [79.901, 79.907]	36 **Kr** krypton 83.798(2)
37 **Rb** rubidium 85.468	38 **Sr** strontium 87.62	39 **Y** yttrium 88.906	40 **Zr** zirconium 91.224(2)	41 **Nb** niobium 92.906	42 **Mo** molybdenum 95.95	43 **Tc** technetium	44 **Ru** ruthenium 101.07(2)	45 **Rh** rhodium 102.91	46 **Pd** palladium 106.42	47 **Ag** silver 107.87	48 **Cd** cadmium 112.41	49 **In** indium 114.82	50 **Sn** tin 118.71	51 **Sb** antimony 121.76	52 **Te** tellurium 127.60(3)	53 **I** iodine 126.90	54 **Xe** xenon 131.29
55 **Cs** caesium 132.91	56 **Ba** barium 137.33	57-71 lanthanoids	72 **Hf** hafnium 178.49(2)	73 **Ta** tantalum 180.95	74 **W** tungsten 183.84	75 **Re** rhenium 186.21	76 **Os** osmium 190.23(3)	77 **Ir** iridium 192.22	78 **Pt** platinum 195.08	79 **Au** gold 196.97	80 **Hg** mercury 200.59	81 **Tl** thallium 204.38 [204.38, 204.39]	82 **Pb** lead 207.2	83 **Bi** bismuth 208.98	84 **Po** polonium	85 **At** astatine	86 **Rn** radon
87 **Fr** francium	88 **Ra** radium	89-103 actinoids	104 **Rf** rutherfordium	105 **Db** dubnium	106 **Sg** seaborgium	107 **Bh** bohrium	108 **Hs** hassium	109 **Mt** meitnerium	110 **Ds** darmstadtium	111 **Rg** roentgenium	112 **Cn** copernicium	113 **Nh** nihonium	114 **Fl** flerovium	115 **Mc** moscovium	116 **Lv** livermorium	117 **Ts** tennessine	118 **Og** oganesson

57 **La** lanthanum 138.91	58 **Ce** cerium 140.12	59 **Pr** praseodymium 140.91	60 **Nd** neodymium 144.24	61 **Pm** promethium	62 **Sm** samarium 150.36(2)	63 **Eu** europium 151.96	64 **Gd** gadolinium 157.25(3)	65 **Tb** terbium 158.93	66 **Dy** dysprosium 162.50	67 **Ho** holmium 164.93	68 **Er** erbium 167.26	69 **Tm** thulium 168.93	70 **Yb** ytterbium 173.05	71 **Lu** lutetium 174.97
89 **Ac** actinium	90 **Th** thorium 232.04	91 **Pa** protactinium 231.04	92 **U** uranium 238.03	93 **Np** neptunium	94 **Pu** plutonium	95 **Am** americium	96 **Cm** curium	97 **Bk** berkelium	98 **Cf** californium	99 **Es** einsteinium	100 **Fm** fermium	101 **Md** mendelevium	102 **No** nobelium	103 **Lr** lawrencium

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

For notes and updates to this table, see www.iupac.org. This version is dated 28 November 2016.

Copyright © 2016 IUPAC, the International Union of Pure and Applied Chemistry.