



3rd EDITION

# Users' Guides to the Medical Literature

ESSENTIALS OF EVIDENCE-BASED CLINICAL PRACTICE

Gordon Guyatt, MD  
Drummond Rennie, MD  
Maureen O. Meade, MD  
Deborah J. Cook, MD

**Mc  
Graw  
Hill**  
Education

**JAMA**evidence®

**Users' Guides**

—— to the ——

**Medical  
Literature**

#### Notice

Medicine is an ever-changing science. As new research and clinical experience broaden our knowledge, changes in treatment and drug therapy are required. The authors and the publisher of this work have checked with sources believed to be reliable in their efforts to provide information that is complete and generally in accord with the standards accepted at the time of publication. However, in view of the possibility of human error or changes in medical sciences, neither the authors nor the publisher nor any other party who has been involved in the preparation or publication of this work warrants that the information contained herein is in every respect accurate or complete, and they disclaim all responsibility for any errors or omissions or for the results obtained from use of the information contained in this work. Readers are encouraged to confirm the information contained herein with other sources. For example and in particular, readers are advised to check the product information sheet included in the package of each drug they plan to administer to be certain that the information contained in this work is accurate and that changes have not been made in the recommended dose or in the contraindications for administration. This recommendation is of particular importance in connection with new or infrequently used drugs.

# Users' Guides to the Medical Literature

## ESSENTIALS OF EVIDENCE-BASED CLINICAL PRACTICE

3rd EDITION

### Editors

**Gordon Guyatt, MD, MSc**  
Departments of Clinical Epidemiology &  
Biostatistics and Medicine  
Faculty of Health Sciences  
McMaster University  
Hamilton, Ontario, Canada

**Maureen O. Meade, MD, FRCPC, MSc**  
Departments of Medicine and Clinical  
Epidemiology & Biostatistics  
Faculty of Health Sciences  
McMaster University  
Hamilton, Ontario, Canada

**Drummond Rennie, MD**  
Former Deputy Editor, *JAMA*  
Chicago, Illinois  
Philip R. Lee Institute for Health Policy  
Studies  
University of California, San Francisco  
San Francisco, California, USA

**Deborah J. Cook, MD, MSc**  
Departments of Medicine and Clinical  
Epidemiology & Biostatistics  
Faculty of Health Sciences  
McMaster University  
Hamilton, Ontario, Canada



New York Chicago San Francisco Athens  
London Madrid Mexico City Milan  
New Delhi Singapore Sydney Toronto



The **JAMA** Network

**JAMA**evidence®  
Using Evidence to Improve Care

Copyright © 2015 American Medical Association. All rights reserved. Except as permitted under the United States Copyright Act of 1976, no part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written permission of the publisher, with the exception that the program listings may be entered, stored, and executed in a computer system, but they may not be reproduced for publication.

ISBN: 978-0-07-180872-9

MHID: 0-07-180872-8

The material in this eBook also appears in the print version of this title: ISBN: 978-0-07-179415-2, MHID: 0-07-183314-5.

eBook conversion by codeMantra  
Version 1.0

All trademarks are trademarks of their respective owners. Rather than put a trademark symbol after every occurrence of a trademarked name, we use names in an editorial fashion only, and to the benefit of the trademark owner, with no intention of infringement of the trademark. Where such designations appear in this book, they have been printed with initial caps.

McGraw-Hill Education eBooks are available at special quantity discounts to use as premiums and sales promotions or for use in corporate training programs. To contact a representative, please visit the Contact Us page at [www.mhprofessional.com](http://www.mhprofessional.com).

#### TERMS OF USE

This is a copyrighted work and McGraw-Hill Education and its licensors reserve all rights in and to the work. Use of this work is subject to these terms. Except as permitted under the Copyright Act of 1976 and the right to store and retrieve one copy of the work, you may not decompile, disassemble, reverse engineer, reproduce, modify, create derivative works based upon, transmit, distribute, disseminate, sell, publish or sublicense the work or any part of it without McGraw-Hill Education's prior consent. You may use the work for your own noncommercial and personal use; any other use of the work is strictly prohibited. Your right to use the work may be terminated if you fail to comply with these terms.

THE WORK IS PROVIDED "AS IS." MCGRAW-HILL EDUCATION AND ITS LICENSORS MAKE NO GUARANTEES OR WARRANTIES AS TO THE ACCURACY, ADEQUACY OR COMPLETENESS OF OR RESULTS TO BE OBTAINED FROM USING THE WORK, INCLUDING ANY INFORMATION THAT CAN BE ACCESSED THROUGH THE WORK VIA HYPERLINK OR OTHERWISE, AND EXPRESSLY DISCLAIM ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. McGraw-Hill Education and its licensors do not warrant or guarantee that the functions contained in the work will meet your requirements or that its operation will be uninterrupted or error free. Neither McGraw-Hill Education nor its licensors shall be liable to you or anyone else for any inaccuracy, error or omission, regardless of cause, in the work or for any damages resulting therefrom. McGraw-Hill Education has no responsibility for the content of any information accessed through the work. Under no circumstances shall McGraw-Hill Education and/or its licensors be liable for any indirect, incidental, special, punitive, consequential or similar damages that result from the use of or inability to use the work, even if any of them has been advised of the possibility of such damages. This limitation of liability shall apply to any claim or cause whatsoever whether such claim or cause arises in contract, tort or otherwise.

*To our students, in many countries, whose interest, passion, and probing questions made possible the development of the methods we use to communicate the concepts of evidence-based medicine.*

GG, MOM, and DJC

*To Deb, who has watched over and tended me while I have watched over and tended this wonderful group, with gratitude for her love and her good humor.*

DR

*This page intentionally left blank*

# CONTENTS

---

Contributors.....	xi
Foreword.....	xvii
Preface.....	xxiii

## The Foundations

---

<b>1</b> How to Use the Medical Literature—and This Book—to Improve Your Patient Care.....	1
<b>2</b> What Is Evidence-Based Medicine?.....	9
<b>3</b> What Is the Question?.....	25
<b>4</b> Finding Current Best Evidence.....	43
<b>5</b> Why Study Results Mislead: Bias and Random Error... ..	85

## Therapy

---

<b>6</b> Therapy (Randomized Trials).....	95
<b>7</b> How to Use a Noninferiority Trial.....	127
<b>8</b> Does Treatment Lower Risk? Understanding the Results.....	149
<b>9</b> Confidence Intervals: Was the Single Study or Meta-analysis Large Enough?.....	165

## Harm (Observational Studies)

---

<b>10</b> Harm (Observational Studies).....	179
---	-----

## Diagnosis

---

<b>11</b> The Process of Diagnosis.....	211
<b>12</b> Diagnostic Tests.....	223



## Prognosis

---

<b>13</b> Prognosis .....	251
---------------------------	-----

## Summarizing the Evidence

---

<b>14</b> The Process of a Systematic Review and Meta-analysis.....	269
<b>15</b> Understanding and Applying the Results of a Systematic Review and Meta-analysis .....	291
<b>16</b> Network Meta-analysis.....	327

## Moving From Evidence to Action

---

<b>17</b> How to Use a Patient Management Recommendation: Clinical Practice Guidelines and Decision Analyses .....	357
<b>18</b> Decision Making and the Patient.....	389
Glossary .....	417
Index .....	493

# JAMAevidence

Using Evidence to Improve Care

## **JAMAevidence: Using Evidence to Improve Care**

Founded around the *Users' Guides to the Medical Literature*, *The Rational Clinical Examination: Evidence-Based Clinical Diagnosis*, and *Care at the Close of Life: Evidence and Experience*, JAMAevidence offers an invaluable online resource for learning, teaching, and practicing evidence-based medicine (EBM). Updated regularly, the site includes fully searchable content of the *Users' Guides to the Medical Literature*, *The Rational Clinical Examination*, and *Care at the Close of Life* and features podcasts from the leading minds in EBM, interactive worksheets, functional calculators, and a comprehensive collection of PowerPoint slides for educators and students.

[www.JAMAevidence.com](http://www.JAMAevidence.com)

Please visit [www.JAMAevidence.com](http://www.JAMAevidence.com) for subscription information.



The JAMA Network



*This page intentionally left blank*

# CONTRIBUTORS

---

**Thomas Agoritsas, MD, Dr Med**

Health Information Research Unit  
Department of Clinical  
Epidemiology & Biostatistics  
McMaster University  
Hamilton, Ontario, Canada

**Elie A. Akl, MD**

Department of Medicine  
American University of Beirut  
Riad-El-Solh, Beirut, Lebanon

**Paul Elias Alexander**

Department of Clinical  
Epidemiology & Biostatistics  
Health Research Methodology  
Graduate Program  
McMaster University  
Hamilton, Ontario, Canada

**Waleed Alhazzani, MD,**

**FRCPC, MSc**

Departments of Medicine and  
Clinical Epidemiology &  
Biostatistics  
McMaster University  
Hamilton, Ontario, Canada

**Pablo Alonso-Coello, MD**

Hospital de la Santa Creu i Sant Pau  
Barcelona, Spain

**Shannon M. Bates, MDCM,  
MSc, FRCP(c)**

Department of Medicine  
McMaster University  
Hamilton, Ontario, Canada

**Heiner C. Bucher, MPH**

Basel Institute for Clinical  
Epidemiology and Biostatistics  
University Hospital Basel  
Basel, Switzerland

**Alonso Carrasco-Labra, DDS,  
Msc, PhD(c)**

Department of Clinical  
Epidemiology & Biostatistics  
McMaster University  
Hamilton, Ontario, Canada  
Evidence-Based Dentistry Unit  
Universidad de Chile  
Santiago, Chile

**Deborah J. Cook, MD, FRCPC,  
MSc, OC**

Departments of Clinical  
Epidemiology & Biostatistics  
and Medicine  
McMaster University  
Hamilton, Ontario, Canada

**PJ Devereaux, MD, PhD, FRCPC**

Departments of Clinical  
Epidemiology & Biostatistics  
and Medicine  
McMaster University  
Hamilton, Ontario, Canada

**Glyn Elwyn, MD, MSc,  
FRCGP, PhD**

The Dartmouth Centre for  
Health Care Delivery  
Science  
Hanover, New Hampshire, USA

**xii** CONTRIBUTORS

**Toshi A. Furukawa, MD, PhD**

Departments of Health  
Promotion and Human  
Behavior and Clinical  
Epidemiology  
Kyoto University Graduate  
School of Medicine  
Kyoto, Japan

**Gordon Guyatt, MD, MSc,  
FRCPC, OC**

Departments of Clinical  
Epidemiology & Biostatistics  
and Medicine  
McMaster University  
Hamilton, Ontario, Canada

**Alfred Theodore (Ted) Haines,  
MD, CCFP, MSc, DOHS,  
FRCPC**

Departments of Clinical  
Epidemiology & Biostatistics  
and Family Medicine  
McMaster University  
Chedoke-McMaster Hospitals  
LAMP Community Health  
Centre  
Occupational Health Clinic for  
Ontario Workers  
Hamilton, Ontario, Canada

**Rose Hatala, MD, MSc**

University of British Columbia  
Vancouver, British Columbia,  
Canada

**R. Brian Haynes, MD, PhD**

Departments of Clinical  
Epidemiology & Biostatistics  
and Medicine

McMaster University  
Hamilton, Ontario, Canada

**Robert Hayward, MD**

Owogo Inc.  
Centre for Health Evidence  
Department of Medicine  
University of Alberta  
Edmonton, Alberta, Canada

**John P. A. Ioannidis, MD, DSc**

Departments of Medicine,  
Health Research and Policy,  
and Statistics  
Stanford Prevention Research  
Center  
Meta-Research Innovation Center  
Stanford University  
Stanford, California, USA

**Cynthia A. Jackevicius, BScPhm,  
PharmD, MSc, BCPS, FCSHP**

Western University of Health  
Sciences  
Pomona, California, USA  
Institute for Clinical Evaluative  
Sciences  
Institute for Health Policy,  
Management and Evaluation  
University of Toronto  
Toronto, Ontario, Canada  
Veterans Affairs Greater Los  
Angeles Healthcare System  
Los Angeles, California, USA

**Roman Jaeschke, MD, MSc,  
FRCPC**

Department of Medicine  
St. Joseph's Healthcare  
Hamilton, Ontario, Canada

**Mitchell Levine, MD, MSc**

Department of Clinical  
Epidemiology & Biostatistics  
McMaster University  
Centre for Evaluation of  
Medicines  
St. Joseph's Healthcare  
Hamilton, Ontario, Canada

**Braden Manns, MD**

Departments of Medicine &  
Community Health Sciences  
University of Calgary  
Calgary, Alberta, Canada

**K. Ann McKibbin, MLS, PhD,  
FMLA**

Department of Clinical  
Epidemiology & Biostatistics  
McMaster University  
Hamilton, Ontario, Canada

**Maureen O. Meade, MD, MSc,  
FRCPC**

Departments of Clinical  
Epidemiology & Biostatistics  
and Medicine  
McMaster University  
Hamilton, Ontario, Canada

**Edward J. Mills, PhD, MSc, MSt**

Global Evaluative Sciences  
Vancouver, British Columbia,  
Canada

**Victor M. Montori, MD, MSc**

Knowledge and Evaluation  
Research Unit  
Mayo Clinic  
Rochester, Minnesota, USA

**Sohail M. Mulla, MSc**

Department of Clinical  
Epidemiology & Biostatistics  
Health Research Methodology  
Graduate Program  
McMaster University  
Hamilton, Ontario, Canada

**M. Hassan Murad, MD, MPH**

Division of Preventive Medicine  
Mayo Clinic  
Rochester, Minnesota, USA

**Reem A. Mustafa, MD**

Department of Medicine  
University of Missouri-Kansas City  
Overland Park, Kansas, USA

**Ignacio Neumann, MD, MSc**

Department of Internal Medicine  
Pontificia Universidad Católica  
de Chile

Santiago, Chile

Department of Clinical  
Epidemiology & Biostatistics  
McMaster University  
Hamilton, Ontario, Canada

**Vlado Perkovic, MBBS, PhD,  
FASN, FRACP**

George Institute for Global  
Health Australia, Medicine  
University of Sydney  
Sydney, New South Wales,  
Australia

**Kameshwar Prasad, MD,  
DM, MMSc**

Department of Neurology  
Neurosciences Centre

**xiv** CONTRIBUTORS

All India Institute of Medical  
Sciences  
New Delhi, India

**Milo A. Puhan, MD, PhD**  
Department of Epidemiology  
and Public Health  
Epidemiology, Biostatistics  
and Prevention Institute  
University of Zurich  
Zurich, Switzerland

**Adrienne G. Randolph, MD, MSc**  
Department of Anaesthesia  
Harvard Medical School  
Department of Anesthesia,  
Perioperative and Pain  
Medicine  
Boston Children's Hospital  
Boston, Massachusetts, USA

**W. Scott Richardson, MD**  
Department of Medicine  
Georgia Regents University-  
University of Georgia  
Medical Partnership  
Athens, Georgia, USA

**David M. Rind, MD**  
Department of Medicine  
Harvard Medical School  
Editorial and Evidence-Based  
Medicine, UpToDate  
Wolters Kluwer Health  
Waltham, Massachusetts, USA

**Bram Rochweg, BSc, MD**  
Department of Medicine  
McMaster University  
Hamilton, Ontario, Canada

**Nancy Santesso, BSc(Hon),  
MLIS, PhD(c)**  
Department of Clinical  
Epidemiology & Biostatistics  
McMaster University,  
Hamilton, Ontario, Canada

**Holger J. Schünemann, MD,  
PhD, MSc, FRCPC**  
Departments of Clinical  
Epidemiology & Biostatistics  
and Medicine  
McMaster University  
Hamilton, Ontario, Canada

**Ian A. Scott, MBBS, FRACP,  
MHA, MED**  
Department of Internal  
Medicine and Clinical  
Epidemiology  
Princess Alexandra Hospital  
Department of Medicine  
University of Queensland  
Brisbane, Queensland,  
Australia

**Sadeesh Srinathan, MD, MSc**  
University of Manitoba Health  
Sciences Centre  
Winnipeg, Manitoba, Canada

**Sharon E. Straus, MSc, MD,  
FRCPC**  
Department of Medicine  
Division of Geriatric Medicine  
University of Toronto  
Li Ka Shing Knowledge  
Institute  
St. Michael's Hospital  
Toronto, Ontario, Canada

**Kristian Thorlund,**  
**MSc, PhD**  
Department of Clinical  
Epidemiology & Biostatistics  
McMaster University  
Hamilton, Ontario, Canada

**Per Olav Vandvik,**  
**MD, PhD**  
Department of Medicine  
University of Oslo  
Norwegian Knowledge Centre  
for the Health Services  
Oslo, Norway

**Michael Walsh, MD, PhD**  
Departments of Medicine  
and Clinical Epidemiology &  
Biostatistics  
Population Health Research  
Institute  
Hamilton Health Sciences and  
McMaster University  
Division of Nephrology  
St. Joseph's Hospital  
Hamilton, Ontario, Canada

**Stephen D. Walter,**  
**PhD, FRSC**  
Department of Clinical  
Epidemiology & Biostatistics  
McMaster University  
Hamilton, Ontario, Canada

**Mark C. Wilson, MD, MPH**  
Department of Internal Medicine  
Graduate Medical Education  
Carver College of Medicine  
University of Iowa Hospitals  
and Clinics  
Iowa City, Iowa, USA

**Peter Wyer, MD**  
Columbia University Medical  
Center  
New York, New York, USA

**John J. You, MD, MSc**  
Departments of Medicine and  
Clinical Epidemiology &  
Biostatistics  
McMaster University  
Hamilton, Ontario, Canada



*This page intentionally left blank*

# FOREWORD

---

When I was attending school in wartime Britain, staples of the curriculum, along with cold baths, mathematics, boiled cabbage, and long cross-country runs, were Latin and French. It was obvious that Latin was a theoretical exercise—the Romans were dead, after all. However, although France was clearly visible just across the Channel, for years it was either occupied or inaccessible, so learning the French language seemed just as impractical and theoretical an exercise. It was unthinkable to me and my teachers that I would ever put it to practical use—that French was a language to be spoken.

This is the relationship too many practitioners have with the medical literature—clearly visible but utterly inaccessible. We recognize that practice should be based on discoveries announced in the medical journals. But we also recognize that every few years the literature doubles in size, and every year we seem to have less time to weigh it,<sup>1</sup> so every day the task of taming the literature becomes more hopeless. The translation of those hundreds of thousands of articles into everyday practice appears to be an obscure task left to others, and as the literature becomes more inaccessible, so does the idea that the literature has any utility for a particular patient become more fanciful.

This book, now in its third edition, is intended to change all that. It is designed to make the clinician fluent in the language of the medical literature in all its forms. To free the clinician from practicing medicine by rote, by guesswork, and by their variably integrated experience. To put a stop to clinicians being ambushed by drug company representatives, or by their patients, telling them of new therapies the clinicians are unable to evaluate. To end their dependence on out-of-date authority. To enable the practitioner to work from the patient and use the literature as a tool to solve the patient's problems. To provide the clinician access to what is relevant and the ability to assess its validity and whether it applies to a specific patient. In other

words, to put the clinician in charge of the single most powerful resource in medicine.

## **The Users' Guides Series in *JAMA***

I have left it to Gordon Guyatt, MD, MSc, the moving force, principal editor, and most prolific coauthor of the Users' Guides to the Medical Literature series in *JAMA*, to describe the history of this series and of this book in the accompanying preface. But where did *JAMA* come into this story?

In the late 1980s, at the invitation of my friend David Sackett, MD, I visited his department at McMaster University to discuss a venture with *JAMA*—a series that examined the evidence behind the clinical history and examination. After these discussions, a series of articles and systematic reviews was developed and, with the enthusiastic support of then *JAMA* Editor in Chief George Lundberg, MD, *JAMA* began publishing The Rational Clinical Examination series in 1992.<sup>2</sup> By that time, I had formed an excellent working relationship with the brilliant group at McMaster. Like their leader, Sackett, they tended to be iconoclastic, expert at working together and forming alliances with new and talented workers, and intellectually exacting. Like their leader, they delivered on their promises.

So, when I heard that they were thinking of updating the wonderful little series of Readers' Guides published in 1981 in the *Canadian Medical Association Journal (CMAJ)*, I took advantage of this working relationship to urge them to update and expand the series for *JAMA*. Together with Sackett, and first with Andy Oxman, MD, and then with Gordon Guyatt taking the lead (when Oxman left to take a position in Oslo), the Users' Guides to the Medical Literature series was born. We began publishing articles in the series in *JAMA* in 1993.<sup>3</sup>

At the start, we thought we might have 8 or 10 articles, but the response from readers was so enthusiastic and the variety of types of article in the literature so great that ever since I have found myself receiving, sending for review, and editing new

articles for the series. Just before the first edition of this book was published in 2002, Gordon Guyatt and I closed this series at 25, appearing as 33 separate journal articles.

The passage of years during the preparation of the original *JAMA* series and the publication of the first edition of this book had a particularly useful result. Some subjects that were scarcely discussed in the major medical journals in the early 1990s but that had burgeoned years later could receive the attention that had become their due. For instance, in 2000, *JAMA* published 2 Users' Guides<sup>4,5</sup> on how readers should approach reports of qualitative research in health care. To take another example, systematic reviews and meta-analyses, given a huge boost by the activities of the Cochrane Collaboration, had become prominent features of the literature, and as Gordon Guyatt points out in his preface, the change in emphasis in the Users' Guides to preappraised resources continues.

## **The Book**

From the start, readers kept urging us to put the series together as a book. That had been our intention right from the start, but each new article delayed its implementation. How fortunate! When the original Readers' Guides appeared in the *CMAJ* in 1981, Gordon Guyatt's phrase "evidence-based medicine" had never been coined, and only a tiny proportion of health care workers possessed computers. The Internet did not exist and electronic publication was only a dream. In 1992, the Web—for practical purposes—had scarcely been invented, the dot-com bubble had not appeared, let alone burst, and the health care professions were only beginning to become computer literate. But at the end of the 1990s, when Guyatt and I approached my colleagues at *JAMA* with the idea of publishing not merely the standard printed book but also Web-based and CD-ROM formats of the book, they were immediately receptive. Putting the latter part into practice has been the notable achievement of Rob Hayward, MD, of the Centre for Health Evidence of the University of Alberta.

The science and art of evidence-based medicine, which this book does so much to reinforce, has developed remarkably during the past 25 years, and this is reflected in every page of this book. Encouraged by the immediate success of the first and second editions of the *Users' Guides to the Medical Literature*, Gordon Guyatt and the Evidence-Based Medicine Working Group have once again brought each chapter up to date for this third edition. They have also added 6 completely new chapters: Evidence-Based Medicine and the Theory of Knowledge, How to Use a Noninferiority Trial, How to Use an Article About Quality Improvement, How to Use an Article About Genetic Association, Understanding and Applying the Results of a Systematic Review and Meta-analysis, and Network Meta-analysis. Some of these chapters appear in the larger Manual version of this book.

An updated Web version of the *Users' Guides to the Medical Literature* will accompany the new edition. As part of the online educational resource, JAMAevidence, the *Users' Guides to the Medical Literature* online is intertwined with the online edition of *The Rational Clinical Examination: Evidence-Based Clinical Diagnosis*. Together they serve as the cornerstones of a comprehensive online educational resource for teaching and learning evidence-based medicine. Interactive calculators and worksheets provide practical complements to the content, and downloadable PowerPoint presentations serve as invaluable resources for instructors. Finally, podcast presentations bring the foremost minds behind evidence-based medicine to medical students, residents, and faculty around the world.

Once again, I thank Gordon Guyatt for being an inspired author, a master organizer, and a wonderful teacher, colleague, and friend. I know personally and greatly admire a good number of his colleagues in the Evidence-Based Medicine Working Group, but it would be invidious to name them, given the huge collective effort this has entailed. This is an enterprise that came about only because of the strenuous efforts of many individuals. On the *JAMA* side, I must thank Annette Flanagan, RN, MA, a wonderfully efficient, creative, and diplomatic colleague

at *JAMA*. All of this was coordinated and kept up to schedule by the energy and meticulous efficiency of Kate Pezalla, MA. My colleague, Edward Livingston, MD, a surgeon and a perceptive critic, is taking over the *Users' Guides to the Medical Literature* series at *JAMA*, and I am confident it will prosper in his hands. In addition, I acknowledge the efforts of our partners at McGraw-Hill Education—James Shanahan, Scott Grillo, Michael Crumsho, and Robert Pancotti.

Finally, I thank my friends Cathy DeAngelis, MD, MPH, and her successor, Howard Bauchner, MD, MPH, former and current Editors in Chief of The *JAMA* Network, for their strong backing of me, my colleagues, and this project. Howard inherited this project. Once I found out that his immediate and enthusiastic acceptance of it was based on his regular use of early articles in the *Users' Guides* series, any concern about its reception vanished. Indeed, Howard was the instigator of Evidence-Based Medicine—An Oral History,<sup>2,3</sup> a video series of personal views on the birth and early growth of evidence-based medicine that has helped put the *Users' Guides* into perspective. Howard's infectious good spirits and sharp intelligence bode well for further editions of this book.

**Drummond Rennie, MD**

*University of California, San Francisco*

## References

1. Durack DT. The weight of medical knowledge. *N Engl J Med*. 1978;298(14):773-775.
2. Smith R, Rennie D. Evidence-based medicine—an oral history. *JAMA*. 2014;311(4):365-367.
3. Evidence-based medicine—an oral history website. <http://ebm.jamanetwork.com>. Accessed August 17, 2014.

*This page intentionally left blank*

# PREFACE

---

Evidence-based medicine (EBM)—as a concept with that particular moniker—is now almost 25 years old. Looking back, periods of infancy, childhood, adolescence,<sup>1</sup> and now a mature adulthood are evident.<sup>2</sup> This third edition of the *Users' Guides to the Medical Literature* firmly establishes the maturity of the EBM movement.

The first articulation of the world view that was to become EBM appeared in 1981 when a group of clinical epidemiologists at McMaster University, led by David Sackett, MD, published the first of a series of articles that advised clinicians on how to read clinical journals.<sup>3</sup> Although a huge step forward, the series had its limitations. After teaching what they then called *critical appraisal* for a number of years, the group became increasingly aware of both the necessity and the challenges of going beyond reading the literature in a browsing mode and instead using research studies to solve patient management problems on a day-to-day basis.

In 1990, I assumed the position of residency director of the Internal Medicine Program at McMaster. Through Dave Sackett's leadership, critical appraisal had evolved into a philosophy of medical practice based on knowledge and understanding of the medical literature supporting each clinical decision. We believed that this represented a fundamentally different style of practice and required a term that would capture this difference.

My mission as residency director was to train physicians who would practice this new approach to medicine. In the spring of 1990, I presented our plans for changing the program to the members of the Department of Medicine, many of whom were unsympathetic. The term suggested to describe the new approach was *scientific medicine*. Those already hostile were incensed at the implication that they had previously been “unscientific.” My second try at a name for our philosophy of medical practice,



*evidence-based medicine*, became extremely popular in a very short time. To use the current vernacular, it went viral.<sup>4</sup>

After that fateful Department of Medicine meeting at McMaster, the term *EBM* first appeared in the autumn of 1990 in an information document for residents entering, or considering application to, the residency program. The relevant passage follows:

Residents are taught to develop an attitude of “enlightened scepticism” towards the application of diagnostic, therapeutic, and prognostic technologies in their day-to-day management of patients. This approach . . . has been called “evidence-based medicine.” . . . The goal is to be aware of the evidence on which one’s practice is based, the soundness of the evidence, and the strength of inference the evidence permits. The strategy employed requires a clear delineation of the relevant question(s); a thorough search of the literature relating to the questions; a critical appraisal of the evidence and its applicability to the clinical situation; a balanced application of the conclusions to the clinical problem.

The first published appearance of the term was in the American College of Physicians’ *Journal Club* in 1991.<sup>5</sup> Meanwhile, our group of enthusiastic evidence-based medical educators at McMaster were refining our practice and teaching of EBM. Believing that we were on to something important, we linked up with a larger group of academic physicians, largely from the United States, to form the first Evidence-Based Medicine Working Group and published an article in *JAMA* that defined and expanded on the description of EBM, labeling it as a “paradigm shift.”<sup>6</sup>

This working group then addressed the task of producing a new set of articles, the successor to the Readers’ Guides, to present a more practical approach to applying the medical literature to clinical practice. With the unflagging support and wise counsel of *JAMA* Deputy Editor Drummond Rennie, MD, the Evidence-Based Medicine Working Group created a 25-part series called the *Users’ Guides to the Medical Literature*, published in *JAMA* between 1993 and 2000.<sup>7</sup> The series continues

to be published in *JAMA*, with articles that address new concepts and applications.

The first edition of the *Users' Guides to the Medical Literature* was a direct descendant of the *JAMA* series. By the time of the book's publication in 2002, EBM had already undergone its first fundamental evolution, the realization that evidence was never sufficient for clinical decision making. Rather, management decisions always involve trade-offs between desirable and undesirable consequences and thus require value and preference judgments. Indeed, in the first edition of the *Users' Guide to the Medical Literature*, the first principle of EBM was presented as Clinical Decision Making: Evidence Is Never Enough, joining the previously articulated principle of a hierarchy of evidence.

It did not take long for people to realize that the principles of EBM were equally applicable for other health care workers, including nurses, dentists, orthodontists, physiotherapists, occupational therapists, chiropractors, and podiatrists. Thus, terms such as *evidence-based health care* and *evidence-based practice* are appropriate to cover the full range of clinical applications of the evidence-based approach to patient care. Because our *Users' Guides* are directed primarily at physicians, we have continued with the term *EBM*.

The second edition incorporated 2 new EBM developments in EBM thinking. First, we had realized that only a few clinicians would become skilled at critically appraising original journal articles and that preappraised evidence would be crucial for evidence-based clinical practice. Second, our knowledge of how best to ensure that clinical decisions were consistent with patient values and preferences was rudimentary and would require extensive study.

This third edition of the *Users' Guides to the Medical Literature* builds on these realizations, most substantially in the revised guide to finding the evidence. The emphasis is now on preappraised resources and particularly on the successor to medical texts: electronic publications that produce updated

evidence summaries as the data appear and provide evidence-based recommendations for practice.

Awareness of the importance of preappraised evidence and evidence-based recommendations is reflected in other changes in the third edition. We have added a fundamental principle to the hierarchy of evidence and the necessity for value and preference judgments: that optimal clinical decision making requires systematic summaries of the best available evidence.

This principle has led to a fundamental revision of the *Users' Guide* to systematic reviews, which now explicitly includes the meta-analyses and acknowledges 2 core considerations. The first is how well the systematic review and meta-analysis were conducted. The second, inspired by the contributions of the GRADE (Grading of Recommendations Assessment, Development and Evaluation) Working Group,<sup>8</sup> demands an assessment of the confidence that one can place in the estimates of effect emerging from the review and meta-analysis. However well done the review, if the primary evidence on which it is based warrants little confidence, inferences from the review will inevitably be very limited.

The third edition of the *Users' Guides to the Medical Literature* incorporates the lessons we have learned in more than 20 years of teaching the concepts of EBM to students with a wide variety of backgrounds, prior preparation, clinical interest, and geographic location. Indeed, among our many blessings is the opportunity to travel the world, helping to teach at EBM workshops. Participating in workshops in Thailand, Saudi Arabia, Egypt, Pakistan, Oman, Kuwait, Singapore, the Philippines, Japan, India, Peru, Chile, Brazil, Germany, Spain, France, Belgium, Norway, the United States, Canada, and Switzerland—the list goes on—provides us with an opportunity to try out and refine our teaching approaches with students who have a tremendous heterogeneity of backgrounds and perspectives. At each of these workshops, the local EBM teachers share their own experiences,

struggles, accomplishments, and EBM teaching tips that we can add to our repertoire.

We are grateful for the extraordinary privilege of sharing, in the form of the third edition of *Users' Guides to the Medical Literature*, what we have learned.

**Gordon Guyatt, MD, MSc**

*McMaster University*

## References

1. Daly J. *Evidence-based Medicine and the Search for a Science of Clinical Care*. Berkeley, CA: Milbank Memorial Fund and University of California Press; 2005.
2. Smith R, Rennie D. Evidence-based medicine—an oral history. *JAMA*. 2014;311(4):365-367.
3. Department of Clinical Epidemiology & Biostatistics, McMaster University. How to read clinical journals, I: why to read them and how to start reading them critically. *Can Med Assoc J*. 1981;124(5):555-558.
4. Evidence-based medicine—an oral history website. <http://ebm.jamanetwork.com>. Accessed August 17, 2014.
5. Guyatt G. Evidence-based medicine. *ACP J Club (Ann Intern Med)*. 1991;114(suppl 2):A-16.
6. Evidence-Based Medicine Working Group. Evidence-based medicine: a new approach to teaching the practice of medicine. *JAMA*. 1992;268(17):2420-2425.
7. Guyatt GH, Rennie D. Users' guides to the medical literature. *JAMA*. 1993;270(17):2096-2097.
8. Guyatt GH, Oxman AD, Vist GE, et al; GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336(7650):924-926.

*This page intentionally left blank*

1

# How to Use the Medical Literature— and This Book—to Improve Your Patient Care

Gordon Guyatt and Maureen O. Meade

## IN THIS CHAPTER

---

**The Structure of the *Users' Guides to the Medical Literature*: The Foundations**

**Advanced Topics**

---

The objective of this book is to help you make efficient use of the published literature in guiding your patient care. What does the published literature comprise? Our definition is broad. You may find *evidence*<sup>a</sup> in a wide variety of sources, including original journal articles, *reviews* and *synopses* of *primary studies*, *clinical practice guidelines*, and traditional and innovative medical textbooks. Increasingly, clinicians can most easily access many of these sources through the Internet.

### THE STRUCTURE OF THE *USERS' GUIDES TO THE MEDICAL LITERATURE*: THE FOUNDATIONS

---

This book is not like a novel that you read from beginning to end. Indeed, the *Users' Guides* are designed so that each part is largely self-contained. Thus, we anticipate that clinicians may be selective in their reading of the core content chapters and will certainly be selective when they move beyond the essentials. On the first reading, you may choose only a few advanced areas that interest you. If, as you use the medical literature, you find the need to expand your understanding of, for instance, studies addressing *screening* tests or the use of *surrogate outcomes*, you can consult the relevant chapters to familiarize or reacquaint yourself with the issues. You may also find the glossary a useful reminder of the formal definitions of terms used herein. Finally, we rely heavily on examples to make our points. You will find examples identified by their blue background.

The Essentials version of this book comprises 18 chapters in 7 sections: The Foundations, Therapy, Harm, Diagnosis, Prognosis, Summarizing the Evidence, and Moving From Evidence to Action (Box 1-1). A larger Manual version of this book includes additional chapters in each section.

---

<sup>a</sup>The italicization, here and in every other chapter, represents the first occurrence in the chapter of a word defined in the glossary.

## BOX 1-1

**Sections of This Book**

The Foundations  
Therapy  
Harm  
Diagnosis  
Prognosis  
Summarizing the Evidence  
Moving From Evidence to Action

The first section of this book introduces the foundations of *evidence-based practice*. Two chapters in this section, What Is Evidence-Based Medicine? and Evidence-Based Medicine and the Theory of Knowledge, present the 3 guiding principles of *evidence-based medicine* (EBM), and place EBM in the context of a humanistic approach to medical practice. The subsequent chapters in this section deal with defining your clinical question, locating the best evidence to address that question, and distinguishing *bias* from *random error* (a key principle of critical appraisal).

Clinicians are primarily interested in making accurate diagnoses and selecting optimal treatments for their patients. They also must avoid exposing patients to *harm* and offer patients prognostic information. Thus, chapters in 4 sections of this book (Therapy, Harm, Diagnosis, and Prognosis) begin by outlining what every medical student, intern and resident, and practicing physician and other clinicians will need to know to use articles that present primary data that address these 4 principal issues in providing patient care.

Increasingly, we have become aware that individual studies are often unrepresentative of all relevant studies (ie, showing larger or smaller *treatment effects* than *pooled estimates* of all relevant



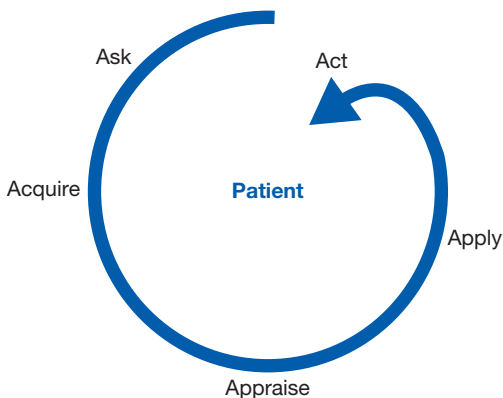
studies), imprecise, or limited in their applicability—so much so that, since the previous edition of this book, we have added the need for systematic summaries of all relevant studies as a core principle of EBM. This has major implications for clinicians looking to use the literature to provide optimal patient care. Efficient and optimally effective evidence-based practice dictates bypassing the critical assessment of primary studies and, if they are available, moving straight to the evaluation of rigorous *systematic reviews*. Even more efficient than using a systematic review is moving directly to an evidence-based recommendation. Ideally, management recommendations—summarized in clinical practice guidelines or *decision analyses*—will incorporate the best evidence and make explicit the value judgments used in moving from evidence to recommendations for action. Unfortunately, many clinical practice guidelines sometimes provide recommendations that are inconsistent with the best evidence or with typical patient *values and preferences*. The last 2 sections of the book, Summarizing the Evidence and Moving From Evidence to Action, provide clinicians with guides for using systematic reviews (with and without *meta-analyses*) and recommendations to optimize their patient care.

Our approach to addressing diagnosis, therapy, harm, and *prognosis* begins when the clinician faces a clinical question (Figure 1-1). Having identified the problem, the clinician then formulates a structured clinical question (the “Ask,” Figure 1-1) (see Chapter 3, What Is the Question?) and continues with finding the best relevant evidence (the “Acquire,” Figure 1-1) (see Chapter 4, Finding Current Best Evidence).

Many chapters of this book include an example of a search for the best evidence. These searches were accurate at the time they were done, but you are unlikely to get exactly the same results if you replicate the searches now. The reasons for this include additions to the literature and occasional structural changes in databases. Thus, you should view the searches as illustrations of searching principles, rather than as currently definitive searches that address the clinical question. Having identified the best evidence, the clinician then proceeds

FIGURE 1-1

## Using the Medical Literature to Provide Optimal Patient Care



through the next 3 steps in evaluating that evidence: appraisal, considering how to apply the results, and acting (Figure 1-1). The appraisal includes 2 questions, “How serious is the *risk of bias*?” and “What are the results?” The first question, “How serious is the risk of bias?” deals with the extent to which the results represent an unbiased estimate of the truth. In the first 2 editions of this book, we referred to risk of bias as *validity* and used the question, “Are the results valid?” We have made this change because “risk of bias” is a more explicit and transparent term. In Chapter 7, How to Use a Noninferiority Trial, limitations of study design related to these topics include issues beyond risk of bias. Therefore, in Chapter 7, we continue to use the term validity and the question “Are the results valid?” to capture the risk of bias and these additional issues.

The second question in the appraisal step is, “What are the results?” For issues of therapy or harm, this will involve assessing the magnitude and precision of the impact of the intervention (a treatment or possible harmful exposure) (see Chapter 6,

Therapy [Randomized Trials]; Chapter 7, How to Use a Noninferiority Trial; Chapter 8, Does Treatment Lower Risk? Understanding the Results; Chapter 9, Confidence Intervals: Was the Single Study or Meta-analysis Large Enough? and Chapter 10, Harm [Observational Studies]). For issues of diagnosis, this will involve generating *pretest probabilities* and then *posttest probabilities* on the basis of test results (see Chapter 11, The Process of Diagnosis, and Chapter 12, Diagnostic Tests). For issues of prognosis, this will involve determining the likelihood of events occurring over time and the precision of those estimates (see Chapter 13, Prognosis).

Once we understand the results, we move to dealing with applicability (Figure 1-1) and ask ourselves the third question: "How can I apply these results to patient care?" This question has 2 parts. First, can you *generalize* (or, to put it another way, *particularize*) the results to your patient? For instance, your confidence in estimates of treatment effect decreases if your patient is too dissimilar from those who participated in the *trial* or trials. Second, what is the significance of the results for your patient? Have the investigators measured all *patient-important outcomes*? What is the tradeoff among the benefits, *risks*, and *burdens* of alternative management strategies?

Often, you will find a systematic review that, if it is done well and includes a meta-analysis (see Chapter 14, The Process of a Systematic Review and Meta-analysis), will have conducted the search and risk of bias appraisals and, further, summarized the results and suggested the confidence you can place in estimates (see Chapter 15, Understanding and Applying the Results of a Systematic Review and Meta-analysis). In addition, you often will find a recommendation that, if developed rigorously, is based on trustworthy systematic reviews of the evidence and explicitly considers patient values and preferences (see Chapter 17, How to Use a Patient Management Recommendation: Clinical Practice Guidelines and Decision Analyses) and provides guidance on the issue of applying the results to your patient. In our discussions of systematic reviews and guidelines, we introduce the

*GRADE (Grading of Recommendations Assessment, Development and Evaluation)* approach to summarizing evidence and developing recommendations, an approach that we believe represents a major advance in EBM (see Chapter 15, Understanding and Applying the Results of a Systematic Review and Meta-analysis).

The final step in using the evidence is action (Figure 1-1). Often, this will involve shared decision making with your patients (see Chapter 18, Decision Making and the Patient), a key part of the EBM process.

We have kept the initial chapters of each part of this book simple and succinct. From an instructor's point of view, these core chapters constitute a curriculum for a short course in using the literature for medical students, resident physicians, or students of other health professions. They also are appropriate for a continuing education program for practicing physicians and other clinicians.

## ADVANCED TOPICS

---

Moving beyond the foundations, the advanced topics in this book will interest clinicians who want to practice EBM at a more sophisticated level. They are organized according to the core issues addressed in the sections on Therapy, Harm, Diagnosis, and Prognosis.

The presentations of advanced topics will deepen your understanding of study methods, statistical issues, and use of the numbers that emerge from medical research. We wrote the advanced chapters mindful of an additional audience: those who teach evidence-based practice. Many advanced entries read like guidelines for an interactive discussion with a group of learners in a tutorial or on the ward. That is natural enough because the material was generated in such small-group settings. Indeed, the Evidence-Based Medicine Working Group has produced materials that specifically discuss the challenges that arise when these concepts are presented in small-group

settings, including a series of 5 articles published in the *Canadian Medical Association Journal*<sup>1</sup> and another 5 articles in the *Journal of General Internal Medicine*.<sup>2</sup>

Experience on the wards and in outpatient clinics, and with the first 2 editions of the *Users' Guides to the Medical Literature*, has taught us that this approach is well suited to the needs of any clinician who is eager to achieve an evidence-based practice.

## References

1. Wyer PC, Keitz S, Hatala R, et al. Tips for learning and teaching evidence-based medicine: introduction to the series. *CMAJ*. 2004;171(4):347-348.
2. Kennedy CC, Jaeschke R, Keitz S, et al; Evidence-Based Medicine Teaching Tips Working Group. Tips for teachers of evidence-based medicine: adjusting for prognostic imbalances (confounding variables) in studies on therapy or harm. *J Gen Intern Med*. 2008;23(3):337-343.

2

# What Is Evidence-Based Medicine?

Gordon Guyatt, Roman Jaeschke,  
Mark C. Wilson, Victor M. Montori,  
and W. Scott Richardson

## IN THIS CHAPTER

---

### Three Fundamental Principles of EBM

- Best Evidence Summaries
- Guides to Confidence in Estimates
- Evidence Is Never Enough to Drive Clinical Decision Making

### Clinical Skills, Humanism, and EBM

### Additional Challenges for EBM

---

*Evidence-based medicine* (EBM) involves conscientiously working with patients to help them resolve (sometimes) or cope with (often) problems related to their physical, mental, and social health. The EBM approach necessitates awareness and understanding of clinical research *evidence*. For those involved in making health care decisions, EBM encompasses creating implementation strategies to ensure practice evidence that is well grounded in best evidence research summaries.

At the core of EBM is a care and respect for patients who will suffer if clinicians fall prey to muddled clinical reasoning and to neglect or misunderstanding of research findings. Practitioners of EBM strive for a clear and comprehensive understanding of the evidence underlying their clinical care and work with each patient to ensure that chosen courses of action are in that patient's best interest. Practicing EBM requires clinicians to understand how uncertainty about clinical research evidence intersects with an individual patient's predicament and preferences. In this chapter, we outline how EBM proposes to achieve these goals and, in so doing, define the nature of EBM.

## THREE FUNDAMENTAL PRINCIPLES OF EBM

---

Conceptually, EBM involves 3 fundamental principles. First, optimal clinical decision making requires awareness of the best available evidence, which ideally will come from systematic summaries of that evidence. Second, EBM provides guidance to decide whether evidence is more or less trustworthy—that is, how confident can we be of the properties of diagnostic tests, of our patients' *prognosis*, or of the impact of our therapeutic options? Third, evidence alone is never sufficient to make a clinical decision. Decision makers must always trade off the benefits and *risks, burden*, and costs associated with alternative management strategies and, in doing so, consider their patients' unique predicament and *values and preferences*.<sup>1</sup>

## Best Evidence Summaries

In 1992, Antman et al<sup>2</sup> published an article that compared the recommendations of experts for management of patients with myocardial infarction to the evidence that was available at the time the recommendations were made. Figures 2-1 and 2-2 summarize their results in *forest plots*. Both are cumulative *meta-analyses*: the first of thrombolytic therapy for myocardial infarction and the second for lidocaine antiarrhythmic therapy. In both cases, the line in the center represents an *odds ratio* of 1.0 (treatment is neither beneficial nor harmful). As in any forest plot, the dots represent the best estimates of treatment effect (often from individual studies; in this case from the totality of accumulated evidence), and the associated lines represent the 95% *confidence intervals* (CIs).

The “Patients” column presents the total number of patients enrolled in all *randomized clinical trials* (RCTs) conducted to the date specified in the “Year” column—the reason we call it a cumulative meta-analysis. In both figures, early on, with relatively few patients, the CIs are wide, but they progressively narrow as new trials were reported.

For the thrombolytic example, by 10 trials and approximately 2500 patients, it appears that thrombolytic therapy reduces mortality, but the CIs are still wide enough to permit residual uncertainty. By 30 trials and more than 6000 patients, the reduction in odds of death of approximately 25% seems secure.

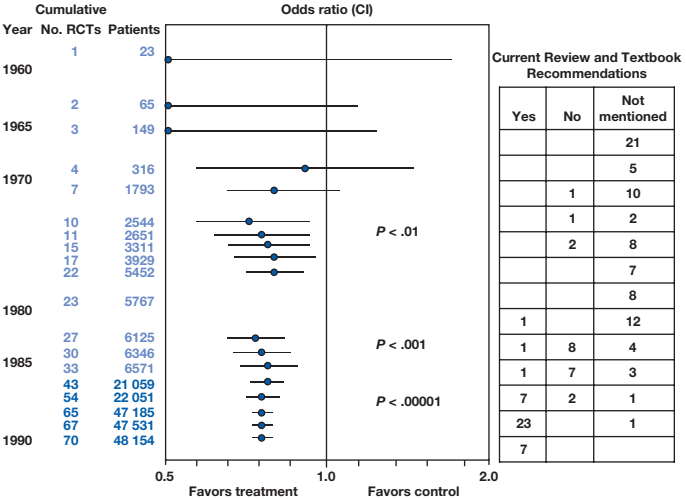
Despite this apparently definitive result, additional trials that enrolled 40 000 patients—half of whom did not receive the benefits of life-prolonging thrombolytic therapy—were conducted. Why was this necessary?

The right side of each figure, which presents the guidance expressed in then-current reviews and textbooks as the data were accumulating, provides the answer to this question. Until approximately a decade after the answer was in, there was considerable disagreement among experts, with many recommending against, or not mentioning, thrombolytic therapy. To the detriment of patients who did not receive thrombolytic



FIGURE 2-1

Thrombolytic Therapy in Acute Myocardial Infarction



Abbreviations: CI, confidence interval; RCTs, randomized clinical trials. This is a cumulative meta-analysis of thrombolytic therapy for myocardial infarction. The line down the center, the odds ratio, equals 1.0. The dots represent best estimates, and the lines around the dots are 95% CIs. The numbers on the left side of the figure are trials and patient totals across trials.

Early on, the CIs are very wide. By 10 trials, it appears therapy reduces mortality, but the effect is still uncertain. By 30 trials, the effect seems secure. However, 40000 more patients were enrolled after the answer was in. Why?

The right side of the figure displays current reviews and textbook recommendations as data accumulated. Recommendations are in favor ("Yes"), against ("No"), or "Not mentioned." Two key points: (1) at the same time, experts disagreed, and (2) it took 10 years for experts to catch up with evidence.

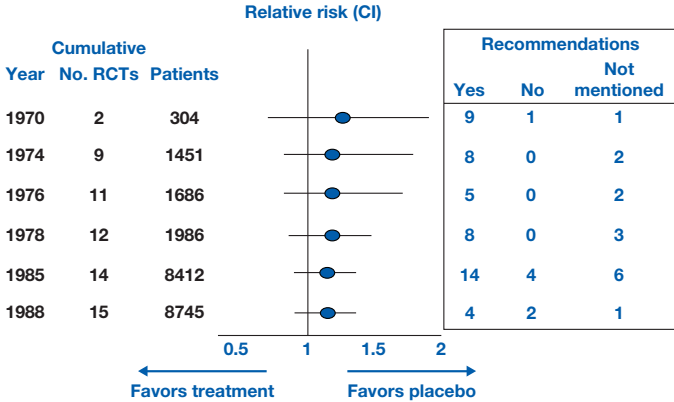
Reproduced from Antman et al.<sup>2</sup>

therapy during this period, it took a decade for the experts to catch up with the evidence.

Figure 2-2 tells a perhaps even more disturbing story. This cumulative meta-analysis reveals that there was never any RCT

FIGURE 2-2

## Prophylactic Lidocaine in Acute Myocardial Infarction



Abbreviations: CI, confidence interval; RCTs, randomized clinical trials.

This figure shows a cumulative meta-analysis of the effect of prophylactic lidocaine in preventing death from myocardial infarction. In this case, there is never any evidence of benefit. Ultimately, harm is not proved, but there clearly is no benefit. Most experts, however, were recommending therapy despite RCT evidence. Also, as in Figure 2-1, there was a lot of disagreement among experts.

Reproduced from Antman et al.<sup>2</sup>

evidence that suggested a lower mortality with prophylactic lidocaine after myocardial infarction—indeed, *point estimates* suggested an increase in death rate. Nevertheless, although we once again see widespread disagreement among the experts, most texts and reviews were recommending prophylactic lidocaine during the 2 decades during which the RCT evidence was accumulating.

Why the expert disagreement, the lag behind the evidence, and the recommendations inconsistent with the evidence? These stories come from the era before *systematic reviews* and meta-analyses were emerging in the late 1980s. If the evidence

summaries presented in the forest plots had been available to the experts, they would have grasped the benefits of thrombolytic therapy far earlier than they did and abandoned prophylactic lidocaine far earlier. Indeed, following EBM principles that limit reliance on biologic rationale and place far more emphasis on empirical evidence, the experts may never have started using lidocaine.

Rational clinical decisions require systematic summaries of the best available evidence. Without such summaries, clinicians—expert or otherwise—will be unduly influenced by their own preconceptions and by unrepresentative and often lower-quality evidence. This, the first principle of EBM, immediately raises another question: “How does one recognize the best evidence?”

### Guides to Confidence in Estimates

Summaries of the best evidence for diagnosis, prognosis, or treatment present evidence, respectively, for how to interpret test results, predict patients' likely fate, or understand the impact of alternative management strategies. Sometimes, such evidence is trustworthy—we have high confidence in estimates of test properties, patients' prognosis, or treatment effects. At other times, limitations in evidence leave us uncertain. Evidence-based medicine provides guidance to distinguish between these situations and the range of confidence between them.

Historically, EBM answered the question, “What is the best evidence?” with *hierarchies of evidence*, the most prominent of which was the hierarchy related to evidence that supported therapeutic interventions (Figure 2-3). Issues of diagnosis or prognosis require different hierarchies. For studies of the accuracy of diagnostic tests, the top of the hierarchy includes studies that enrolled patients about whom clinicians had diagnostic uncertainty and that undertook a *blind* comparison between the candidate test and a *criterion standard* (see Chapter 12, Diagnostic Tests, and Chapter 13, Prognosis). For prognosis, prospective observational studies accurately documenting *exposures* and outcomes and

FIGURE 2-3

### Hierarchy of Evidence



Because we would like to optimally individualize patient care, n-of-1 randomized clinical trials are at the top of the hierarchy of study designs, followed by conventional randomized trials. Next in the hierarchy are observational studies; we should try to find studies that focus on outcomes important to the patient. Next, if there are no clinical studies available, we may look at basic scientific research, although caution must be used in extrapolating the results to the clinical setting. Clinical experience is at the bottom of the hierarchy, either your own or that of colleagues or experts.

following up all patients during relevant periods would sit atop the hierarchy.

Returning to the hierarchy of therapy, noting the limitations of human intuition,<sup>3</sup> EBM places the unsystematic observations of individual clinicians lowest on the hierarchy. Noting that predictions based on physiologic experiments are often right but sometimes disastrously wrong, EBM places such experiments at the next step up in the hierarchy. Observational studies

that measure the apparent impact on *patient-important outcomes* and RCTs constitute the next 2 steps up the hierarchy of evidence.

All of the sources of evidence mentioned thus far involve generalizations from groups of patients to an individual, and all are limited in this regard. The same strategies that minimize *bias* in conventional therapeutic trials that involve multiple patients, however, can guard against misleading results in studies that involve single patients.<sup>4</sup> In the *n-of-1 RCT*, a patient and clinician are blind to whether that patient is receiving active or *placebo* medication. The patient makes quantitative ratings of troublesome symptoms during each period, and the *n-of-1 RCT* continues until both the patient and the clinician conclude that the patient is or is not obtaining benefit from the target intervention. An *n-of-1 RCT* can provide definitive evidence of treatment effectiveness in individual patients<sup>5,6</sup> and is thus at the top of the evidence hierarchy. Unfortunately, *n-of-1 RCTs* are restricted to chronic conditions with treatments that act and cease acting quickly and are subject to considerable logistic challenges. We therefore must usually rely on studies of other patients to make inferences regarding our patient.

This hierarchy is far from absolute, and a more sophisticated framework has emerged for judging confidence in estimates of effect. [Table 2-1](#) summarizes that framework, formulated by the *GRADE (Grading of Recommendations Assessment, Development and Evaluation)* Working Group, originally to provide an approach to the development of *clinical practice guidelines*.<sup>7,8</sup> The *GRADE* approach involves rating our confidence in estimates of the effects of health care interventions (also referred to as *quality of evidence*) as high, moderate, low, or very low. Consistent with the previous hierarchy approach, in the *GRADE* guidance, RCTs begin as high confidence and observational studies begin as low confidence. We lose confidence in a body of RCT evidence, however, if studies have major problems in design and execution (*risk of bias*); results are imprecise, inconsistent, or indirect (eg, the population of

TABLE 2-1

Confidence Assessment Criteria<sup>a</sup>

Study Design	Confidence in Estimates	Lower If <sup>a</sup>	Higher If <sup>a</sup>
Randomized trial	High	Risk of bias	
		-1 Serious	
	-2 Very serious		
	Moderate	Inconsistency	
-1 Serious		Large effect	
Observational study	Low	-2 Very serious	+1 Large
		Indirectness	+2 Very large
	-1 Serious	Dose response	
	-2 Very serious	+1 Evidence of a gradient	
	Very low	Imprecision	
		-1 Serious	
-2 Very serious			
Publication bias			
		-1 Likely	
		-2 Very likely	

<sup>a</sup>Minus and plus signs refer, respectively, to rating down and rating up confidence in estimates. The 1 refers to rating down or up by 1 level (eg, from high to moderate or moderate to high), and the 2 refers to rating down or up by 2 levels (eg, high to low or low to high).

interest differs from the population studied); or we have a high suspicion of publication bias (see Chapter 15, Understanding and Applying the Results of a Systematic Review and Meta-analysis). When a body of RCT evidence suffers from a number of these limitations, the confidence in estimates may be low or even very low.

Similarly, if treatment effects are sufficiently large and consistent, the GRADE approach allows for moderate or even high confidence ratings from carefully conducted observational studies.

For example, observational studies have allowed extremely strong inferences about the efficacy of insulin in diabetic ketoacidosis or that of hip replacement in patients with debilitating hip osteoarthritis.

The EBM approach implies a clear course of action for clinicians addressing patient problems. They should seek the highest-quality evidence available to guide their clinical decisions. This approach makes it clear that any claim that there is no evidence for the effect of a particular treatment is a non sequitur. The available evidence may warrant very low confidence—it may be the unsystematic observation of a single clinician or physiologic studies that point to mechanisms of action that are only indirectly related—but there is always evidence.

### **Evidence Is Never Enough to Drive Clinical Decision Making**

First, picture a woman with chronic pain from terminal cancer. She has come to terms with her condition, resolved her affairs, said her good-byes, and wishes to receive only palliative care. She develops severe pneumococcal pneumonia. Evidence that antibiotic therapy reduces morbidity and mortality from pneumococcal pneumonia warrants high confidence. This evidence does not, however, dictate that this patient should receive antibiotics. Her values—emerging from her comorbidities, social setting, and beliefs—are such that she would prefer to forgo treatment.

Now picture a second patient, an 85-year-old man with severe dementia who is mute and incontinent, is without family or friends, and spends his days in apparent discomfort. This man develops pneumococcal pneumonia. Although many clinicians would argue that those responsible for his decision making should elect not to administer antibiotic therapy, others would suggest that they should. Again, evidence of treatment effectiveness does not automatically imply that treatment should be administered.

Finally, picture a third patient, a healthy 30-year-old mother of 2 children who develops pneumococcal pneumonia. No clinician would doubt the wisdom of administering antibiotic therapy to this patient. This does not mean, however, that an underlying value judgment has been unnecessary. Rather, our values are sufficiently concordant, and the benefits so overwhelm the risk of treatment that the underlying value judgment is unapparent.

By values and preferences, we mean the collection of goals, expectations, predispositions, and beliefs that individuals have for certain decisions and their potential outcomes. The explicit enumeration and balancing of benefits and risks that are central to EBM bring the underlying value judgments involved in making management decisions into bold relief.

Acknowledging that values play a role in every important patient care decision highlights our limited understanding of how to ensure that decisions are consistent with individual and, where appropriate, societal values. As we discuss further in the final section of this chapter, developing efficient processes for helping patients and clinicians work together toward optimal decisions consistent with patient values and preferences remains a frontier for EBM.

Next, we comment on additional skills that clinicians must master for optimal patient care and the relation of those skills to EBM.

## CLINICAL SKILLS, HUMANISM, AND EBM

---

In summarizing the skills and attributes necessary for *evidence-based practice*, [Box 2-1](#) highlights how EBM complements traditional aspects of clinical expertise. One of us, an intensive care specialist, developed a lesion on his lip shortly before an important presentation. He was concerned and, wondering whether he should take acyclovir, proceeded to spend the next 30 minutes searching for and evaluating the highest-quality evidence. When he began to discuss his remaining uncertainty



**BOX 2-1****Knowledge and Skills Necessary for Optimal Evidence-Based Practice**

- Diagnostic expertise
- In-depth background knowledge
- Effective searching skills
- Effective critical appraisal skills
- Ability to define and understand benefits and risks of alternatives
- In-depth physiologic understanding that allows application of evidence to the individual
- Sensitivity and communication skills required for full understanding of patient context
- Ability to elicit and understand patient values and preferences and work with patients in shared decision making

with his partner, an experienced dentist, she cut short the discussion by exclaiming, “But, my dear, that isn’t herpes!”

This story illustrates the necessity of obtaining the correct diagnosis before seeking and applying research evidence regarding optimal treatment. After making the diagnosis, the clinician relies on experience and background knowledge to define the relevant management options. Having identified those options, the clinician can search for, evaluate, and apply the best evidence regarding patient management.

In applying evidence, clinicians rely on their expertise to define features that affect the applicability of the results to the individual patient. The clinician must judge the extent to which differences in treatment (for instance, local surgical expertise or the possibility of patient *nonadherence*) or patient characteristics (such as age, comorbidity, or the patient’s

personal circumstances) may affect estimates of benefit and risk that come from the published literature.

We note that some of these skills—the sensitivity to the patient’s unique predicament and the communication skills necessary for shared decision making—are often not typically associated with EBM. We believe they are, in fact, at the core of EBM. Understanding the patient’s personal circumstances is of particular importance and requires advanced clinical skills, including listening skills and compassion. For some patients, incorporation of patient values for major decisions will mean a full enumeration of the possible benefits, risks, and inconveniences associated with alternative management strategies. For some patients and problems, this discussion should involve the patient’s family. For other problems—the discussion of *screening* with prostate-specific antigen with older male patients, for instance—attempts to involve family members might violate cultural norms.

Some patients are uncomfortable with an explicit discussion of benefits and risk and object to clinicians placing what they perceive as excessive responsibility for decision making on their shoulders. In such cases, it is the physician’s responsibility to develop insight to ensure that choices will be consistent with the patient’s values and preferences while remaining sensitive to the patient’s preferred role in decision making.

## ADDITIONAL CHALLENGES FOR EBM

---

Busy clinicians—particularly those early in their development of the skills needed for *evidence-based practice*—will find that they often perceive time limitations as the biggest challenge to evidence-based practice. This perception may arise from having inadequate access to various evidence-based resources. Fortunately, a tremendous array of sophisticated evidence-based information is now available for clinicians working in high-income countries, and the pace of innovation remains extremely rapid (see Chapter 4, Finding Current Best Evidence).

Access to preprocessed information cannot, however, address other skills required for efficient evidence-based practice. These skills include formulating focused clinical questions, matching prioritized questions to the most appropriate resources, assessing confidence in estimates, and understanding how to apply results to clinical decision making. Although these skills take time to learn, the reward in terms of efficient and effective practice can more than compensate.

Another challenge for evidence-based practice is ensuring that management strategies are consistent with patients' values and preferences. In a time-constrained environment, how can we ensure that patients' involvement in decision making has the form and extent that they desire and that the outcome reflects their needs and desires? Evidence-based medicine leaders are now making progress in addressing these challenges.<sup>9,10</sup>

This book deals primarily with decision making at the level of the individual patient. Evidence-based approaches can also inform health care policy making, day-to-day decisions in public health, and systems-level decisions, such as those facing hospital managers. In each of these areas, EBM can support the appropriate goal of gaining the greatest health benefit from limited resources.

In the policy arena, dealing with differing values poses even more challenges than in the arena of individual patient care. Should we restrict ourselves to alternative resource allocation within a fixed pool of health care resources, or should we consider expanding health care services at the cost, for instance, of higher tax rates for individuals or corporations? How should we deal with the large body of observational studies that suggest that social and economic factors may have a larger influence on the health of populations than health care provision? How should we deal with the tension between what may be best for a person and what may be optimal for the society of which that person is a member? The debate about such issues is at the core of evidence-based policy making in health care; it also has implications for decision making at the individual patient level.

## References

1. Napodano R. *Values in Medical Practice*. New York, NY: Humana Sciences Press; 1986.
2. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: treatments for myocardial infarction. *JAMA*. 1992;268(2): 240-248.
3. Nisbett R, Ross L. *Human Inference*. Englewood Cliffs, NJ: Prentice-Hall; 1980.
4. Guyatt G, Sackett D, Taylor DW, Chong J, Roberts R, Pugsley S. Determining optimal therapy—randomized trials in individual patients. *N Engl J Med*. 1986;314(14):889-892.
5. Guyatt GH, Keller JL, Jaeschke R, Rosenbloom D, Adachi JD, Newhouse MT. The n-of-1 randomized controlled trial: clinical usefulness: our three-year experience. *Ann Intern Med*. 1990;112(4):293-299.
6. Larson EB, Ellsworth AJ, Oas J. Randomized clinical trials in single patients during a 2-year period. *JAMA*. 1993;270(22):2708-2712.
7. Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schünemann HJ; GRADE Working Group. What is “quality of evidence” and why is it important to clinicians? *BMJ*. 2008;336(7651):995-998.
8. Balshem H, Helfand M, Schünemann HJ, et al. GRADE guidelines, 3: rating the quality of evidence. *J Clin Epidemiol*. 2011;64(4):401-406.
9. Montori VM, Guyatt GH. Progress in evidence-based medicine. *JAMA*. 2008;300(15):1814-1816.
10. Stiggelbout AM, Van der Weijden T, De Wit MP, et al. Shared decision making: really putting patients at the centre of healthcare. *BMJ*. 2012;344:e256.

*This page intentionally left blank*

3

# What Is the Question?

Gordon Guyatt, Maureen O. Meade,  
Thomas Agoritsas, W. Scott Richardson,  
and Roman Jaeschke

## IN THIS CHAPTER

---

### **Three Ways to Use the Medical Literature**

Staying Alert to Important New Evidence

Problem Solving

Asking Background and Foreground Questions

### **Clarifying Your Question**

The Structure: Patients, Exposures, Outcome

Five Types of Foreground Clinical Questions

Finding a Suitably Designed Study for Your  
Question Type

Three Examples of Question Clarification

Example 1: Diabetes and Target Blood Pressure

Example 2: Transient Loss of Consciousness

Example 3: Squamous Cell Carcinoma

### **Conclusion: Defining the Question**

---

## THREE WAYS TO USE THE MEDICAL LITERATURE

---

Consider a medical student, early in her training, seeing a patient with newly diagnosed type 2 diabetes mellitus. She will ask questions such as the following: “What is type 2 diabetes mellitus?” “Why does this patient have polyuria?” “Why does this patient have numbness and pain in his legs?” “What treatment options are available?” These questions address normal human physiology and the pathophysiology associated with a medical condition.

Traditional medical textbooks, whether in print or online, that describe underlying pathophysiology or epidemiology of a disorder provide an excellent resource for addressing these *background questions*. In contrast, the sorts of *foreground questions* that experienced clinicians usually ask require different resources. Formulating a question is a critical and generally unappreciated skill for *evidence-based practice*. The following ways to use the medical literature provide opportunities to practice that skill.

### Staying Alert to Important New Evidence

A general internist is checking e-mails on a smartphone while riding public transit to work. While screening a weekly e-mail alert from EvidenceUpdates (<http://plus.mcmaster.ca/EvidenceUpdates>, Figure 3-1), the internist sees an article titled, Cardiovascular Effects of Intensive Lifestyle Intervention in Type 2 Diabetes,<sup>1</sup> recently published and rated by internist colleagues as newsworthy and highly relevant for practice.

This internist is in the process of addressing a question that clinicians at all stages of training and career development are constantly posing: “What important new evidence should I know to optimally treat patients?” Clinicians traditionally addressed this question by attending rounds and conferences

FIGURE 3-1

## Example of E-mail Alert From EvidenceUpdates

EvidenceUPDATES  
FROM THE BMJ EVIDENCE CENTRE

Dear Dr. Agoritsas:

New articles: colleagues in your discipline have identified the following article(s) as being of interest:

Article Title	Discipline	Relevance	News-worthiness
Cardiovascular Effects of Intensive Lifestyle Intervention in Type 2 Diabetes, N Engl J Med	General Practice(GP)/Family Practice(FP) General Internal Medicine-Primary Care(US)	7 7	6 6
Comparison of different regimens of proton pump inhibitors for acute peptic ulcer bleeding, Cochrane Database Syst Rev	Hospital Doctor/Hospitalists Internal Medicine	6 6	6 6
Effects of Combined Application of Muscle Relaxants and Celecoxib Administration After Total Knee Arthroplasty (TKA) on Early Recovery: A Randomized, Double-Blind, Controlled Study, J Arthroplasty	Hospital Doctor/Hospitalists Internal Medicine	6 6	4 4
Corticosteroids for acute bacterial meningitis, Cochrane Database Syst Rev	Hospital Doctor/Hospitalists Internal Medicine	6 6	4 4

and by subscribing to target medical journals in which articles relevant to their practice appear. They kept up-to-date by skimming the table of contents and reading relevant articles.

This traditional approach to what we might call the browsing mode of using the medical literature has major limitations of inefficiency and its resulting frustration. Many screened articles may prove of little relevance or newsworthiness or fail to meet the critical appraisal criteria that are presented in this book. To make matters worse, the volume of research is markedly increasing,<sup>2</sup> and relevant studies appear in a large variety of journals.<sup>3</sup> *Evidence-based medicine* offers solutions to these problems.

The most efficient strategy for ensuring you are aware of recent developments relevant to your practice is to subscribe to e-mail alerting systems, such as EvidenceUpdates, used by the internist in this example. This free service has research staff screening approximately 45 000 articles per year in more than 125 clinical journals for methodologic quality and a worldwide panel of practicing physicians rating them for clinical relevance and newsworthiness.<sup>4</sup> You can tailor alerting systems to your information needs (clinical disciplines and frequency of alerts)



and identify the 20 to 50 articles per year that will influence your practice.<sup>5</sup> Several other free or subscription-based alerting systems are available, both for a wide scope of disciplines (eg, NEJM Journal Watch, <http://www.jwatch.org>) and for specific subspecialties (eg, OrthoEvidence, <http://www.myortho.evidence.com>).

An alternative to alerting systems are *secondary evidence-based journals*. For example, in internal and general medicine, *ACP Journal Club* (<http://acpjc.acponline.org>) publishes *synopses* of articles that meet criteria of both high clinical relevance and methodologic quality. We describe such secondary journals in more detail in Chapter 4, Finding Current Best Evidence. If you prefer browsing to receiving alerts, such preappraised sources of *evidence* may increase your efficiency.

Some specialties (primary care and mental health care) and subspecialties (cardiology, oncology, and obstetrics and gynecology) already have specialty-devoted secondary journals; others do not. The New York Academy of Medicine keeps a current list of available secondary journals in many health care disciplines ([http://www.nyam.org/fellows-members/ebhc/eb\\_publications.html](http://www.nyam.org/fellows-members/ebhc/eb_publications.html)). If your specialty does not yet have its own journal, you can apply your own relevance and methodologic screening criteria to articles in your target specialty or subspecialty journals. When you have learned the skills, you will be surprised at the small proportion of studies to which you need attend and the efficiency with which you can identify them.

## Problem Solving

Experienced clinicians managing a patient with type 2 diabetes mellitus will ask questions such as “In patients with new-onset type 2 diabetes mellitus, which clinical features or test results predict the development of diabetic complications?” “In patients with type 2 diabetes mellitus requiring drug therapy, does starting with metformin treatment yield improved diabetes control and reduce long-term complications better than other initial treatments?” Here, clinicians are defining specific

questions raised in caring for patients and then consulting the literature to resolve these questions.

## Asking Background and Foreground Questions

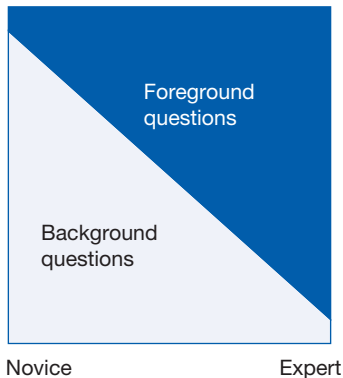
One can think of the first set of questions, those of the medical student, as background questions and of the browsing and problem-solving sets as foreground questions. In most situations, you need to understand the background thoroughly before it makes sense to address foreground issues.

Experienced clinicians may occasionally require background information when a new condition or medical *syndrome* (eg, Middle East respiratory syndrome coronavirus), a new diagnostic test (eg, molecular diagnosis), or a new treatment modality (eg, dipeptidyl peptidase 4 inhibitors) appears in their clinical arena.

Figure 3-2 represents the evolution of the questions we ask as we progress from being novices posing background questions to experts posing foreground questions. This book explores how clinicians can use the medical literature to solve their foreground questions.

**FIGURE 3-2**

### Background and Foreground Questions



## CLARIFYING YOUR QUESTION

---

### The Structure: Patients, Exposures, Outcome

Clinical questions often spring to mind in a form that makes finding answers in the medical literature a challenge. Dissecting the question into its component parts to facilitate finding the best evidence is a fundamental skill. One can divide questions of therapy or *harm* into 4 parts following the *PICO* framework: patients or population, intervention(s) or exposure(s), comparator, and outcome (Box 3-1). For questions of *prognosis*, you can use 1 of 2 alternative structures. One has only 3 elements: patients, exposure (time), and outcome. An alternative focuses on patient-related factors, such as age and sex, that can modify prognosis: patients, exposure (eg, older age or male), comparison (eg, younger age or female), and outcome. For diagnostic tests, the structure we suggest is patients, exposure (test), and outcome (*criterion standard*).<sup>6</sup>

### BOX 3-1

#### Framing Clinical Questions: PICO

**Patients or Population:** Who are the relevant patients?

**Intervention(s) or Exposure(s):** For example, diagnostic tests, foods, drugs, surgical procedures, time, or risk factors. What are the management strategies we are interested in comparing or the potentially harmful exposures about which we are concerned?

**Comparator:** For issues of therapy, prevention, or harm, there will always be both an experimental intervention or putative harmful exposure and a control, alternative, or comparison intervention.

**Outcome:** What are the patient-relevant consequences of the exposures in which we are interested? We may also be interested in the consequences to society, including cost or resource use. It may also be important to specify the period of interest.

## Five Types of Foreground Clinical Questions

In addition to clarifying the population, intervention or exposure, and outcome, it is productive to label the nature of the question that you are asking. There are 5 fundamental types of clinical questions:

1. Therapy: determining the effect of interventions on *patient-important outcomes* (*symptoms*, function, morbidity, mortality, and costs)
2. Harm: ascertaining the effects of potentially harmful agents (including therapies from the first type of question) on patient-important outcomes
3. *Differential diagnosis*: in patients with a particular clinical presentation, establishing the frequency of the underlying disorders
4. Diagnosis: establishing the *power* of a test to differentiate between those with and without a *target condition* or disease
5. Prognosis: estimating a patient's future course

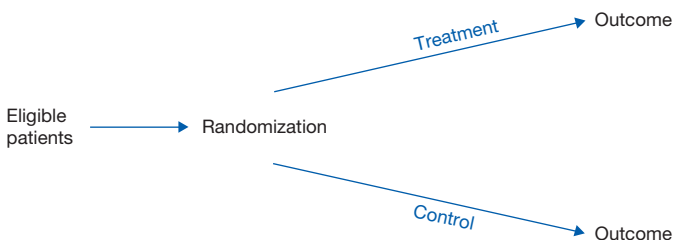
## Finding a Suitably Designed Study for Your Question Type

You need to correctly identify the category of study because, to answer your question, you must find an appropriately designed study. If you look for a *randomized trial* to inform the properties of a diagnostic test, you will not find the answer you seek. We will now review the study designs associated with the 5 major types of questions.

To answer questions about a therapeutic issue, we seek studies in which a process analogous to flipping a coin determines participants' receipt of an *experimental treatment* or a control or standard treatment: a randomized trial (see Chapter 6, Therapy [Randomized Trials]). Once investigators allocate participants to treatment or *control groups*, they follow them forward in time to determine whether they have, for instance, a stroke or

FIGURE 3-3

## Structure of Randomized Trials



myocardial infarction—what we call the outcome of interest (Figure 3-3). When randomized trials are not available, we look to *observational studies* in which—rather than randomization—clinician or patient preference, or happenstance, determines whether patients receive an intervention or alternative (see Chapter 5, Why Study Results Mislead: Bias and Random Error).

Ideally, we would also look to randomized trials to address issues of harm. For most potentially harmful exposures, however, randomly allocating patients is neither practical nor ethical. For instance, one cannot suggest to potential study participants that an investigator will decide by the flip of a coin whether or not they smoke during the next 20 years. For exposures such as smoking, the best one can do is identify observational studies (often subclassified as *cohort* or *case-control studies*) that provide less trustworthy evidence than randomized trials (see Chapter 10, Harm [Observational Studies]).

Figure 3-4 depicts a common observational study design in which patients with and without the exposures of interest are followed forward in time to determine whether they experience the outcome of interest. For smoking, an important outcome would likely be the development of cancer.

For sorting out differential diagnosis, we need a different study design (Figure 3-5). Here, investigators collect a group of patients with a similar presentation (eg, painless jaundice,

FIGURE 3-4

### Structure of Observational Cohort Studies

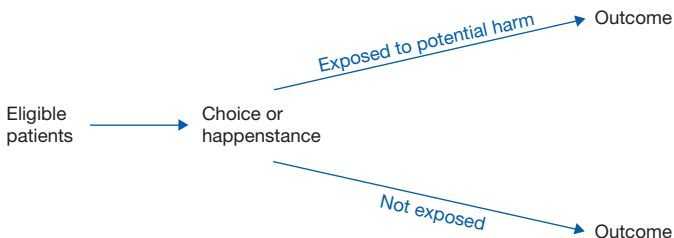
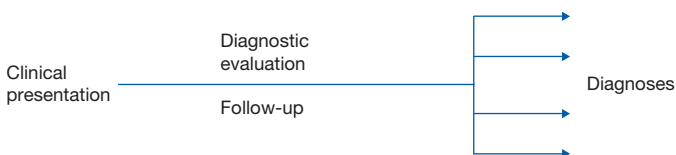


FIGURE 3-5

### Structure of Studies of Differential Diagnosis



syncope, or headache), conduct an extensive battery of tests, and, if necessary, follow patients forward in time. Ultimately, for each patient the investigators hope to establish the underlying cause of the symptoms and *signs* with which the patient presented.

Establishing the performance of a diagnostic test (ie, the test's properties or operating characteristics) requires a slightly different design (Figure 3-6). In diagnostic test studies, investigators identify a group of patients among whom they suspect a disease or condition of interest exists (such as tuberculosis, lung cancer, or iron deficiency anemia), which we call the target condition. These patients undergo the new diagnostic test and a *reference standard* (also referred to as *gold standard* or *criterion standard*). Investigators evaluate the diagnostic test by comparing its classification of patients with that of the reference standard (Figure 3-6).

A final type of study examines a patient's prognosis and may identify factors that modify that prognosis. Here, investigators identify patients who belong to a particular group (such as pregnant women, patients undergoing surgery, or patients with cancer) with or without factors that may modify their prognosis (such as age or *comorbidity*). The exposure here is time, and investigators follow up patients to determine whether they experience the *target outcome*, such as an adverse obstetric or neonatal event at the end of a pregnancy, a myocardial infarction after surgery, or survival in cancer (Figure 3-7).

FIGURE 3-6

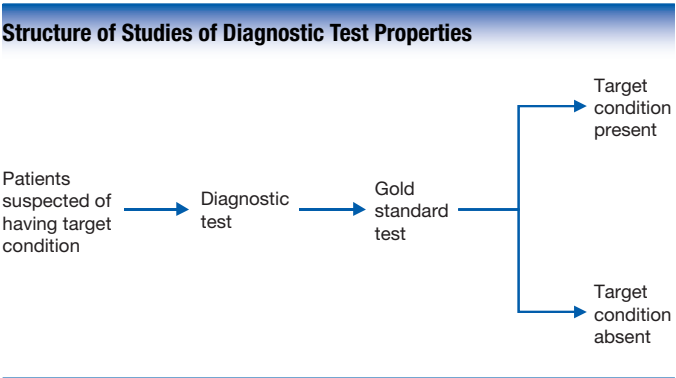
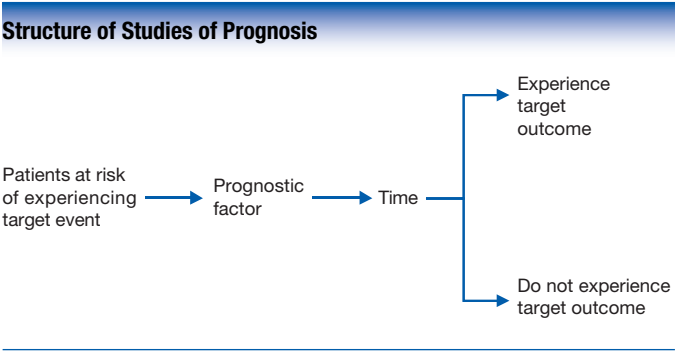


FIGURE 3-7



## Three Examples of Question Clarification

We will now provide examples of the transformation of unstructured clinical questions into the structured questions that facilitate the use of the medical literature.

### Example 1: Diabetes and Target Blood Pressure

A 55-year-old white woman presents with type 2 diabetes mellitus and hypertension. Her glycemic control is excellent with metformin, and she has no history of complications. To manage her hypertension, she takes a small daily dose of a thiazide diuretic. During a 6-month period, her blood pressure is near 155/88 mm Hg.

**Initial Question:** When treating hypertension, at what target blood pressure should we aim?

**Digging Deeper:** One limitation of this formulation of the question is that it fails to specify the population in adequate detail. The benefits of tight control of blood pressure may differ among patients with diabetes vs those without diabetes, in type 1 vs type 2 diabetes, and among patients with and without diabetic complications.

The detail in which we specify the patient population is a double-edged sword. On the one hand, being very specific (middle-aged women with uncomplicated type 2 diabetes) will ensure that the answer we get is applicable to our patient. We may, however, fail to find any studies that restrict themselves to this population. The solution is to start with a specific patient population but be ready to remove specifications to find a relevant article. In this case, we may be ready to remove the “female,” “middle-aged,” “uncomplicated,” and “type 2,” in that order. If we suspect that the optimal target blood pressure may be similar among patients



with and without diabetes, and if it proves absolutely necessary, we might remove “diabetes” from the question.

The order in which we remove the patient specifications depends on how likely it is that those characteristics will influence response to treatment. We suggest removing “female” first because we think it likely that optimal target blood pressure will be similar in men and women. Similarly, younger, middle-aged, and elderly individuals are likely to have the same optimal targets (although here we are not quite so sure). As our doubts about the same optimal targets across populations becomes progressively greater (uncomplicated vs complicated diabetes, type 1 vs type 2, or patients with diabetes vs those without), we become increasingly reluctant to remove the particular patient characteristic from the question.

We may wish to specify that we are interested in the addition of a specific antihypertensive agent. Alternatively, the intervention of interest may be any antihypertensive treatment. Furthermore, a key part of the intervention will be the target for blood pressure control. For instance, we might be interested in knowing whether it makes any difference if our target diastolic blood pressure is less than 80 mm Hg vs less than 90 mm Hg. Another limitation of the initial question formulation is that it fails to specify the criteria (the outcomes of interest) by which we will judge the appropriate target for our hypertensive treatment.

### Improved (Searchable) Question: A Question About Therapy

- *Patients:* Patients with hypertension and type 2 diabetes without diabetic complications.
- *Intervention/Exposure:* Any antihypertensive agent that aims at a target diastolic blood pressure of 90 mm Hg.

- *Comparator:* Target diastolic blood pressure of 80 mm Hg.
- *Outcomes:* Stroke, myocardial infarction, cardiovascular death, and total mortality.

### Example 2: Transient Loss of Consciousness

A previously well, although a heavy drinker, 55-year-old man presents to the emergency department with an episode of transient loss of consciousness. On the evening of presentation, he had his usual 5 beers and started to climb the stairs at bedtime. The next thing he remembers is being woken by his son, who found him lying near the bottom of the stairs. The patient took about a minute to regain consciousness and remained confused for another 2 minutes. His son did not witness any shaking, and there had not been any incontinence. Physical examination findings were unremarkable; the electrocardiogram revealed a sinus rhythm with a rate of 80/min and no abnormalities. Glucose, sodium, and other laboratory results were normal, and a blood alcohol test result was negative.

**Initial Question:** How extensively should I investigate this patient?

**Digging Deeper:** The initial question gives us little idea of where to look in the literature for an answer. As it turns out, there are a host of questions that could be helpful in choosing an optimal investigational strategy. We could, for instance, pose a question of differential diagnosis: If we knew the distribution of ultimate diagnoses in such patients, we could choose to investigate the more common and omit investigations targeted at remote possibilities.

Other information that would help us would be the properties of individual diagnostic tests. If an electroencephalogram were extremely accurate for diagnosing a

seizure or a 24-hour Holter monitor for diagnosing arrhythmia, we would be far more inclined to order these tests than if they missed patients with the underlying problems or falsely identified patients as not having the problems.

Alternatively, we could ask a question of prognosis. If patients had benign prognoses, we might be much less eager to investigate extensively than if patients tended to have poor outcomes. Finally, the ultimate answer to how intensively we should investigate might come from a randomized trial in which patients similar to this man were allocated to more vs less intensive investigation.

### Improved (Searchable) Questions: A Question About Differential Diagnosis

- *Patients*: Middle-aged patients presenting with transient loss of consciousness.
- *Intervention/Exposure*: Thorough investigation and follow-up for common and less common diagnoses.
- *Comparator*: Minimal investigation and follow-up.
- *Outcomes*: Frequency of underlying disorders, such as vasovagal syncope, seizure, arrhythmia, and transient ischemic attack.

### A Question About Diagnosis

- *Patients*: Middle-aged patients presenting with transient loss of consciousness.
- *Intervention/Exposure*: Electroencephalogram.
- *Outcomes*: Reference standard investigation (probably long-term follow-up).

### A Question About Prognosis

- *Patients*: Middle-aged patients presenting with transient loss of consciousness.

- *Exposure/Comparison:* Time.
- *Outcomes:* Morbidity (complicated arrhythmias or seizures, strokes, or serious accidents) and mortality in the year after presentation.

### A Question About Diagnostic Impact

You can think of this also as a question of therapy; the principles of critical appraisal are the same.

- *Patients:* Middle-aged patients presenting with loss of consciousness.
- *Intervention/Exposure:* Comprehensive investigation.
- *Comparator:* Minimal investigation.
- *Outcomes:* Morbidity and mortality in the year after presentation.

### Example 3: Squamous Cell Carcinoma

A 60-year-old man with a 40-pack-year smoking history presents with hemoptysis. A chest radiograph shows a parenchymal mass with a normal mediastinum, and a fine-needle aspiration and biopsy of the mass reveals non-small cell carcinoma. Aside from hemoptysis, the patient is asymptomatic, and the physical examination results are normal.

**Initial Question:** What investigations should we undertake before deciding whether to offer this patient surgery?

**Digging Deeper:** The key defining features of this patient are his non-small cell carcinoma and the fact that his medical history, physical examination, and chest radiograph indicate no evidence of intrathoracic or extrathoracic metastatic disease. Alternative investigational strategies address 2 issues: Does the patient have occult mediastinal disease, and does he have occult extrathoracic metastatic

disease? Investigational strategies for addressing the possibility of occult mediastinal disease include undertaking a mediastinoscopy or performing computed tomography (CT) of the chest and proceeding according to the results of this investigation. Investigational strategies for extrathoracic disease include brain and abdominal CT and bone scanning. Positron emission tomography–CT (PET-CT) represents an alternative approach for both intrathoracic and extrathoracic disease.

What outcomes are we trying to influence in our choice of investigational approach? We would like to prolong the patient's life, but the extent of his underlying tumor is likely to be the major determinant of survival, and our investigations cannot change that. We wish to detect occult mediastinal metastases if they are present because, if the cancer has spread, resectional surgery is unlikely to benefit the patient. Thus, in the presence of mediastinal metastatic disease, patients will usually receive palliative approaches and avoid an unnecessary thoracotomy.

We could frame our structured clinical question in 2 ways. One would be asking about the usefulness of the PET-CT scan for identifying metastatic disease. More definitive would be to ask a question of diagnostic impact, analogous to a therapy question: What investigational strategy would yield superior patient-important outcomes?

### **Improved (Searchable) Questions: A Question About Diagnosis**

- *Patients:* Newly diagnosed non–small cell lung cancer with no evidence of extrapulmonary metastases.
- *Intervention:* PET-CT scan of the chest.
- *Outcome:* Mediastinal spread at mediastinoscopy.

### A Question About Diagnostic Impact (Therapy)

- *Patients:* Newly diagnosed non–small cell lung cancer with no evidence of extrapulmonary metastases.
- *Intervention:* PET-CT.
- *Comparator:* Alternative diagnostic strategies.
- *Outcome:* Unnecessary thoracotomy.

## CONCLUSION: DEFINING THE QUESTION

---

Constructing a searchable and answerable question that allows you to use the medical literature to solve problems is no simple matter. It requires a detailed understanding of the clinical issues involved in patient management. The 3 examples in this chapter illustrate that each patient encounter may trigger a number of clinical questions and that you must give careful thought to what you really want to know. Bearing the structure of the question in mind—patient or population, intervention or exposure, outcome, and, for therapy or harm questions, comparison—is helpful in arriving at an answerable question. Identifying the type of questions—therapy, harm, differential diagnosis, diagnosis, and prognosis—will not only ensure you choose the right question structure but also ensure that you are looking for a study with an appropriate design.

Careful definition of the question will provide another benefit: you will be less likely to be misled by a study that addresses a question related to that in which you are interested, but with 1 or more important differences. For instance, making sure that the study compares experimental treatment to current optimal care may highlight the limitations of trials that use a *placebo* comparator rather than an alternative

active agent. Specifying that you are interested in patient-important outcomes (such as long bone fractures) identifies the limitations of studies that focus on *substitute* or *surrogate end points* (such as bone density). Specifying that you are primarily interested in avoiding progression to dialysis will make you appropriately wary of a *composite end point* of progression to dialysis or doubling of serum creatinine level. You will not reject such studies out of hand, but the careful definition of the question will help you to critically apply the results to your patient care.

A final crucial benefit from careful consideration of the question is that it sets the stage for efficient and effective literature searching to identify and retrieve the current best evidence (see Chapter 4, Finding Current Best Evidence). Specifying a structured question and identifying an appropriate study design to answer it will allow you to select and use searching resources efficiently and thus enhance your evidence-based practice.

## References

1. Wing RR, Bolin P, Brancati FL, et al; Look AHEAD Research Group. Cardiovascular effects of intensive lifestyle intervention in type 2 diabetes. *N Engl J Med*. 2013;369(2):145-154.
2. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Med*. 2010;7(9):e1000326.
3. McKibbin KA, Wilczynski NL, Haynes RB. What do evidence-based secondary journals tell us about the publication of clinically important articles in primary healthcare journals? *BMC Med*. 2004;2:33.
4. Haynes RB, Cotoi C, Holland J, et al; McMaster Premium Literature Service (PLUS) Project. Second-order peer review of the medical literature for clinical practitioners. *JAMA*. 2006;295(15):1801-1808.
5. Haynes RB. ACP Journal Club: the best new evidence for patient care. *ACP J Club*. 2008;148(3):2.
6. Agoritsas T, Merglen A, Courvoisier DS, et al. Sensitivity and predictive value of 15 PubMed search strategies to answer clinical questions rated against full systematic reviews. *J Med Internet Res*. 2012;14(3):e85.

4

# Finding Current Best Evidence

Thomas Agoritsas, Per Olav Vandvik, Ignacio Neumann, Bram Rochweg, Roman Jaeschke, Robert Hayward, Gordon Guyatt, and K. Ann McKibbin

## IN THIS CHAPTER

---

### Introduction

Searching for Evidence: A Clinical Skill

Start by Clarifying the Question

Searching the Medical Literature Is Sometimes Futile

### How Evidence Is Processed and Organized Into EBM Resources

Hierarchy of Evidence

Levels of Processing

Pyramid of EBM Resources

### Three Criteria for Choosing an EBM Resource

Based on Current Best Evidence

Coverage and Specificity

Availability and Access

(continued on following page)



## **Using the Pyramid of EBM Resources to Answer Your Questions**

Summaries and Guidelines

Preappraised Research

Alerts to Important New Evidence

Nonpreappraised Research

Searching All Levels of the Pyramid at the Same Time

When to Use Google

## **Translating a Question Into Search Terms**

How to Choose and Combine Search Terms

Broad vs Narrow Searches

Finding Related Articles

Getting Help

## **Conclusion: Improving Your Searching Skills in Daily Practice**

---

## INTRODUCTION

---

### Searching for Evidence: A Clinical Skill

Searching for current best evidence in the medical literature has become a central skill in clinical practice.<sup>1,2</sup> On average, clinicians have 5 to 8 questions about individual patients per daily shift<sup>3-5</sup> and regularly use online *evidence-based medicine* (EBM) resources to answer them.<sup>6-9</sup> Some now even consider that “the use of search engines is as essential as the stethoscope.”<sup>10</sup>

However, because of the increasing volume of new literature and speed of new research, finding useful evidence efficiently remains challenging. Approximately 2000 new articles are indexed in PubMed every day,<sup>10</sup> and although few of them directly inform clinical practice, as many as 75 are *randomized clinical trials* and 11 are *systematic reviews*.<sup>11</sup> These numbers explain why searching in PubMed is not the most efficient way to look for evidence-based answers. For example, when typing “stroke prevention in atrial fibrillation” in PubMed, you will see that current best evidence is literally lost in an output of almost 4000 citations, with a mix of trials, reviews, guidelines, and editorials that are impossible to screen for relevance during your daily practice.

Fortunately, numerous EBM resources now provide shorter and more efficient paths. These resources select, process, and organize the evidence; some, however, are more trustworthy than others. This chapter will help you navigate through existing EBM resources, distinguish the trustworthy from the less trustworthy, and maximize your chances of quickly finding answers based on current best evidence.

### Start by Clarifying the Question

As we have seen in Chapter 3, What Is the Question? framing the question appropriately is an important prerequisite to any search. An initial distinction to make is whether you are asking a *background question* (eg, definition or pathophysiology

of a syndrome or mechanism of a treatment modality) or a *foreground question* (eg, targeted questions of therapy, *harm*, diagnosis, or *prognosis* that provide the evidentiary basis for decision making). Although some EBM resources also answer background questions, this chapter, and the *Users' Guides to the Medical Literature* overall, focuses on efficiently finding answers to foreground questions.

Foreground questions often arise in a form that does not facilitate finding an answer (see Chapter 3, What Is the Question?). A first step is to translate and structure the question into its components, using the *PICO* framework, which accounts for the patient or population, the intervention or exposure, the comparator, and the outcomes (see Chapter 3, Box 3-1). When framing your question, remember to consider all *patient-important outcomes*. Doing so will guide you in selecting the body of evidence that adequately addresses your patient's dilemma between benefits and harms that matter to your patient's decision.

Structuring the question will not only clarify what you are looking for but also help you formulate relevant search terms and combine them into search strategies, adapted to each type of EBM resource. We explore, toward the end of this chapter (see Translating a Question Into Search Terms), how the issues of question formulation and choice of search strategies become particularly crucial when evidence is harder to find using pre-appraised resources and you need to search in larger databases, such as PubMed. Finally, clarifying your question will help you search for appropriate study designs (see Chapter 3, What Is the Question?) and select corresponding search filters (eg, Clinical Queries) to reduce the number of citations in search outputs and enhance your chances of finding the best relevant evidence.

### Searching the Medical Literature Is Sometimes Futile

Consider the following clinical question: "In patients with pulmonary embolism, to what extent do those with pulmonary

infarction have a poorer *health outcome* than those without pulmonary infarction?”

Before beginning your search to answer this question, you should think about how investigators would differentiate between those with and without infarction. Because there is no definitive method, short of autopsy, to make this differentiation, our literature search is doomed before we begin.

This example illustrates that the medical literature will not help you when no feasible study design or measurement tools exist that investigators could use to resolve an issue. Your search also will be futile if no one has conducted and published the necessary study. Before embarking on a search, carefully consider whether the yield is likely to be worth the time expended.

## HOW EVIDENCE IS PROCESSED AND ORGANIZED INTO EBM RESOURCES

---

Evidence-based medicine resources are rapidly evolving and provide innovative solutions to deal with the production, summary, and appraisal of the evidence.<sup>1</sup> Numerous EBM resources are currently available. To clearly see how to navigate across available resources, we offer 3 classification systems: (1) *hierarchy of evidence* in primary studies, (2) level of processing of the evidence, and (3) categories of EBM resources (Figure 4-1). Together, these 3 classification systems describe the flow of evidence from primary studies to existing EBM resources.

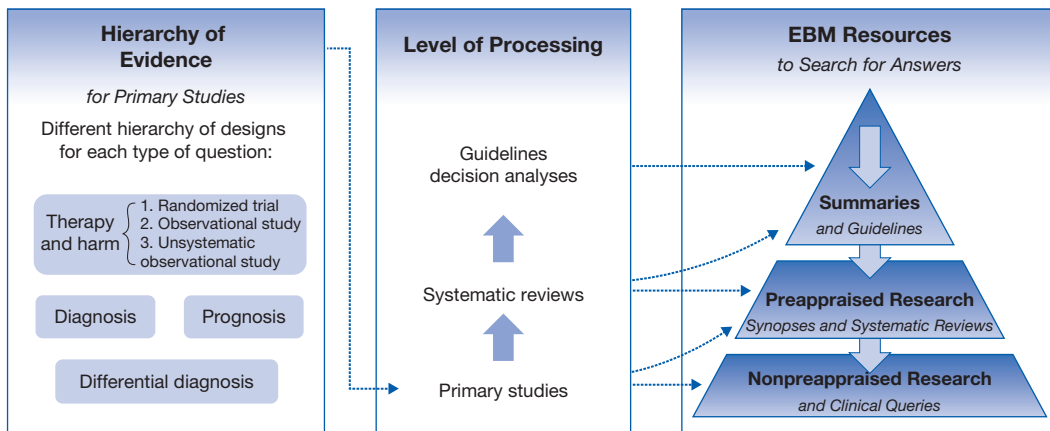
### Hierarchy of Evidence

At the level of *primary studies*, our first classification relates to the hierarchy of evidence (Figure 4-1, left box). For each type of question, EBM suggests a hierarchy of research designs to minimize the *risk of bias*. For questions regarding therapy or harm, well-conducted randomized clinical trials are superior



**FIGURE 4-1**

**From Evidence to Evidence-Based Resources**



to *observational studies*, which are superior to unsystematic clinical observations. Questions of diagnostic test properties, *differential diagnosis*, or prognosis require different hierarchies of study design (see Chapter 2, What Is Evidence-Based Medicine?).

Furthermore, within each type of design, some studies provide evidence of higher quality than others. The ideal EBM resource should facilitate access to studies with the most appropriate design and lowest risk of bias.

## Levels of Processing

A second classification refers to the level of processing of the evidence (Figure 4-1, middle box). Primary studies can stand alone or be processed into systematic reviews. On the basis of clear eligibility criteria, authors of a systematic review conduct a comprehensive search for all primary studies, critically appraise their quality, and, when it is considered appropriate, provide a summary estimate of effects across studies. Well-conducted systematic reviews are far more useful than single primary studies because they represent the entire body of relevant evidence (see Chapter 14, The Process of a Systematic Review and Meta-analysis). Searching for systematic reviews instead of primary studies will save you substantial time and effort.

A further level of processing is to move from evidence (ideally systematic reviews) to recommendations for practice, as in *clinical practice guidelines* (see Moving From Evidence to Action). Providing recommendations requires judging the relative desirability of alternative courses of action. Therefore, this level of processing requires looking at the entire body of evidence, integrating and appraising the evidence from systematic reviews for each patient-important outcome, taking into account patient *values and preferences*, and being mindful of resource considerations. *Decision analyses* (see Chapter 17, How to Use a Patient Management Recommendation: Clinical Practice Guidelines and Decision Analyses) and health technology assessment

reports also may provide a similar level of processing of the evidence. As for primary studies, some guidelines are more trustworthy than others, and the ideal EBM resources should provide access to the more trustworthy ones.

## Pyramid of EBM Resources

Although the 2 previous classifications—the hierarchy of evidence and level of processing—help you decide what type of evidence is likely to answer your question, they do not inform you of where to search for the evidence. For example, you may wonder where to search for high-quality systematic reviews. Should you start your search in the Cochrane Library, use review filters in PubMed, or look in the reference list of an online summary such as UpToDate? To make that choice, you need to understand how evidence is organized into a third classification: the *pyramid of EBM resources* (Figure 4-1, right box). From a practical perspective, resources can be viewed in 3 broad categories: summaries and guidelines, preappraised research, and nonpreappraised research.

Table 4-1 outlines these categories of EBM resources. Box 4-1 and the subsequent paragraph provide a fuller description of each category with examples of resources.

You can navigate efficiently across these different types of resources, as well as search all 3 categories simultaneously, using *federated search engines*, such as ACCESSSS (<http://plus.mcmaster.ca/accessss>), Trip (<http://www.tripdatabase.com>), Sum Search (<http://sumsearch.org>), or Epistemonikos (<http://www.epistemonikos.org>). Before we describe these search engines in detail, we will look at general criteria that will help clinicians choose which EBM resources to select given their question and which to avoid.

To complement resources that help you answer clinical questions, additional resources can link the evidence with your daily practice, such as *clinical decision support systems*<sup>15</sup> or context-specific access to online resources within electronic

**TABLE 4-1****Categories of EBM Resources**

<b>Category</b>	<b>Layers<sup>a</sup></b>	<b>Description</b>	<b>Examples</b>
Summaries and guidelines	Online summary resources Databases of clinical practice guidelines	Summary of the body of evidence at a topic-level (not limited to a question, intervention, or outcome) Often with actionable recommendations for clinical decision making Regularly updated	UpToDate DynaMed Clinical Evidence Best Practice US National Guidelines Clearinghouse
Preappraised research	Synopses of systematic reviews Systematic reviews Synopses of studies	Structured abstracts or 1-page summaries of selected systematic reviews or studies Various degrees of preappraisal – Selection according to methodologic criteria – Clinicians’ ratings – Clinicians’ comments – Experts’ structured appraisal Continuously updated Source of evidence alerts	ACP Journal Club McMaster <i>PLUS</i>  DARE Cochrane  Evidence Updates

*(Continued)*





TABLE 4-1

**Categories of EBM Resources (Continued)**

Category	Layers <sup>a</sup>	Description	Examples
Nonreappraised research	Filtered studies	All primary studies with no reappraisal	PubMed (MEDLINE) CINAHL CENTRAL
	Unfiltered studies	Automatic filtering of databases for specific study designs or clinical content	Filters: Clinical Queries in PubMed
Federated searches	All layers of resources searched at once	Search engines that retrieve evidence from summaries and reappraised and nonreappraised research, and organize the results accordingly	ACCESSSS Trip SumSearch Epistimonikos

Abbreviations: ACCESSSS, ACCess to Evidence-based Summaries, Synopses, Systematic Reviews and Studies; CENTRAL, Cochrane Central Register of Controlled Trials; CINAHL, Cumulative Index to Nursing and Allied Health Literature; DARE, Database of Abstracts of Reviews of Effects.

<sup>a</sup>These layers correspond to the 6-S pyramid from Haynes et al.<sup>1,2</sup>



**BOX 4-1****Overview of EBM Resources****1. Summaries and guidelines.**

Summaries are regularly updated online resources that aim to integrate the body of evidence at a topic level for several related questions. For example, a topic such as “treatment of type 2 diabetes mellitus in the elderly patient” will usually summarize evidence for drug therapy, strategies to control glycemic levels and avoid hypoglycemia, and lifestyle modification and the reduction of cardiovascular risk. These summaries often provide actionable recommendations for practice. Current examples widely used by clinicians include UpToDate (<http://www.uptodate.com>), DynaMed (<https://dynamed.ebscohost.com>), and Best Practice (<http://bestpractice.bmj.com>).

Guidelines follow a similar approach, usually focused on a specific topic or disease (eg, “antithrombotic therapy and prevention of thrombosis”<sup>12</sup>). Even more than summaries, guidelines are focused on providing recommendations for optimal patient management. Searching for available guidelines is more challenging because they are scattered across specialty journals and organization websites. A useful resource to search for guidelines is the US National Guideline Clearinghouse (<http://www.guideline.gov>), which includes guidelines from many countries.

**2. Preappraised research.**

When summaries or guidelines do not provide a satisfactory answer (eg, they provide an answer that is apparently not based on current best evidence or do not provide an answer at all), you must look directly at research findings, first from systematic reviews and then, if necessary, from primary studies. Many resources can prevent the unpleasant experience of searching the whole medical literature (at the risk of getting lost) or having to screen and read articles as PDFs. These resources select only

systematic reviews and studies that meet defined methodologic criteria and provide synopses—a 1-page structured abstract or description of reviews or studies. The degree and quality of pre-appraisal vary across resources. Some provide clinicians' ratings or short comments on relevance or newsworthiness, whereas others include a structured appraisal from experts. An example of the former is McMaster *PLUS* (Premium Literature Service<sup>13,14</sup>; <http://plus.mcmaster.ca/evidenceupdates>), and examples of the latter are *ACP Journal Club* (<http://acpjic.acponline.org>) and DARE (Database of Abstracts of Reviews of Effects; [www.crd.york.ac.uk/crdweb](http://www.crd.york.ac.uk/crdweb)). You can access preappraised research in 2 complementary ways: by searching these specific databases for a given question and, for some of them, by subscribing to an e-mail alerting system. Personalized alerts are an efficient way to remain up-to-date on important new research in your area of interest (see, for example, BMJ EvidenceUpdates; <http://plus.mcmaster.ca/evidenceupdates>).

### 3. Nonpreappraised research.

Only when other sources have failed to provide an answer should you search for primary studies in the larger databases, such as MEDLINE (<http://www.ncbi.nlm.nih.gov/pubmed>) or CINAHL (<http://www.cinahl.com>). Because these databases include millions of articles, using them efficiently requires more advanced searching skills. Limiting your search with filters, such as Clinical Queries (<http://www.ncbi.nlm.nih.gov/pubmed/clinical>), provides a useful way to reduce the number of abstracts you need to review to identify the best evidence to address your clinical question.

medical records.<sup>16</sup> However, although some clinical decision support systems have the potential to improve processes of care or patient outcomes,<sup>17</sup> most cover only a limited range of clinical problems, are not necessarily based on current best evidence, and are often “homebuilt” so that their use is questionable.<sup>1</sup>

## THREE CRITERIA FOR CHOOSING AN EBM RESOURCE

All EBM resources are not equally trustworthy, and none provide answers to all questions. Efficient searching involves choosing the appropriate resources for your clinical question—in much the same way you choose diagnostic tests appropriate for your patient’s symptoms. [Table 4-2](#) offers an initial guideline for making resource choices.

### Based on Current Best Evidence

Many online summaries and guideline websites promote themselves as “evidence-based,” but few have explicit links to research

**TABLE 4-2**

**Selection Criteria for Choosing or Evaluating EBM Resources**

Criterion	Description of Criterion
On the basis of current best evidence	How strong is the commitment to evidence to support inference?
	Does it have citations to references to all evidence summaries and recommendations?
	Is the process for keeping it up-to-date transparent and trustworthy?
	Is the quality of the evidence assessed?
	Is the strength of recommendations reported?
Coverage and specificity	Are numerical effect estimates reported for patient important outcomes?
	Does the resource cover my discipline and specific area of practice adequately?
Availability and access	Does it cover questions of the type I am asking (eg, therapy, diagnosis, prognosis, harm)?
	Is it readily available in all locations in which I would use it?
	Can I easily afford it?

findings. To judge the strength of the commitment to evidence to support inference, check whether you can distinguish statements that are based on high-quality vs low-quality evidence. If you cannot make this distinction, dismiss the resource altogether. Resources should provide citations to references to relevant research findings. Currency is important, and a simple way to judge whether the evidence is up-to-date is to look for the date of the most recent reference cited: if it is more than 2 years old, it is possible that future studies lead to a different conclusion.<sup>1,18,19</sup> Generally, the process for keeping a resource up-to-date should be transparent and trustworthy. A date stamp should accompany each summarized topic or piece of evidence (eg, "This topic last updated: Sep 17, 2013"), along with access to the explicit mechanism used to screen for related new findings. An opaque process should raise a red flag that the evidence may be partial, biased, or already outdated.

A summary or guideline should use a rating system to assess the risk of bias of cited studies and the quality of reviews. Resources that provide recommendations should be based on the entire body of existing evidence, ideally summarized in systematic reviews, and provide the benefits and harms of available options. The resources also should use an appropriate system to grade strength of recommendations and provide explicit judgments concerning underlying values and preferences (see *Moving From Evidence to Action*). Finally, to be actionable, the recommendations should report numerical effect estimates for patient-important outcomes to support clinical judgment and shared decision making at the point of care. For example, the ninth edition of the *Antithrombotic Therapy and Prevention of Thrombosis* guideline issued a weak recommendation for aspirin for primary prevention of cardiovascular events in people older than 50 years, based on moderate confidence in estimates of effect (grade 2B).<sup>20</sup> The authors provide numerical estimates: for example, in people at moderate risk of cardiovascular events, prophylactic aspirin resulted in 19 fewer myocardial infarctions per 1000 (from 26 fewer to 12 fewer) but 16 more major extra-cranial bleeds per 1000 (from 7 more to 20 more).

## Coverage and Specificity

An ideal resource will cover most of the questions relevant to your practice—and not much more. However, few, if any, resources are sufficient as such a one-stop shop for the evidence you need,<sup>18</sup> and resources from the 3 levels of the pyramid of EBM resources are often complementary. The higher you look in the pyramid, the more time it takes for the resource developers to process and summarize the evidence at a topic level, making these resources potentially out of date. To be comprehensive in your searching, you will need to look for preappraised research for more recent evidence. Conversely, the lower you look in the pyramid, the larger, and often less specific, the resource. Thus, preappraised research limited to your area of practice, such as collections of synopses designed to help you keep up with information on the latest developments in a specific field or specialty—eg, *Evidence-Based Mental Health* (<http://ebmh.bmj.com>) or *Evidence-Based Nursing* (<http://ebn.bmj.com>)—may serve your needs efficiently.

The type of question also will affect your choice of a specific resource. For example, resources that focus on management issues informed mainly by randomized clinical trials, such as the Cochrane Database of Systematic Reviews, may not be ideal to answer questions of harm or rare adverse events. Similarly, background questions are more likely to be answered by summaries (eg, UpToDate or DynaMed) than preappraised research (eg, systematic reviews or synopses). For example, if you have background questions about the Middle East respiratory syndrome coronavirus, both UpToDate and DynaMed have a dedicated entry on the topic that summarizes its case definition and the incidence of recent clusters.

## Availability and Access

The most trustworthy and efficient resources are frequently expensive, particularly those at the top of the pyramid of EBM resources. For example, an individual subscription to an online

summary often costs more than \$250 annually. To establish your information resource regimen, you can map the EBM resources that are accessible to you through your university, school, or clinical institution and check whether they meet your information needs. Academic clinicians typically have access to the resources of their academic institution or hospital libraries, including the full texts of many studies and reviews.

Clinicians in private practice in high-income countries may have access to some resources through their professional associations but otherwise may be burdened by the cost of subscriptions. Some countries have national libraries that centralize access to many resources. Often, the institutional choice of resources is not made by practicing clinicians and may be guided by financial constraints. If an important resource is not available, make the case for it to your librarian (and suggest which other resources are less useful in practice).<sup>1</sup> If your institution is not willing to pay a license, consider subscribing individually. Health professionals in lower-income countries may have institutional access to information resources through the World Health Organization's Health InterNetwork Access to Research Initiative (<http://www.who.int/hinari/en>) or other organizations but otherwise face even greater financial obstacles to information resources. Additional strategies include seeking open-access journals, writing to authors for a reprint or e-print of their article, and contacting colleagues in academic centers who have access to more extensive library facilities.

Preappraised resources are sometimes expensive as well, and therefore we further describe how searching federated search engines, such as ACCESSSS or Trip, can give you an overview of the clinical content of various resources to help you make subscription decisions.

Free e-mail systems, such as BMJ EvidenceUpdates (<http://plus.mcmaster.ca/evidenceupdates>), can alert you to important new findings, although access to full texts will vary according to your institutional or personal licenses. An increasing number of full-text articles are accessible through PubMed or Google

Scholar or directly via open-access journals (eg, *CMAJ*, PLOS journals, and BioMed Central journals; see <http://www.doaj.org> for a directory of open-access journals). Many other journals provide free access to full-text articles 6 to 12 months after publication (eg, *BMJ*, *JAMA*, and *Mayo Clinic Proceedings*) or a portion of their content at the time of publication. However, focusing on free full-text articles and free Internet resources may give a partial and potentially biased view of current best evidence.<sup>21</sup>

Finally, ask your institution or professional organization how to access EBM resources at the point of care and obtain proxy server permission or remote access at home (eg, using a VPN connection). This will give you direct access to evidence on your smartphone and tablets and considerably enhance your *evidence-based practice*.

## USING THE PYRAMID OF EBM RESOURCES TO ANSWER YOUR QUESTIONS

---

Numerous EBM resources are available, including many providers of summaries at the top level of the pyramid. Each has a different clinical scope, as well as different methodologic and editorial processes. No single portal lists them all, but many can be found through the New York Academy of Medicine ([http://www.nyam.org/fellows-members/ebhc/eb\\_resources.html](http://www.nyam.org/fellows-members/ebhc/eb_resources.html)) or the Cochrane Collaboration (<http://www.cochrane.org/about-us/webliography-evidence-based-health-care-resources>) websites.

It is beyond the scope of this chapter to discuss the pros and cons of each resource. Instead, we will focus on how to navigate across the pyramid of EBM resources and discuss how these resources can complement each other. We provide examples of resources to illustrate important aspects both from research on evidence retrieval and from our own practice but do not aim to be comprehensive or prescriptive on which resource to use.



## Summaries and Guidelines

Start your searches by using resources at the top of the pyramid for summaries and guidelines that address your question. These resources can provide a comprehensive view of the body of evidence at a topic level. Imagine, for example, that you are looking for antithrombotic therapies most appropriate for prevention of stroke in patients with atrial fibrillation. Available options include aspirin; other antiplatelet agents, such as clopidogrel; a combination of aspirin plus other antiaggregants; warfarin; or new anticoagulants, such as direct thrombin inhibitors or factor Xa inhibitors. To fully address your question from lower levels of the pyramid, you would need to retrieve, read, and integrate several systematic reviews or trials that cover all of the relevant comparisons and important outcomes. Summaries and guidelines aim to integrate this body of evidence and also often provide actionable recommendations for practice.

Table 4-3 lists examples of 10 widely used online summaries and their corresponding URLs. A recent analytical survey compared them on 3 aspects: the timeliness of updates, coverage of clinical topics, and quality of processing and reporting of the evidence.<sup>19</sup> At the time of this assessment (2011), the mean time since update ranged from 3.5 months (DynaMed) to 29 months (First Consult), and the percentage of clinical topics covered ranged from 25% (Clinical Evidence) to 83% (UpToDate). Quality substantially varied across the resources. For example, despite its limited coverage, the authors rated Clinical Evidence as the highest-quality resource. Because EBM resources continuously evolve, these numbers may be outdated but illustrate that online summaries can be complementary. Summaries also differ on their methods and commitment to providing actionable recommendations (eg, UpToDate now formulates recommendations using the *GRADE [Grading of Recommendations Assessment, Development and Evaluation]* framework, whereas

**TABLE 4-3****Rank Orderings of 10 Online Summaries Compared on 3 Aspects<sup>19</sup>**

Summary Resource	URL	Updates	Coverage, No. (%)	Quality
DynaMed	<a href="https://dynamed.ebscohost.com">https://dynamed.ebscohost.com</a>	1	3 (70)	2
UpToDate	<a href="http://www.uptodate.com">http://www.uptodate.com</a>	5	1 (83)	2
Micromedex	<a href="http://www.micromedex.com">http://www.micromedex.com</a>	2	8 (47)	2
Best Practice	<a href="http://bestpractice.bmj.com">http://bestpractice.bmj.com</a>	3	4 (63)	7
Essential Evidence Plus	<a href="http://www.essentialevidenceplus.com">http://www.essentialevidenceplus.com</a>	7	7 (48)	2
First Consult	<a href="http://www.firstconsult.com">http://www.firstconsult.com</a>	9	5 (60)	2
Medscape Reference	<a href="http://reference.medscape.com">http://reference.medscape.com</a>	6	2 (82)	9
Clinical Evidence	<a href="http://clinicalevidence.bmj.com">http://clinicalevidence.bmj.com</a>	8	10 (25)	1
ACP PIER	<a href="http://acpjc.acponline.org">http://acpjc.acponline.org</a>	4	9 (33)	7
PEPID	<a href="http://www.pepidonline.com">http://www.pepidonline.com</a>	NA	6 (58)	10

Abbreviation: NA, data not available.

Reproduced with permission from the *Journal of Clinical Epidemiology*.<sup>19</sup>



Clinical Evidence focuses more on the summary of evidence, also using GRADE) and their editorial style (eg, structured bullet points in DynaMed and Best Practice vs textbook-like structured chapters in UpToDate).

Unlike summaries, most guidelines are scattered across journals or websites from individual countries or health organizations. One of the most comprehensive portals to search for guidelines is the US National Guideline Clearinghouse (<http://www.guideline.gov>). It includes the full text of many US guidelines and thousands of international guidelines. Searching is easy, although initial retrievals are often relatively large. Other international guidelines can be found through the UK National Institute for Health and Care Excellence (<https://www.evidence.nhs.uk>) or the Guideline International Network (<http://www.g-i-n.net/library/international-guidelines-library>).

Perhaps even more than other types of preappraised evidence, practice guidelines are extremely variable in their trustworthiness.<sup>22,23</sup> When you conduct your search, look for guidelines that are transparent in how they process the evidence and formulate recommendations (see Chapter 17, How to Use a Patient Management Recommendation: Clinical Practice Guidelines and Decision Analyses). The US National Guideline Clearinghouse website also allows side-by-side comparisons of the guideline process and components for guidelines on the same topic.

Finally, the top of the EBM pyramid also includes decision analyses, which process a body of evidence in a similar way to guidelines, map out the options with outcomes and probabilities, and help you judge the benefits and harms of different treatment options for a specific patient (see Chapter 17, How to Use a Patient Management Recommendation: Clinical Practice Guidelines and Decision Analyses). These decision analyses often can be found in stand-alone studies, *economic evaluation* reports, and health technology assessment reports. An efficient way to locate decision analyses is through the Centre for Reviews and Dissemination at the UK University of York (<http://www.crd.york.ac.uk/crdweb>) by selecting the search filters “HTA” and “NHS EED” (for economic evaluation).

## Preappraised Research

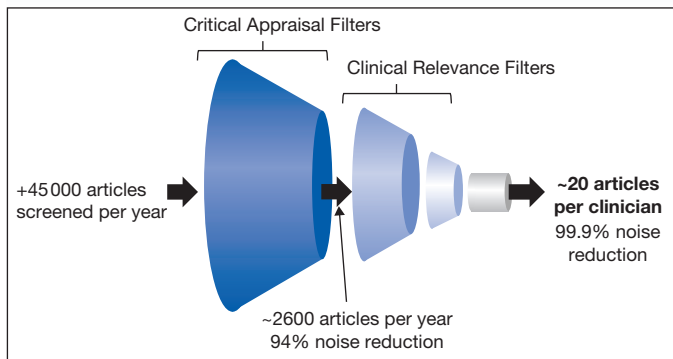
If you do not find a satisfactory answer in summaries or guidelines, either because your question is not covered or because you have reasons to doubt what you found, you may need to look for preappraised research. You also might search preappraised research to look for more recent evidence published since the summary or guideline was last updated.<sup>24</sup> You might wonder how often this additional searching would be worth the trouble. A recent study of the quality of online summaries found that, on average, new high-quality evidence providing potentially different conclusions than existing summaries was available for approximately 52% of the topics evaluated in UpToDate, 60% in Best Practice, and 23% in DynaMed.<sup>18</sup> This potential discrepancy between newly published evidence and existing recommendations would occur more frequently, and likely with greater adverse consequences, for most clinical practice guidelines, which are usually updated every 2 to 8 years.<sup>25</sup>

Consider, for example, the question of whether cardiac resynchronization therapy (CRT) reduces mortality in patients with heart failure and a narrow QRS complex. An initial search in mid-September 2013 in DynaMed or UpToDate provided an excellent summary of available evidence on the efficacy of CRT according to the degree of heart failure and the QRS duration but did not yet identify a more recent trial published in the *New England Journal of Medicine*.<sup>26</sup> This trial found that CRT did not reduce the composite rate of death or hospitalization for heart failure and actually may increase mortality. This important new evidence will of course be included in subsequent updates, but this process typically takes a couple of months to up to 29 months, depending on the online summary.<sup>19</sup>

A quick and efficient way to find preappraised research is to search specific databases, which include only studies and reviews that are more likely to be methodologically sound and clinically relevant. [Figure 4-2](#) shows a typical example of this improved selection process from McMaster *PLUS* (Premium

FIGURE 4-2

## Example of Preappraised Research: McMaster PLUS



Reproduced with permission of the Health Information Research Unit, McMaster University.

LiteratUre Service), a large database created by the McMaster Health Knowledge Refinery ([http://hiru.mcmaster.ca/hiru/HIRU\\_McMaster\\_PLUS\\_Projects.aspx](http://hiru.mcmaster.ca/hiru/HIRU_McMaster_PLUS_Projects.aspx)). The selection process used is as follows: trained research staff continually critically appraise more than 45 000 articles per year, from more than 125 empirically selected, high-quality clinical journals, and identify studies and systematic reviews that meet prespecified methodologic standards. For example, studies of prevention or therapy must have random allocation, a follow-up rate of at least 80%, and at least 1 patient-important outcome. These selected articles are then rated for relevance and newsworthiness by frontline clinicians from around the globe.<sup>27</sup> McMaster PLUS is thus a continuously updated database of more than 32 000 highly selective articles (with approximately 3300 added every year) that also feeds several other EBM resources and journals (eg, *ACP Journal Club*, *Clinical Evidence*, and *DynaMed*). A simple way to access McMaster PLUS is through the free search

engine of BMJ EvidenceUpdates (<http://plus.mcmaster.ca/EvidenceUpdates/QuickSearch.aspx>) or through the McMaster search engine, ACCESSSS, which we discuss further below (see Searching All Levels of the Pyramid at the Same Time). McMaster PLUS also has distinct databases for nursing (<http://plus.mcmaster.ca/np>) and rehabilitation studies (<http://plus.mcmaster.ca/rehab>).

In a further level of preappraisal, the more clinically relevant studies and systematic reviews are selected to become *synopses* (<1% of the initial selection). These synopses are usually a 1-page, structured summary of the research findings, along with a brief commentary from an expert in the field. You can find various types of synopses in specialized evidence-based *secondary evidence-based journals*. Figure 4-3 shows an example of a synopsis of a systematic review from *ACP Journal Club* (<http://acpjic.acponline.org>) on the impact of eplerenone on mortality compared with other aldosterone antagonists in heart failure. The abstract summarizes salient elements of the methods and results and an expert provides a commentary. This appraisal is not always systematic or as thorough as a full critical appraisal, but it usually provides the gist of the strengths and weaknesses of a study. Similar resources include *Evidence-Based Medicine* (<http://ebm.bmj.com>), *Evidence-Based Mental Health* (<http://ebmh.bmj.com>), *Evidence-based Oncology* ([www.sciencedirect.com/science/journal/13634054](http://www.sciencedirect.com/science/journal/13634054)), or POEMs (Patient-Oriented Evidence that Matters) ([www.essentialevidenceplus.com/content/poems](http://www.essentialevidenceplus.com/content/poems)). The New York Academy of Medicine keeps a current list of specialized EBM journals in many health care disciplines ([www.nyam.org/fellows-members/ebhc/eb\\_publications.html](http://www.nyam.org/fellows-members/ebhc/eb_publications.html)).

When searching preappraised research, make synopses of systematic reviews your first priority because they summarize the body of evidence on a question. In addition to evidence-based journals, you can find synopses of systematic reviews in DARE (Database of Abstracts of Reviews of Effects) (<http://www.cochrane.org/editorial-and-publishing-policy-resource/database-abstracts-reviews-effects-dare>). If no synopses answer

## FIGURE 4-3

Example of Synopsis of a Systematic Review From *ACP Journal Club*

## Therapeutics

## Review: Eplerenone is not more effective for reducing mortality than other aldosterone antagonists

Chatterjee S, Meoller C, Shah N, et al. Eplerenone is not superior to older and less expensive aldosterone antagonists. *Am J Med.* 2012; 125:817-25.

Clinical impact ratings: ● ★★★★★ ☆ ● ★★★★★ ☆

## Question

In patients with left ventricular (LV) dysfunction, what is the relative efficacy of eplerenone and other aldosterone antagonists (AAs)?

## Review scope

Included studies compared eplerenone or other AAs with control (placebo, angiotensin-converting enzyme inhibitor, angiotensin-receptor blocker, or  $\beta$ -blocker) in patients > 18 years of age with symptomatic or asymptomatic LV dysfunction, had  $\geq$  8 weeks of follow-up, and reported  $\geq$  1 outcome of interest. Studies comparing AAs with each other were excluded. Outcomes were all-cause mortality, cardiovascular (CV) mortality, gynecostasia [per trial definition in individual studies], and hyperkalemia (serum potassium > 5.5 mEq/L).<sup>1</sup>

## Review methods

MEDLINE, EMBASE/Excerpta Medica, CINAHL, and Cochrane Central Register of Controlled Trials (all to Jul 2011); reference lists; and reviews were searched for randomized controlled trials (RCTs). 16 RCTs ( $n = 12,505$ , mean age 55 to 69 y, 54% to 87% men) met selection criteria. 4 RCTs included patients after acute myocardial infarction LV dysfunction, and 12 included patients with heart failure. Study drugs were spironolactone (10 RCTs), canrenone (3 RCTs), and eplerenone (3 RCTs). Risk for bias (Cochrane criteria) was low for 8 RCTs, intermediate for 7, and high for 1.

## Main results

Eplerenone and other AAs reduced all-cause mortality and CV mortality compared with no AA (Table). Eplerenone increased risk for hyperkalemia, and other AAs increased risk for gynecostasia, compared with no AA (Table). Based on an indirect comparison, other AAs reduced mortality more than eplerenone ( $P = 0.009$ ).

## Conclusion

Based on an indirect comparison, eplerenone is not more effective for reducing mortality in adults with left ventricular dysfunction than other aldosterone antagonists.

<sup>1</sup>Information provided by author.

Source of funding: No external funding.

For correspondence: Dr. S. Chatterjee, Maimonides Medical Center, Brooklyn, NY, USA. E-mail sauravchatterjee@gmail.com. ■

## Commentary

In their thorough review of the use of AAs in systolic heart failure, Chatterjee and colleagues conclude that data are insufficient to recommend eplerenone over spironolactone. Only 3 large outcome trials actually address the issue: RALES, assessing spironolactone (1), and EPHEsus (2) and EMPHASIS-HF (3), assessing eplerenone. Although the populations evaluated in each study were quite different, the relative reductions in mortality were similar (25%, 14%, and 19%, respectively). Indirect comparisons of drug efficacy across clinical trials with different patient populations and study protocols are challenging. Without head-to-head trials of AAs, we should not draw conclusions about their relative efficacy.

Chatterjee and colleagues confirm that spironolactone increases risk for gynecostasia. Hyperkalemia is a known adverse effect of any AA, although potassium increases were "not clinically important" in RALES (1). After RALES was published, however, there was a marked increase in the number of spironolactone prescriptions, with an increase in hyperkalemia and associated mortality (4). Gynecostasia can be distressing to male patients, but hyperkalemia may be fatal to either sex.

A strict, evidence-based practitioner would base drug and dosage selection on the clinical trial most closely matching a patient's presentation. While waiting for a definitive head-to-head trial—noting that benefits seem similar in the studied populations—I start with the less expensive spironolactone, switching to eplerenone if troublesome sexual adverse effects develop (while closely monitoring potassium).

Ellis Luder, MD, FACC

Mid Valley Cardiology, New York University School of Medicine  
Kingston, New York, USA

## References

- Pitt B, Zannad F, Remme WJ, et al. The effect of spironolactone on morbidity and mortality in patients with severe heart failure. Randomized Aldactone Evaluation Study Investigators. *N Engl J Med.* 1999;341:709-17.
- Pitt B, Remme W, Zannad F, et al; Eplerenone Post-Acute Myocardial Infarction Heart Failure Efficacy and Survival Study Investigators. Eplerenone, a selective aldosterone blocker, in patients with left ventricular dysfunction after myocardial infarction. *N Engl J Med.* 2003;348:1309-21.
- Zannad F, McMurray JJ, Krum H, et al; EMPHASIS-HF Study Group. Eplerenone in patients with systolic heart failure and mild symptoms. *N Engl J Med.* 2011;364:11-21.
- Junttila DN, Mandani MM, Lee DS, et al. Rates of hyperkalemia after publication of the Randomized Aldactone Evaluation Study. *N Engl J Med.* 2004;351:543-51.

Eplerenone or other AAs vs control in patients with left ventricular dysfunction<sup>1</sup>

Outcomes	Number of trials (n)	Weighted event rates		At 2 to 24 mo	
		Eplerenone	Control <sup>2</sup>	RRR (95% CI)	NNT (CI)
All-cause mortality	2 (9368)	14%	16%	15% (1 to 23)	41 (27 to 88)
CV mortality	2 (9368)	12%	14%	17% (8 to 25)	42 (29 to 88)
Gynecostasia	2 (9361)	0.49%	0.60%	26% (-27 to 57)	NS
				RRR (CI)	NNH (CI)
Hyperkalemia	3 (9489)	6.1%	3.8%	72% (39 to 147)	37 (19 to 140)
				Other AAs	RRR (95% CI)
All-cause mortality	12 (3560)	19%	25%	25% (17 to 34)	16 (12 to 24)
CV mortality	4 (2553)	26%	34%	25% (16 to 33)	12 (9 to 19)
				RRR (CI)	NNH (CI)
Gynecostasia	6 (2278)	5.4%	0.86%	52% (28 to 1057)	23 (11 to 49)
Hyperkalemia	10 (3342)	8.1%	4.5%	80% (-17 to 291)	NS

AA = aldosterone antagonist; CV = cardiovascular; NS = not significant; other abbreviations defined in Glossary. Weighted event rates, RRR, RRI, NNT, NNH, and CI calculated from overall event rates and risk ratios in article using a random-effects model.

<sup>1</sup>Placebo, angiotensin-converting enzyme inhibitor, angiotensin-receptor blocker, or  $\beta$ -blocker.

<sup>2</sup>Other AAs were spironolactone or canrenone.

your question, move to a direct search for other systematic reviews. A useful resource is the Cochrane Library (<http://www.thecochranelibrary.com>).

Regardless of the resources you use, remember that preappraisal and the collection of these synopses can only increase the likelihood of finding sound evidence efficiently. It does not guarantee it. You should also apply your own critical appraisal to the research findings that are summarized, as explained throughout the *Users' Guides to the Medical Literature*.

## Alerts to Important New Evidence

In addition to building continuously updated databases of preappraised research, an increasing number of resources offer e-mail *alerting services*. To make the volume of new evidence manageable, these alerts are usually tailored to your information needs when you register (eg, clinical disciplines, quality choices, and frequency of alerts).

For example, the whole process leading to McMaster *PLUS*, including clinicians' ratings for relevance and newsworthiness, results in up to a 99.9% noise (non-clinically relevant) reduction and produces a manageable stream of approximately 20 to 50 key articles per year in a clinical area that may influence your practice (Figure 4-2).<sup>28</sup> You can receive these alerts by subscribing to BMJ EvidenceUpdates or ACCESSSS. Several other free or fee-based alerting systems are available for both a wide scope of disciplines (eg, NEJM Journal Watch, <http://www.jwatch.org>) and specific subspecialties (eg, OrthoEvidence, <http://www.myorthovidence.com>). When using any of these alerting resources, check whether their process of selecting and appraising the evidence is explicit, trustworthy, and meeting your needs.

## Nonpreappraised Research

Only when summaries, guidelines, and preappraised research have failed to provide an answer should you search among the



tens of millions of nonreappraised research articles. They are stored in many different databases (the ones usually searched in systematic reviews), such as PubMed's MEDLINE, EMBASE, CINAHL, or Web of Science. These databases can be accessed directly or through different search engines. Some search engine companies, such as Ovid (<http://www.ovid.com>), are designed to facilitate complex search strategies, such as those done by medical librarians or authors of systematic reviews. For clinical purposes, PubMed is the most popular search engine, providing free access to the entire MEDLINE database (<http://www.ncbi.nlm.nih.gov/pubmed>).

Consider, for example, the question of whether statins can prevent dementia. Summaries and reappraised research provide limited or selected evidence to answer that question. Because of its volume, searching PubMed to find relevant evidence requires more advanced searching skills, particularly in the choice and combination of search terms. Simple searches typically yield large outputs with few easily identified relevant studies in the first pages.

To limit irrelevant studies in the outputs, use methodologic filters, such as Clinical Queries. As shown in [Figure 4-4](#), instead of typing your search terms on the main page of PubMed, select Clinical Queries or go directly to <http://www.ncbi.nlm.nih.gov/pubmed/clinical>. Empirically validated "methods" search terms are added to your search, according to your type of question. For example, [Table 4-4](#) lists the filters used for questions of therapy that facilitate the retrieval of randomized clinical trials.<sup>29</sup> Two filters are available for each search category, 1 broad (sensitive) and 1 narrow (specific), the latter being more adapted to clinical practice. Use of a filter will increase the proportion of relevant studies from approximately 2% to 30% in the first 2 pages of PubMed's output (first 40 citations).<sup>2</sup> Similar filters are available for questions of diagnosis, etiology, prognosis, and *clinical prediction rules*.

[Table 4-5](#) lists similar broad and narrow filters to find systematic reviews from PubMed.<sup>30</sup> In contrast with Clinical Queries,



FIGURE 4-4

### Clinical Queries in PubMed: Accessing From Main Page and Choosing of Filter (Category and Scope)

The screenshot displays the PubMed Clinical Queries interface. At the top, there is a search bar with 'PubMed' entered and a search button. Below the search bar, the 'PubMed Clinical Queries' section is visible, showing results for 'mucolytics COPD'. The 'Clinical Study Categories' section has 'Therapy' selected for Category and 'Broad' for Scope. A blue arrow points to the 'Clinical Queries' link in the 'PubMed Tools' section.

Reproduced with permission of the US National Library of Medicine and PubMed.



TABLE 4-4

Clinical Queries "Therapy" Filter: Performance and Strategy Used<sup>a</sup>

	Sensitivity, %	Specificity, %	PubMed Equivalent
Broad filter	99	70	((clinical[Title/Abstract] AND trial[Title/Abstract]) OR clinical trials[MeSH Terms] OR clinical trial[Publication Type] OR random*[Title/Abstract] OR random allocation[MeSH Terms] OR therapeutic use[MeSH Subheading])
Narrow filter	93	97	(randomized controlled trial[Publication Type] OR (randomized[Title/Abstract] AND controlled[Title/ Abstract] AND trial[Title/ Abstract]))

Abbreviation: MeSH, medical subject headings.

<sup>a</sup>Reproduced with permission of the US National Library of Medicine and PubMed.

these filters are not implemented in PubMed; the search strategy needs to be copied and pasted right after your search. Going back to our example of the search phrase "statins for the prevention of dementia," an unfiltered search retrieves hundreds of citations that cannot be reliably screened in clinical practice. When adding the narrow filter of Table 4-5 to your search, the output shrinks to 19 citations (in October 2013), and a quick review will identify 6 systematic reviews, including 1 Cochrane Review, updated in 2009, and the most recent review, published in *Mayo Clinic Proceedings* in September 2013, Statins and Cognition: A Systematic Review and Meta-analysis of Short- and Long-term Cognitive Effects. The University of York keeps a comprehensive list of available filters and the publications that describe

TABLE 4-5

Filters to Retrieve Systematic Reviews From PubMed<sup>a30</sup>

	Sensitivity, %	Specificity, %	PubMed Equivalent
Broad filter	99.9	52	search*[Title/ Abstract] OR meta analysis[Publication Type] OR meta analysis[Title/ Abstract] OR meta analysis[MeSH Terms] OR review[Publication Type] OR diagnosis[MeSH Subheading] OR associated[Title/Abstract]
Narrow filter	71	99	MEDLINE[Title/Abstract] OR (systematic[Title/ Abstract] AND review[Title/ Abstract]) OR meta analysis[Publication Type]

Abbreviation: MeSH, medical subject headings.

<sup>a</sup>These filters are not implemented in PubMed; the search strategy needs to be copied and pasted right after the search to optimally filter systematic reviews. Reproduced with permission from the *BMJ*.

their development and validations. For example, in addition to the ones we have already discussed, you will find filters for adverse events, economic evaluation, observational studies, and even qualitative studies (<https://sites.google.com/a/york.ac.uk/issg-search-filters-resource/home/search-filters-by-design>).

Another useful database for clinical practice is the Cochrane Controlled Trials Registry, the largest electronic compilation of controlled trials, built from MEDLINE, EMBASE, and other sources, including hand searches of most major health care journals. Because it includes only trials, this registry is the fastest, most reliable method of determining whether a controlled trial has been published on any topic. You can search the registry

in the Cochrane Library's advanced search function (<http://onlinelibrary.wiley.com/cochranelibrary/search>; select "Search Limits," then "Trials"). However, to access the full text of articles, you will need a subscription to the Cochrane Library or several Ovid Evidence-Based Medicine Review packages of databases (<http://www.ovid.com/site/catalog/DataBase/904.jsp>).

### Searching All Levels of the Pyramid at the Same Time

At this point, you may wonder if you can search across all levels of the pyramid of resources, instead of having sequential searches in different resources to get the current best evidence. Federated search engines do this easily. One of the most comprehensive and transparent federated resources is ACCESSSS (<http://plus.mcmaster.ca/accessss>). Typing a single question in ACCESSSS will run parallel searches in major resources from each level of the pyramid, from summaries to all types of preappraised research and all Clinical Queries filters in PubMed. [Table 4-6](#) presents the resources searched by ACCESSSS. Results are given in 1 page organized by level in the pyramid of EBM resources, with the most relevant and useful for clinical practice on the top (see [Figure 4-5](#)). Subscribing to ACCESSSS is free, although access to the full text of some resources will depend on institutional or personal subscriptions. To directly link your own subscriptions to all features of ACCESSSS, you can ask to add your institution to its list.

Other interesting and free federated searches that similarly search multiple resources at more or less each level of the pyramid are available. Instead of looking into summaries at the top, Trip (<http://www.tripdatabase.com>) has an algorithm to retrieve clinical practice guidelines, classified by country, along with many sources of synopses and other preappraised and nonpreappraised research. Its navigation is easy, and additional interesting features include the ability to structure your search with PICO (patient, intervention, comparator, outcome) and tailor your search to issues in developing countries. SumSearch (<http://sumsearch.org>) shares similar

TABLE 4-6

**Example of a Federated Search: EBM Resources Searched in Parallel in ACCESSSS<sup>a</sup>**

Summaries	DynaMed UpToDate Best Practice ACP PIER
Preappraised research	
Synopses of systematic reviews	<i>ACP Journal Club</i> DARE McMaster <i>PLUS</i> (including Cochrane reviews)
Systematic reviews	
Synopses of studies	McMaster <i>PLUS</i>
Nonpreappraised research	
Filtered studies	Clinical Queries in PubMed
Unfiltered studies	PubMed (MEDLINE)

Abbreviations: ACCESSSS, ACCess to Evidence-based Summaries, Synopses, Systematic Reviews and Studies; DARE, Database of Abstracts of Reviews of Effects; EBM, evidence-based medicine.

<sup>a</sup>Reproduced with permission of the Health Information Research Unit, McMaster University.

features, particularly for the retrieval of practice guidelines, but it organizes output according to level of processing (original studies, systematic reviews, and guidelines; Figure 4-1, middle box). SumSearch is linked to alerts from NEJM JournalWatch (<http://www.jwatch.org>). Finally, Epistemonikos (<http://www.epistemonikos.org>) is innovative both in simultaneously searching multiple resources and in indexing and interlinking relevant evidence. For example, Epistemonikos connects systematic reviews and their included studies and thus allows clustering of systematic reviews based on the primary studies they have in common. Epistemonikos is also unique in offering an appreciable multilingual user interface, multilingual search, and translation of abstracts in more than 9 languages.

FIGURE 4-5

## Output of a Federated Search in ACCESSSS

**ACCESSSS Federated Search**

History: dabigatran atrial fibrillation Search Advanced Options

Current PLUS Database: Physician

Resource Portal: McMaster University

Add your institution: Change

**Summaries \*\*\*\*\***

- UpToDate**
  - Antithrombotic therapy to prevent embolization in atrial fibrillation
  - Dabigatran: Drug Information
  - More Results...
- DynaMed**
  - Atrial fibrillation
  - Dabigatran
  - More Results...

**Synopses of Syntheses \*\*\*\*\***

- ACP Journal Club (selected via PLUS)**
  - Review: New oral anticoagulants reduced stroke and systemic embolism compared with warfarin in AF
  - Review: Dabigatran increases MI and reduces mortality compared with warfarin, enoxaparin, or placebo

**Syntheses \*\*\*\*\***

- PLUS Syntheses**
  - Stroke Prevention in Atrial Fibrillation (Systematic Review)
  - Combined anticoagulation and aspirin/platelet therapy for high-risk patients with atrial fibrillation: a systematic review (Systematic Review)

**Synopses of Studies \*\*\*\*\***

- ACP Journal Club (selected via PLUS)**
  - CHA<sub>2</sub>DS<sub>2</sub> score predicted bleeding and death in atrial fibrillation treated with anticoagulants
  - Dabigatran led to less major bleeding than warfarin in younger but not older patients with atrial fibrillation

**Studies (pre-appraised by these criteria) \*\*\*\*\***

- PLUS Studies**
  - Dabigatran versus warfarin in patients with mechanical heart valves (Original Study)
  - Intracranial hemorrhage in atrial fibrillation patients during anticoagulation with warfarin or dabigatran: the RE-LY trial (Original Study)
  - More Results...

*Below this bar you must do your own critical appraisal. (and can use these criteria if you wish)*

- PubMed Clinical Queries**

These results are yielded from your search term combined with Search Filters which are a modified version of our PubMed Clinical Queries.

  - Systematic Reviews**
    - Meta-Analysis of Randomized Controlled Trials on Risk of Myocardial Infarction from the Use of Oral Direct Thrombin Inhibitors.
    - Cost-effectiveness of pharmacogenetic guided warfarin therapy versus alternative anticoagulation in atrial fibrillation.
    - More Results...
  - Therapy**
    - The new oral anti-coagulants and the phase 3 clinical trials - a systematic review of the literature.
    - Safety and Efficacy of Dabigatran Compared With Warfarin for Patients Undergoing Radiofrequency Catheter Ablation of Atrial Fibrillation: A Meta-analysis.

**McMaster PLUS**

ES model explained  
Criteria for articles in PLUS

- Summaries \*\*\*\*\***
  - UpToDate
  - DynaMed
  - Best Practice
  - StatRef PER
- Synopses of Syntheses \*\*\*\*\***
  - ACP Journal Club (via PLUS)
  - DARE
- Syntheses \*\*\*\*\***
  - PLUS Syntheses
- Synopses of Studies \*\*\*\*\***
  - ACP Journal Club (via PLUS)
- Studies \*\*\*\*\***
  - PLUS Studies
- Non-Appraised \*\*\*\*\***
  - Pubmed CQ
  - Pubmed

Reproduced with permission of the Health Information Research Unit, McMaster University.

## When to Use Google

Google (<http://www.google.com>) has brought a revolution in the way we search the Internet. Its powerful algorithm retrieves answers to any type of question. Many factors seem to influence its output, including the relevance to your query but also

the number of times a specific website has been previously accessed or cited, the computer IP and server from which you conduct your search, your nationality, and possibly other financial and nonfinancial interests. Because of its lack of transparency, Google is not a reliable way to filter current best evidence from unsubstantiated or nonscientific supervised sources. When searching the Web, be aware that you are not searching defined databases but rather surfing the constantly shifting seas of electronic communications. The material you need that is supported by evidence may not float to the surface at any particular time.

On the other hand, “Googling” can be useful for defined purposes. It is often the fastest way to answer general background questions, often through multilingual resources such as Wikipedia (<http://www.wikipedia.org>), or research new topics, conditions, or treatments that have attracted media attention before being included in any EBM resources (eg, at the time of viral outbreaks around the globe, you may have wondered what Middle East respiratory syndrome coronavirus is). Google also can help you refine the wording of your search terms by rapidly finding 1 relevant citation. For example, you might want to learn whether incretins are associated with pancreatic cancer, but you are unclear about the different types of incretins. By searching Google and Wikipedia, you will rapidly remember how to spell (or copy and paste) dipeptidyl peptidase 4 inhibitor or glucagon-like peptide 1 analogs. Finally, Google can be a surprisingly powerful tool to search for uncommon patterns of symptoms and findings by simply typing them together as a query. These uncommon combinations would usually retrieve little or no information in most medical databases. Google can sometimes find the rare citation that would give you a clue about that syndrome.

A better alternative to Google for answering foreground questions is Google Scholar, which applies Google algorithms to scholarly literature (<http://www.google.com/scholar>). Although Google Scholar’s search algorithms are not transparent,



comparisons have found Google Scholar to be comparable to other databases,<sup>31</sup> and an analysis has found increasing evidence that Google Scholar retrieves twice as many relevant articles as PubMed, with almost 3 times greater access to free full-text articles,<sup>32</sup> as well as access to conference abstracts that might be useful for rare topics. Google Scholar has a complex searching system, and the help feature provides useful guidance in refining your searches (<http://scholar.google.com/intl/en/scholar/help.html>).

## TRANSLATING A QUESTION INTO SEARCH TERMS

---

### How to Choose and Combine Search Terms

Table 4-7 illustrates how you can break down a question into its PICO components and corresponding search terms. You next choose and combine search terms into a variety of search strategies, adapted to each resources. One advantage of searching the top EBM resources is that you can keep searches simple because the databases are highly selective and relatively small. One or 2 search terms for the population or problem and for your intervention or exposure will find most relevant topics. For example, if you are interested in the impact of mucolytics on patients with chronic obstructive pulmonary disease (COPD) who are stable, simply searching with the terms “COPD mucolytic” in summaries (eg, UpToDate) and preappraised research (eg, DARE) will usually suffice. Being too specific in your search can cause you to lose important information. In contrast, searching nonpreappraised research (eg, PubMed) usually requires more specific and structured searches.

To find the evidence you need in large databases, your search terms should closely relate to the components of your PICO question (see Chapter 3, What Is the Question?). For some components, the corresponding search terms are straightforward.

TABLE 4-7

## Combining Search Terms Into Different Search Strategies

PICO Components	Potential Search Terms
P Patients with stable chronic bronchitis	COPD OR (chronic bronchitis)
I Any mucolytic agent	Mucolytic
C Placebo (and current best care)	Placebo
O Number of exacerbation, mortality	Exacerbation OR mortality
Level of the Pyramid	Examples of Search Strategies <sup>a</sup>
Summaries and preappraised research	Chronic bronchitis mucolytic COPD mucolytic
Nonpreappraised research	COPD mucolytic exacerbation (COPD OR (chronic bronchitis)) AND mucolytic (COPD OR (chronic bronchitis)) AND mucolytic AND exacerbation (COPD OR (chronic bronchitis)) AND mucolytic AND (exacerbation OR mortality)

Abbreviation: COPD, chronic obstructive pulmonary disease; PICO, patient or population, intervention or exposure, comparator, and outcome.

<sup>a</sup>OR and AND are Boolean operators in these searches.

For example, if your population is patients with diabetes, you may simply use “diabetes” or “diabetic.” Other components of PICO may prove more challenging, such as “antithyroid drug therapy” as an intervention. Indeed, you might choose “anti-thyroid” as a single term or consider combining several drugs, such as “carbimazole OR propylthiouracil OR methimazole.” Notice that the latter example combines search terms with “OR” in capital letters to signify this is a *Boolean operator*: the search

will retrieve studies for either of these treatments. In contrast, adding no operator actually corresponds to linking search terms with “AND.” For example, typing “neuraminidase inhibitors” is equivalent to typing “neuraminidase AND inhibitors” and will retrieve only studies that include both terms, instead of all studies that include any type of inhibitor.

Efficient wording of search terms is based in part on your familiarity with the topic but is also based on trial and error. The *Medical Subject Headings* (MeSH) Thesaurus (<http://www.nlm.nih.gov/mesh/MBrowser.html>) can help you find words generally used by indexers for a given medical concept. A quick Google search often can give you a sense of appropriate wording in a faster way. If you are surprised that a search yields little relevant evidence, ask yourself if you misspelled a term or were too specific (eg, adding too many words that will automatically be linked with “AND”). Definitions also can differ. For example, in MeSH, “ventilation” refers to “supplying a building or house, their rooms and corridors, with fresh air.” “Pulmonary ventilation” is the preferred term for clinicians because it indicates “the total volume of gas inspired or expired per unit of time, usually measured in liters per minute.”

## Broad vs Narrow Searches

Table 4-8 indicates how to refine your search. If you initially found little evidence, you can broaden your search (eg, increase its sensitivity) by adding synonyms for each concept or using truncated terms (eg, *diabet\** will retrieve diabetes, diabetic, and many other similar terms with different endings). Conversely, if your initial search retrieved too many citations to be screened, you can narrow your search (eg, increase its specificity) by linking more PICO components with “AND” or by adding limits and methodologic filters (eg, narrow Clinical Queries; <http://www.ncbi.nlm.nih.gov/pubmed/clinical>). More sophisticated approaches include entering PICO components sequentially

TABLE 4-8

Refining the Search Strategy<sup>1,19,30,34</sup>

Ways to Increase Sensitivity	Ways to Increase Specificity
Many search terms for a similar PICO component, linked with “OR”	More PICO concepts linked with “AND”: (P) AND (I) AND (C) AND (O)
Truncated terms, wildcards (eg, diabet*, wom?n)	Use of NOT to exclude irrelevant terms
Synonyms (pressure sore, decubitus ulcer)	Use of NOT as Boolean operator
Variant spelling (tumour, tumor)	Limits (date, age group, etc)
Explosion of MeSH terms	Methodologic filters (Clinical Queries)
Use of PubMed “Related citations” or bibliography of relevant articles	Content filters (topic or disease specific)

Abbreviation: MeSH, medical subject headings; PICO, patient or population, intervention or exposure, comparator, and outcome.

according to their importance to obtain a manageable number of articles in large databases, such as PubMed.<sup>33</sup>

## Finding Related Articles

When your PubMed search seems laborious, a useful trick is to find at least 1 potentially relevant article to your question and use the “Related citations” feature, as highlighted in [Figure 4-6](#). It will automatically look for other articles that are similar in their titles, abstracts, and index terms. You then can screen the new output and select “Related citations” for each potentially relevant article you find. To keep track of potentially relevant citations, send them to the PubMed clipboard as you screen, and they will be labeled as items in the clipboard ([Figure 4-6](#)). This strategy may help you gather relevant articles rapidly in a snowball sampling.

FIGURE 4-6

## Features in PubMed: Related Citations and Clipboard

The screenshot shows the PubMed interface with search results for 'High-dose N-acetylcysteine in stable COPD'. A 'Send to:' dialog box is open, showing 'Clipboard' as the selected destination. Three red arrows labeled A, B, and C point to 'Related citations' links in the search results.

**Search Results:**

- 1. [High-dose N-acetylcysteine in stable COPD: the 1-year double-blind randomized placebo-controlled trial](#)  
Tse HN, Raiteri L, Wong KY, Yee KS, Ng LY, Wai KY, Loo CK, Chan MH.  
Chest. 2013 Jul;144(1):106-18. doi: 10.1378/chest.12-2357.  
PMID: 23348145 [PubMed - indexed for MEDLINE]  
[Related citations](#) ← **A**
- 2. [Effect of N-acetylcysteine on air trapping in COPD: a randomized placebo-controlled trial](#)  
Stav D, Raz M.  
Chest. 2009 Aug;136(2):381-6. doi: 10.1378/chest.09-0421. Epub 2009 May 15.  
PMID: 19447919 [PubMed - indexed for MEDLINE]  
[Related citations](#)
- 3. [Nebulized arformoterol in patients with COPD: a 12-week multicenter randomized double-blind double-dummy placebo- and active-controlled trial](#)  
Baumgartner RA, Hanania NA, Calhoun WJ, Sahn SA, Sciarappa K, Hanrahan JP.  
Cin Ther. 2007 Feb;29(2):261-78.  
PMID: 17472819 [PubMed - indexed for MEDLINE]  
[Related citations](#) [Item in clipboard](#) ← **C**

**Send to: Dialog Box:**

Choose Destination:

- File
- Collections
- Order
- Citation manager
- Clipboard
- E-mail
- My Bibliography

Add 1 items:

[Add to Clipboard](#) ← **B**

A, Link to “Related citations” from a relevant article. B, Dialogue box allowing user to send relevant articles to the clipboard. C, After having sent an article to the clipboard, it is labeled so in the output.

Reproduced with permission of the US National Library of Medicine and PubMed.

## Getting Help

Finally, because of the complexity and interconnections of medical databases, some searches simply require the help of information specialists. In anticipation of such cases in your clinical practice, befriend your medical librarians. They can be a great resource to help answer difficult questions or those that require elaborate search strategies.

### BOX 4-2

#### Tips to Help Improve Searching Skills

With the pyramid of EBM resources in mind, map the EBM resources that are accessible to you through your affiliations or personal subscriptions.

Choose which resources you would like to explore next, according to your information needs and the criteria described in this chapter.

Bookmark these resources in the browsers of all of your devices—your desktop computer, smartphone, or tablet. Find out if you can get remote access from your institution and implement it so that access is automatic.

Subscribe to an e-mail alerting system for newly published evidence that is transparent and trustworthy.

Train yourself on questions that are familiar to you and compare EBM resources.

Keep track of your questions. It can enhance your learning and help you reflect back on your evidence-based practice.

Finally, always keep the patient perspective. This will help you focus on the appropriate body of evidence that informs all patient important outcomes, instead of being driven by the evidence that is first presented to you.

## CONCLUSION: IMPROVING YOUR SEARCHING SKILLS IN DAILY PRACTICE

---

Box 4-2 presents a few practical tips to help you improve your searching skills in daily practice. Because of the continuous flow of new research findings of variable quality, finding current best evidence is challenging. However, this process has been greatly facilitated by the development of numerous EBM resources that can provide fast answers at the point of care. No resource is sufficient for all information needs, and you will need to use several in combination to find current best evidence. This chapter provides guidance on how to navigate across the pyramid of resources efficiently, ideally by using federated search engines.

### References

1. Straus SE. *Evidence-Based Medicine: How to Practice and Teach EBM*. 4th ed. New York, NY: Elsevier/Churchill Livingstone; 2011.
2. Agoritsas T, Merglen A, Courvoisier DS, et al. Sensitivity and predictive value of 15 PubMed search strategies to answer clinical questions rated against full systematic reviews. *J Med Internet Res*. 2012;14(3):e85.
3. Green ML, Ciampi MA, Ellis PJ. Residents' medical information needs in clinic: are they being met? *Am J Med*. 2000;109(3):218-223.
4. González-González AI, Dawes M, Sánchez-Mateos J, et al. Information needs and information-seeking behavior of primary care physicians. *Ann Fam Med*. 2007;5(4):345-352.
5. Graber MA, Randles BD, Ely JW, Monnahan J. Answering clinical questions in the ED. *Am J Emerg Med*. 2008;26(2):144-147.
6. Hoogendam A, Stalenhoef AF, Robbé PF, Overbeke AJ. Answers to questions posed during daily patient care are more likely to be answered by UpToDate than PubMed. *J Med Internet Res*. 2008;10(4):e29.
7. Hoogendam A, Stalenhoef AF, Robbé PF, Overbeke AJ. Analysis of queries sent to PubMed at the point of care: observation of search behaviour in a medical teaching hospital. *BMC Med Inform Decis Mak*. 2008;8:42.
8. Thiele RH, Piro NC, Scalzo DC, Nemergut EC. Speed, accuracy, and confidence in Google, Ovid, PubMed, and UpToDate: results of a randomised trial. *Postgrad Med J*. 2010;86(1018):459-465.

9. McKibbin KA, Fridsma DB. Effectiveness of clinician-selected electronic information resources for answering primary care physicians' information needs. *J Am Med Inform Assoc.* 2006;13(6):653-659.
10. Glasziou P, Burls A, Gilbert R. Evidence based medicine and the medical curriculum. *BMJ.* 2008;337:a1253.
11. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Med.* 2010;7(9):e1000326.
12. Guyatt GH, Akl EA, Crowther M, Schunemann HJ, Gutterman DD, Zelman Lewis S. Introduction to the ninth edition: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest.* 2012;141(2 suppl):48S-52S.
13. Haynes RB, Cotoi C, Holland J, et al; McMaster Premium Literature Service (PLUS) Project. Second-order peer review of the medical literature for clinical practitioners. *JAMA.* 2006;295(15):1801-1808.
14. Holland J, Haynes RB; McMaster PLUS Team Health Information Research Unit. McMaster Premium Literature Service (PLUS): an evidence-based medicine information service delivered on the Web. *AMIA Annu Symp Proc.* 2005;2005:340-344.
15. Garg AX, Adhikari NK, McDonald H, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *JAMA.* 2005;293(10):1223-1238.
16. Del Fiol G, Curtis C, Cimino JJ, et al. Disseminating context-specific access to online knowledge resources within electronic health record systems. *Stud Health Technol Inform.* 2013;192:672-676.
17. Roshanov PS, Fernandes N, Wilczynski JM, et al. Features of effective computerised clinical decision support systems: meta-regression of 162 randomised trials. *BMJ.* 2013;346:f657.
18. Jeffery R, Navarro T, Lokker C, Haynes RB, Wilczynski NL, Farjou G. How current are leading evidence-based medical textbooks? an analytic survey of four online textbooks. *J Med Internet Res.* 2012;14(6):e175.
19. Prorok JC, Iserman EC, Wilczynski NL, Haynes RB. The quality, breadth, and timeliness of content updating vary substantially for 10 online medical texts: an analytic survey. *J Clin Epidemiol.* 2012;65(12):1289-1295.
20. Vandvik PO, Lincoff AM, Gore JM, et al. Primary and secondary prevention of cardiovascular disease: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest.* 2012;141(2 suppl):e637S-e668S.
21. Wentz R. Visibility of research: FUTON bias. *Lancet.* 2002;360(9341):1256.
22. Kung J, Miller RR, Mackowiak PA. Failure of clinical practice guidelines to meet institute of medicine standards: two more decades of little, if any, progress. *Arch Intern Med.* 2012;172(21):1628-1633.
23. Vandvik PO, Brandt L, Alonso-Coello P, et al. Creating clinical practice guidelines we can trust, use, and share: a new era is imminent. *Chest.* 2013;144(2):381-389.



24. Banzi R, Cinquini M, Liberati A, et al. Speed of updating online evidence based point of care summaries: prospective cohort analysis. *BMJ*. 2011;343:d5856.
25. Martínez García L, Arévalo-Rodríguez I, Solà I, Haynes RB, Vandvik PO, Alonso-Coello P; Updating Guidelines Working Group. Strategies for monitoring and updating clinical practice guidelines: a systematic review. *Implement Sci*. 2012;7:109.
26. Ruschitzka F, Abraham WT, Singh JP, et al; EchoCRT Study Group. Cardiac-resynchronization therapy in heart failure with a narrow QRS complex. *N Engl J Med*. 2013;369(15):1395-1405.
27. Haynes RB, Holland J, Cotoi C, et al. McMaster PLUS: a cluster randomized clinical trial of an intervention to accelerate clinical use of evidence-based information from digital libraries. *J Am Med Inform Assoc*. 2006;13(6):593-600.
28. Lader E. Review: Eplerenone is not more effective for reducing mortality than other aldosterone antagonists. *Ann Intern Med*. 2012;157:JC6-10.
29. Haynes RB. ACP Journal Club: the best new evidence for patient care. *ACP J Club*. 2008;148(3):2.
30. Haynes RB, McKibbin KA, Wilczynski NL, Walter SD, Werre SR; Hedges Team. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. *BMJ*. 2005;330(7501):1179.
31. Montori VM, Wilczynski NL, Morgan D, Haynes RB; Hedges Team. Optimal search strategies for retrieving systematic reviews from Medline: analytical survey. *BMJ*. 2005;330(7482):68.
32. Kulkarni AV, Aziz B, Shams I, Busse JW. Comparisons of citations in Web of Science, Scopus, and Google Scholar for articles published in general medical journals. *JAMA*. 2009;302(10):1092-1096.
33. Shariff SZ, Bejaimal SA, Sontrop JM, et al. Retrieving clinical evidence: a comparison of PubMed and Google Scholar for quick clinical searches. *J Med Internet Res*. 2013;15(8):e164.
34. Dans AL, Dans LF, Silvestre MAA. Literature searches. In: Dans AL, Dans LF, Silvestre MAA, eds. *Painless Evidence-Based Medicine*. Chichester, England: John Wiley & Sons; 2008:115-136.
35. DiCenso A, Bayley L, Haynes RB. ACP Journal Club. Editorial: Accessing preappraised evidence: fine-tuning the 5S model into a 6S model. *Ann Intern Med*. 2009;151(6):JC3-JC2, JC3-JC3.

5

# Why Study Results Mislead: Bias and Random Error

Gordon Guyatt, Roman Jaeschke,  
and Maureen O. Meade

## IN THIS CHAPTER

---

Random Error

Bias

Strategies for Reducing the Risk of Bias

---

Our clinical questions have correct answers that correspond to an underlying reality or truth. For instance, there is a true underlying magnitude of the impact of  $\beta$ -blockers on mortality in patients with heart failure, the impact of inhaled corticosteroids on exacerbations in patients with asthma, the impact of reamed vs unreamed nailing of tibial fractures, the *prognosis* of patients with hip osteoarthritis, and the diagnostic properties of a pregnancy test. Research studies attempt to estimate that underlying truth. Unfortunately, however, we will never know the exact truth. Studies may be flawed in their design or conduct and introduce *systematic error* (or *bias*). Even if a study could be perfectly designed and executed, the estimated *treatment effect* may miss the mark because of *random error*. The next section explains why.

## RANDOM ERROR

---

Consider a perfectly balanced coin. Every time we flip the coin, the *probability* of it landing with its head up or tail up is equal—50%. Assume, however, that we as investigators do not know that the coin is perfectly balanced—in fact, we have no idea how well balanced it is, and we would like to find out. We can state our question formally: What is the true underlying probability of a resulting head or tail on any given coin flip? Our first experiment addressing this question is a series of 10 coin flips; the result: 8 heads and 2 tails. What are we to conclude? Taking our result at face value, we infer that the coin is very unbalanced (ie, biased in such a way that it yields heads more often than tails) and that the probability of heads on any given flip is 80%.

Few would be happy with this conclusion. The reason for our discomfort is that we know that the world is not constructed so that a perfectly balanced coin will always yield 5 heads and 5 tails in any given set of 10 coin flips. Rather, the result is subject to the play of chance, otherwise

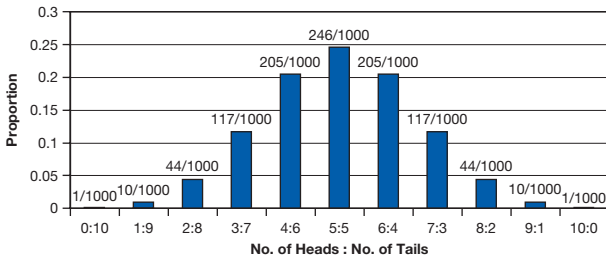
known as random error. Some of the time, 10 flips of a perfectly balanced coin will yield 8 heads. On occasion, 9 of 10 flips will turn up heads. On rare occasions, we will find heads on all 10 flips. Figure 5-1 shows the actual distribution of heads and tails in repeated series of coin flips.

What if the 10 coin flips yield 5 heads and 5 tails? Our awareness of the play of chance leaves us uncertain that the coin is balanced: a series of 10 coin flips of a very biased coin (a true probability of heads of 0.8, for instance) could, by chance, yield 5 heads and 5 tails.

Let us say that a funding agency, intrigued by the results of our first small experiment, provides us with resources to conduct a larger study. This time, we increase the sample size of our experiment markedly, conducting a series of 1000 coin flips. If we end up with 500 heads and 500 tails, are we ready to conclude that we are dealing with a true coin? We are much more confident but still not completely sure. The reason for our remaining uncertainty is that we know that, were the true underlying probability of heads 51%, we would sometimes see 1000 coin flips yield the result we have just observed.

**FIGURE 5-1**

**Theoretical Distribution of Results of an Infinite Number of Repetitions of 10 Flips of an Unbiased Coin**



We can apply the above logic to the results of studies that address questions of prognosis, diagnosis, and *harm*, and also to *randomized clinical trials* (RCTs) that address treatment issues. For instance, an RCT finds that 10 of 100 treated patients die during treatment, as do 20 of 100 control patients. Does treatment really reduce the death rate by 50%? Maybe, but awareness of chance will leave us with some degree of uncertainty about the magnitude of the *treatment effect*—and perhaps about whether treatment helps at all.

In a study of congestive heart failure, 228 of 1320 patients (17%) with moderate to severe heart failure allocated to receive *placebo* died, as did 156 of 1327 (12%) allocated to receive bisoprolol.<sup>1</sup> Although the true underlying reduction in the *relative risk* of dying is likely to be in the vicinity of the 32% suggested by the study, we must acknowledge that appreciable uncertainty remains about the true magnitude of the effect (see Chapter 9, Confidence Intervals: Was the Single Study or Meta-analysis Large Enough?).

We have now addressed the question with which we started: “Why is it that no matter how powerful and well designed a study, we will never be sure of the truth?” The answer is that chance is directionless, and it is equally likely, for instance, to overestimate or underestimate treatment effects.

## BIAS

---

Bias is the term we use for the other reason study results may be misleading. In contrast to random error, bias leads to systematic deviations (ie, the error has direction) from the underlying truth. In studies of prognosis, bias leads us to falsely optimistic or pessimistic conclusions about a patient’s fate. In studies of diagnosis, bias leads us to an overly optimistic (usually) or

pessimistic assessment of a test's value in differentiating between those with and without a target condition. In treatment or harm studies, bias leads to either an underestimate or an overestimate of the underlying benefit or harm (Box 5-1).

Bias may intrude as a result of differences, other than the *experimental intervention*, between patients in treatment and *control groups* at the time they enter a study. At the start of a study, each patient, if left untreated, is destined to do well—or poorly. To do poorly means to have an adverse event (eg, a stroke) during the study. We often refer to the adverse event that is the focus of a study as the *target outcome* or *target event*. Bias will result if treated and control patients differ in their prognosis (ie, their likelihood of experiencing the target outcome) at the start of the study. For instance, if patients in the control group have more severe atherosclerosis or are older than their counterparts, their destiny will be to have a greater proportion of adverse events than those in the intervention or treatment group, and the results of the study will be biased in favor of the treatment group; that is, the study will yield a larger treatment

## BOX 5-1

### How Can a Study of an Intervention (Treatment) Be Biased?

Intervention and control groups may be different at the start

Example: patients in control group are sicker or older

Intervention and control groups may, independent of the experimental treatment, become different as the study proceeds

Example: patients in the intervention group receive effective additional medication

Intervention and control groups may differ, independent of treatment, at the end

Example: more sick patients lost to follow-up in the intervention group

effect than would be obtained were the study groups prognostically similar at baseline.

Even if patients in the intervention and control groups begin the study with the same prognosis, the result may still be biased. This will occur if, for instance, effective interventions are differentially administered to treatment and control groups. For instance, in a study of a novel agent for the *prevention* of complications of atherosclerosis, the intervention group might receive more intensive statin therapy than the control group.

Finally, patients may begin prognostically similar, and stay prognostically similar, but the study may end with a biased result. This could occur if, for example, the study loses patients to *follow-up* (see Chapter 6, Therapy [Randomized Trials]) or because a study is *stopped early* because of an apparent large treatment effect.

## STRATEGIES FOR REDUCING THE RISK OF BIAS

---

This book teaches you how to recognize *risk of bias* not only in studies that address issues of therapy and harm but also in studies of prognosis and diagnosis. In studies of prognosis, investigators can reduce bias by enrolling a representative sample and ensuring they are completely followed up. In studies of diagnosis, investigators can ensure that they have chosen an appropriate *criterion* or *gold standard* for diagnosis and that those interpreting test results are unaware of the gold standard findings. In the remainder of this chapter, however, we focus on issues of therapy and harm.

We have noted that bias arises from differences in *prognostic factors* in treatment and control groups at the start of a study or from differences in prognosis that arise as a study proceeds. What can investigators do to reduce these biases? [Table 5-1](#) summarizes the available strategies to reduce biases in RCTs and *observational studies*.



TABLE 5-1

**Ways of Reducing Bias in Studies of Therapy and Harm**

Source of Bias	Therapy: Strategy for Reducing Bias	Harm: Strategy for Reducing Bias
<b>Differences Observed at the Start of the Study</b>		
Treatment and control patients differ in prognosis	Randomization	Statistical adjustment for prognostic factors in the analysis of data
	Randomization with stratification	Matching
<b>Differences That Arise as the Study Proceeds</b>		
Placebo effects	Blinding of patients	Choice of outcomes (such as mortality) less subject to placebo effects
Cointervention	Blinding of caregivers	Documentation of treatment differences and statistical adjustment
Bias in assessment of outcome	Blinding of assessors of outcome	Choice of outcomes (such as mortality) less subject to observer bias

*(Continued)*



**TABLE 5-1****Ways of Reducing Bias in Studies of Therapy and Harm (Continued)**

<b>Source of Bias</b>	<b>Therapy: Strategy for Reducing Bias</b>	<b>Harm: Strategy for Reducing Bias</b>
<b>Differences at the Completion of the Study</b>		
Loss to follow-up	Ensuring complete follow-up	Ensuring complete follow-up
Stopping study early because of large effect	Completing study as initially planned by sample size calculation	Not applicable
Omitting patients who did not receive assigned treatment	Including all patients for whom data are available in the arm to which they were randomized	Not applicable

When studying new treatments, investigators can implement a large number of strategies to limit the risk of bias. They can reduce the likelihood of differences in the prognostic features in treated and untreated patients at baseline by *randomly allocating* patients to the 2 groups. They can balance placebo effects by administering identical-appearing but biologically inert treatments—placebos—to patients in the control group. *Blinding* clinicians to whether patients are receiving active or placebo therapy can eliminate the risk of important *cointerventions*, and blinding outcome assessors minimizes bias in the assessment of event rates.

Investigators studying either treatment effects or harm using observational study designs have far less control over the risk of bias. They must be content to compare patients whose *exposure* is determined by their choice or circumstances, and they can address potential differences in patients' fate only by statistical adjustment for known prognostic factors. Blinding is impossible, so their best defense against placebo effects and bias in outcome assessment is to choose *end points*, such as death, that are less subject to these biases. Investigators who address both sets of questions can reduce bias by minimizing loss to follow-up (Table 5-1).

Note that when investigators choose observational study designs to study treatment issues, clinicians must apply the risk of bias criteria developed primarily for questions of harm. Similarly, if the potentially harmful exposure is a drug with beneficial effects, investigators may be able to randomize patients to intervention and control groups. In this case, clinicians can apply the risk of bias criteria designed primarily for therapy questions. Whether for issues of therapy or harm, the strength of inference from RCTs will almost invariably be greater than the strength of inference from observational studies.

## Reference

1. CIBIS-II Investigators and Committees. The Cardiac Insufficiency Bisoprolol Study II (CIBIS-II): a randomised trial. *Lancet*. 1999;353(9146):9-13.

*This page intentionally left blank*

6

# Therapy (Randomized Trials)

Michael Walsh, Vlado Perkovic, Braden Manns, Sadeesh Srinathan, Maureen O. Meade, PJ Devereaux, and Gordon Guyatt

## IN THIS CHAPTER

---

### Clinical Scenario

A Patient With Peripheral Artery Disease: How Can I Improve Physical Function and Walking?

### Finding the Evidence

The Users' Guides

### How Serious Is the Risk of Bias?

Did Intervention and Control Groups Start With the Same Prognosis?

Was Prognostic Balance Maintained as the Study Progressed?

Were the Groups Prognostically Balanced at the Study's Completion?

### What Are the Results?

How Large Was the Treatment Effect?

How Precise Was the Estimate of the Treatment Effect?

(continued on following page)

### **How Can I Apply the Results to Patient Care?**

Were the Study Patients Similar to the Patient in My Practice?

Were All Patient-Important Outcomes Considered?

Are the Likely Treatment Benefits Worth the Potential Harm and Costs?

### **Clinical Scenario Resolution**

---

## CLINICAL SCENARIO

### A Patient With Peripheral Artery Disease: How Can I Improve Physical Function and Walking?

**Y**ou are a general internist following up a 62-year-old man with a history of type 2 diabetes mellitus, hypertension, and hyperlipidemia who is taking oral hypoglycemics, a statin, and a thiazide-like diuretic. A vascular surgeon recently evaluated the patient for intermittent claudication and made a diagnosis of peripheral artery disease. The surgeon prescribed low-dose aspirin and pentoxifylline to reduce the patient's risk of vascular events and improve his ability to walk, citing 2 *systematic reviews*: a review of antiplatelet agents in peripheral artery disease that found a decrease in the odds of vascular events (*odds ratio* [OR], 0.78; 95% *confidence interval* [CI], 0.63-0.96) and an increase in walking distance of 59 m (95% CI, 37-81 m) and another review of pentoxifylline in peripheral artery disease that increased maximum walking distance by 59 m (95% CI, 37-81 m).<sup>1,2</sup> Despite the new treatments, the patient is unable to walk more than 2 minutes without pain and finds his quality of life substantially impaired.

Listening to the patient's story of poor response to treatment and ongoing symptoms, you recall seeing an article that may be relevant. You ask him to return in a week for further review of his medications.

## FINDING THE EVIDENCE

You formulate the relevant question for this individual: in a patient with debilitating peripheral vascular disease treated with antiplatelet therapy and not a candidate for surgery,

how can we improve symptom-free walking? To conduct a rapid search focused on the most recent preappraised research (see Chapter 4, Finding Current Best Evidence), you opt for *ACP Journal Club*, directly accessible through your institution (<http://acpjc.acponline.org>). Typing the terms “peripheral vascular disease” and “intermittent claudication” identifies 7 preappraised editorial summaries of studies, one of which turns out to be your target: Ramipril Improved Walking Times and QOL in Peripheral Artery Disease and Intermittent Claudication.<sup>3</sup> You print a copy of the summary<sup>3</sup> and the original full-text article that reports the results of the trial, Effect of Ramipril on Walking Times and Quality of Life Among Patients With Peripheral Artery Disease and Intermittent Claudication.<sup>4</sup>

This article describes a trial that includes 212 patients with peripheral artery disease and a history of stable intermittent claudication. Participants were *randomly allocated* to ramipril, 10 mg daily, or *placebo* for 24 weeks. The primary outcomes were pain-free walking time and maximum walking time.

## THE USERS' GUIDES

---

Box 6-1 presents our usual 3-step approach to using an article from the medical literature to guide your practice. You will find these criteria useful for a variety of therapy-related questions, including treating symptomatic illnesses (eg, asthma or arthritis), *prevention* of distant complications of illness (eg, cardiovascular death after myocardial infarction), *screening* for silent but treatable disease (eg, colon cancer screening), and choosing the optimal diagnostic approach (as in randomized trials of alternative diagnostic strategies that address *patient-important outcomes*).

## BOX 6-1

**Users' Guides for an Article About Therapy****How serious was the risk of bias?**

Did intervention and control groups start with the same prognosis?

Were patients randomized?

Was randomization concealed?

Were patients in the study groups similar with respect to known prognostic factors?

Was prognostic balance maintained as the study progressed?

To what extent was the study blinded?

Were the groups prognostically balanced at the study's completion?

Was follow-up complete?

Were patients analyzed in the groups to which they were randomized?

Was the trial stopped early?

**What are the results?**

How large was the treatment effect?

How precise was the estimate of the treatment effect?

**How can I apply the results to patient care?**

Were the study patients similar to my patient?

Were all patient-important outcomes considered?

Are the likely treatment benefits worth the potential harm and costs?

If the answer to one key question (“Were patients randomized?”) is no, some of the other questions (“Was randomization *concealed*?” “Were patients analyzed in the groups to which they were randomized?”) become irrelevant. Nonrandomized *observational studies* typically yield far weaker inferences than



*randomized clinical trials* (RCTs). Nevertheless, clinicians must use the best *evidence* available in managing their patients, even if the quality of that evidence is limited (see Chapter 2, What Is Evidence-Based Medicine?). The criteria in Chapter 10 (Harm [Observational Studies]) will help you assess an observational study that addresses a potential treatment that has not yet been evaluated in an RCT.

## HOW SERIOUS IS THE RISK OF BIAS?

---

### Did Intervention and Control Groups Start With the Same Prognosis?

#### Were Patients Randomized?

Consider the question of whether hospital care prolongs life. A study finds that more sick people die in the hospital than in the community. We would easily reject the naive conclusion that hospital care kills people because we recognize that hospitalized patients are sicker than patients in the community.

Although the logic of prognostic balance is vividly clear in comparing hospitalized patients with those in the community, it may be less obvious in other contexts. Many people believe that a diet rich in  $\omega$ 3 fatty acids will decrease their risk of a cardiovascular event. This belief arose from many observational studies in which people who ingested larger quantities of  $\omega$ 3 fatty acids had fewer cardiovascular events than those that who ate lesser quantities.<sup>5</sup> However, large randomized trials did not find any benefits with  $\omega$ 3 fatty acid supplementation.<sup>6,7</sup>

Other surprises generated by randomized trials include the demonstration that antioxidant vitamins fail to reduce gastrointestinal cancer<sup>8</sup>—and one such agent, vitamin E, may actually increase all-cause mortality<sup>9</sup>—and that a variety of initially promising drugs increase mortality in patients with heart failure.<sup>10-12</sup> Such surprises occur periodically when investigators conduct randomized trials to test the observations from studies

in which patients and physicians determine which treatment a patient receives.

The reason that studies in which patient or physician preference determines whether a patient receives treatment or control (observational studies) often yield misleading results is that morbidity and mortality result from many causes. Treatment studies attempt to determine the impact of an intervention on events such as stroke, myocardial infarction, and death—occurrences that we call the trial's *target outcomes*. A patient's age, the underlying severity of illness, the presence of *comorbidity*, and a host of other factors typically determine the frequency with which a trial's target outcome occurs (*prognostic factors* or *determinants of outcome*). If prognostic factors—either those we know about or those we do not know about—prove unbalanced between a trial's treatment and *control groups*, the study's outcome will be biased, either underestimating or overestimating the treatment's effect. Because known prognostic factors often influence clinicians' recommendations and patients' decisions about taking treatment, observational studies often yield biased results that may get the magnitude or even the direction of the effect wrong.

Observational studies can theoretically match patients, either in the selection of patients for study or in the subsequent statistical analysis, for known *prognostic factors* (see Chapter 10, Harm [Observational Studies]). However, not all prognostic factors are easily measured or characterized, and in many diseases only a few are known. Therefore, even the most careful patient selection and statistical methods are unable to completely address the *bias* in the estimated *treatment effect*. The power of randomization is that treatment and control groups are more likely to have a balanced distribution of known and unknown prognostic factors.

Consider again our example of the  $\omega$ 3 fatty acid studies. What was the cause of bias in the  $\omega$ 3 fatty acids observational studies? People who eat larger amounts of  $\omega$ 3 fatty acids may typically have a higher socioeconomic status than those who eat smaller amounts. In addition, patients who eat larger amounts

of  $\omega 3$  fatty acids may eat fewer unhealthy foods and may be more careful with other important risk factors (eg, smoking and exercise). Their apparent benefit from  $\omega 3$  fatty acids may reflect their healthier lifestyle. Whatever the explanation, we are now confident that it was their previous *prognosis*, rather than the  $\omega 3$  fatty acids, that led to lower rates of cardiovascular events.

Although randomization is a powerful technique, it does not always succeed in creating groups with similar prognosis. Investigators may make mistakes that compromise randomization, or randomization may fail because of chance—unlikely events sometimes happen. The next 2 sections address these issues.

When those enrolling patients are unaware and cannot control the arm to which the patient is allocated, we refer to randomization as concealed. In unconcealed trials, those responsible for recruitment may systematically enroll sicker—or less sick—patients to either a treatment or control group. This behavior will compromise the purpose of randomization, and the study will yield a biased result.<sup>13-15</sup> Careful investigators will ensure that randomization is concealed through strategies such as remote randomization, in which the individual recruiting the patient makes a call to a methods center to discover the arm of the study to which the patient is assigned.

Consider, for instance, a trial of  $\beta$ -blockers vs angiotensin-converting enzyme (ACE) inhibitors for hypertension treatment that used opaque numbered envelopes to conceal randomization.<sup>16</sup> At the time the study was conducted, evidence suggested that  $\beta$ -blockers were better for patients with heart disease. Significantly more patients with heart disease were assigned to receive  $\beta$ -blockers ( $P = .037$ ). In addition, evidence suggested that ACE inhibitors were better for patients with diabetes mellitus. Significantly more patients with diabetes were assigned to receive ACE

inhibitors ( $P = .048$ ). It is possible that clinicians were opening envelopes and violating the randomization to ensure patients received what the clinicians believed was the best treatment. Thus, the prognostic balance that randomization could have achieved was prevented.

### **Were Patients in the Treatment and Control Groups Similar With Respect to Known Prognostic Factors?**

The purpose of randomization is to create groups whose prognosis, with respect to the target outcomes, is similar. Sometimes, through bad luck, randomization will fail to achieve this goal. The smaller the sample size, the more likely the trial will have prognostic imbalance.

Picture a trial testing a new treatment for heart failure that is enrolling patients classified as having New York Heart Association functional class III and class IV heart failure. Patients with class IV heart failure have a much worse prognosis than those with class III heart failure. The trial is small, with only 8 patients. One would not be surprised if all 4 patients with class III heart failure were allocated to the treatment group and all 4 patients with class IV heart failure were allocated to the control group. Such a result of the allocation process would seriously bias the study in favor of the treatment. Were the trial to enroll 800 patients, one would be startled if randomization placed all 400 patients with class III heart failure in the treatment arm. The larger the sample size, the more likely randomization will achieve its goal of prognostic balance.

You can check how effectively randomization has balanced known prognostic factors by looking for a display of patient characteristics of the treatment and control groups at the study's commencement—the baseline or entry prognostic features. Although we will never know whether similarity exists for the unknown prognostic factors, we are reassured when the known prognostic factors are well balanced.

All is not lost if the treatment groups are not similar at baseline. Statistical techniques permit adjustment of the study result for baseline differences. When both *adjusted analyses* and unadjusted analyses generate the same conclusion, clinicians gain confidence that the *risk of bias* is not excessive.

## Was Prognostic Balance Maintained as the Study Progressed?

### To What Extent Was the Study Blinded?

If randomization succeeds, treatment and control groups begin with a similar prognosis. Randomization, however, provides no guarantees that the 2 groups will remain prognostically balanced. *Blinding* is the optimal strategy for maintaining prognostic balance.

Box 6-2 describes 5 groups involved in clinical trials that, ideally, will remain unaware of whether patients are receiving the *experimental therapy* or control therapy. Patients who take a treatment that they believe is effective may feel and perform better than those who do not, even if the treatment has no biologic activity. Although the magnitude and consistency of this *placebo effect* remain uncertain,<sup>17-20</sup> investigators interested in determining the biologic impact of a treatment will ensure patients are blind to treatment allocation. Similarly, rigorous research designs will ensure blinding of those caring for participants, as well as those collecting, evaluating, and analyzing data (Box 6-2). Demonstrations of bias introduced by unblinding, such as the results of a trial in multiple sclerosis in which a treatment benefit judged by unblinded outcome assessors

**BOX 6-2****Five Groups That Should, if Possible, Be Blind to Treatment Assignment**

Patients	To avoid placebo effects
Clinicians	To prevent differential administration of therapies that affect the outcome of interest (cointervention)
Data collectors	To prevent bias in data collection
Adjudicators of outcome	To prevent bias in decisions about whether or not a patient has had an outcome of interest
Data analysts	To avoid bias in decisions regarding data analysis

disappeared when adjudicators of outcome were blinded,<sup>21</sup> highlight the importance of blinding. The more subjectivity involved in judging whether a patient has had a target outcome, the more important blinding becomes. For example, blinding of an outcome assessor is unnecessary when the outcome is all-cause mortality.

Finally, differences in patient care other than the intervention under study—*cointerventions*—can, if they affect study outcomes, bias the results. Effective blinding eliminates the possibility of either conscious or unconscious differential administration of effective interventions to treatment and control groups. When effective blinding is not possible, documentation of potential cointerventions becomes important.

**Were the Groups Prognostically Balanced at the Study's Completion?**

It is possible for investigators to effectively conceal and blind treatment assignment and still fail to achieve an unbiased result.

### Was Follow-up Complete?

Ideally, at the conclusion of a trial, investigators will know the status of each patient with respect to the target outcome. The greater the number of patients whose outcome is unknown—patients *lost to follow-up*—the more a study is potentially compromised. The reason is that patients who are lost to follow-up often have different prognoses from those who are retained—they may disappear because they have adverse outcomes or because they are doing well and so did not return for assessment.<sup>22</sup> The magnitude of the bias may be substantial. A systematic review suggested that up to a third of positive trials reported in high-impact journals may lose significance given plausible assumptions regarding differential loss to follow-up in treatment and control groups.<sup>23</sup>

When does loss to follow-up pose a serious risk of bias? Although you may run across thresholds such as 20% for a serious risk of bias, such rules of thumb are misleading. Consider 2 hypothetical randomized trials, each of which enters 1000 patients into both the treatment and control groups, of whom 30 (3%) are lost to follow-up (Table 6-1). In trial A, treated patients die at half the rate of the control group (200 vs 400), a *relative risk* (RR) of 50%. To what extent does the loss to follow-up threaten our inference that treatment reduces the death rate by half? If we assume the worst (ie, that all treated patients lost to follow-up died), the number of deaths in the *experimental group* would be 230 (23%). If there were no deaths among the control patients who were lost to follow-up, our best estimate of the effect of treatment in reducing the RR of death decreases from 200/400, or 50%, to 230/400, or 58%. Thus, even assuming the worst makes little difference to the best estimate of the magnitude of the treatment effect. Our inference is therefore secure.

Contrast this with trial B. Here, the RR of death is also 50%. In this case, however, the total number of deaths is much lower; of the treated patients, 30 die, and the number of deaths in control patients is 60. In trial B, if we make the same worst-case assumption about the fate of the patients lost to follow-up, the results would change markedly. If we assume that all patients initially allocated to treatment—but subsequently lost to follow-up—die, the number of deaths among treated patients increases from 30 to 60, which is equal to the number of control group deaths. If this assumption is accurate, we would have 60 deaths in both the treatment and control groups, and the effect of treatment would decrease to 0. Because of this marked change in the treatment effect (50% RR if we ignore those lost to follow-up; 100% RR if we assume all patients in the treatment group who were lost to follow-up died), the 3% loss to follow-up in trial B threatens our inference about the magnitude of the RR.

Of course, this worst-case scenario is unlikely. When a worst-case scenario, were it true, substantially alters the results, you must judge the plausibility of a markedly different outcome *event rate* in the treatment and control group patients lost to follow-up. Ideally, investigators would conduct *sensitivity analyses* to deal with this issue. Because they seldom do, guidelines are available to help you should you choose to make your own judgment of the trial's vulnerability to loss to follow-up.<sup>23</sup>

Thus, loss to follow-up may substantially increase the risk of bias. If assuming a worst-case scenario does not change the inferences arising from study results, then loss to follow-up is unlikely a problem. If such an assumption would significantly alter the results, the extent to which bias is introduced depends



TABLE 6-1

## When Does Loss to Follow-up Seriously Increase Risk of Bias?

	Trial A		Trial B	
	Treatment	Control	Treatment	Control
No. of patients randomized	1000	1000	1000	1000
No. (%) lost to follow-up	30 (3)	30 (3)	30 (3)	30 (3)
No. (%) of deaths	200 (20)	400 (40)	30 (3)	60 (6)
RR not counting patients lost to follow-up	0.2/0.4 = 0.50		0.03/0.06 = 0.50	
RR for worst-case scenario <sup>a</sup>	0.23/0.4 = 0.58		0.06/0.06 = 1	

Abbreviation: RR, relative risk.

<sup>a</sup>The worst-case scenario assumes that all patients allocated to the treatment group and lost to follow-up died and all patients allocated to the control group and lost to follow-up survived.

on how likely it is that treatment patients lost to follow-up fared badly, whereas control patients lost to follow-up fared well. That decision is a matter of judgment.

### Was the Trial Stopped Too Early?

Stopping trials early (ie, before enrolling the planned sample size) when one sees an apparent large benefit is risky and may compromise randomization. These *stopped early trials* run the risk of greatly overestimating the treatment effect.<sup>24</sup>

A trial designed with too short a follow-up also may compromise crucial information that adequate length of follow-up would reveal. For example, consider a trial that randomly assigned patients with an abdominal aortic aneurysm to either an open surgical repair or a less invasive, endovascular repair technique.<sup>25</sup> At the end of the 30-day follow-up, mortality was

significantly lower in the endovascular technique group (*relative risk reduction* [RRR], 0.61; 95% CI, 0.13-0.82). The investigators followed up participants for an additional 2 years and found that there was no difference in mortality between groups after the first year. Had the trial ended earlier, the endovascular technique may have been considered substantially better than the open surgical technique.

### Were Patients Analyzed in the Groups to Which They Were Randomized?

Investigators will undermine the benefits of randomization if they omit from the analysis patients who do not receive their assigned treatment or, worse yet, count events that occur in *non-adherent* patients who were assigned to treatment against the control group. Such analyses will bias the results if the reasons for nonadherence are related to prognosis. In a number of randomized trials, patients who did not adhere to their assigned drug regimens fared worse than those who took their medication as instructed, even after taking into account all known prognostic factors.<sup>26-31</sup> When adherent patients are destined to have a better outcome, omitting those who do not receive assigned treatment undermines the unbiased comparison provided by randomization. Investigators prevent this bias when they follow the *intention-to-treat* principle and analyze all patients in the group to which they were randomized irrespective of what treatment they actually received.<sup>32</sup> Following the intention-to-treat principle does not, however, reduce bias associated with loss to follow-up.<sup>33</sup>

## USING THE GUIDE

**R**eturning to our opening clinical scenario, did the experimental and control groups begin the study with a similar prognosis? The study was randomized and allocation was concealed; 212 patients participated and 95% were followed up.<sup>4</sup> The investigators followed the

intention-to-treat principle, including all patients they had followed up in the arm to which they were randomized, and stopped when they reached the planned sample size. There were more patients who had occlusive arterial disease (39.6% vs 22.7%) in the ramipril group. This finding could bias the results in favor of the placebo group, and the investigators do not provide an adjusted analysis for the baseline differences. Clinicians, patients, data collectors, outcomes assessors, and data analysts were all blind to allocation.

The final risk of bias assessment represents a continuum from studies that are at very low risk of bias to others that are at very high risk of yielding a biased estimate of effect. Inevitably, where a study lies in this continuum involves some judgment. In this case, despite uncertainty about baseline differences between the groups, we conclude that the risk of bias is low.

## WHAT ARE THE RESULTS?

---

### How Large Was the Treatment Effect?

Most frequently, RCTs monitor *dichotomous* outcomes (eg, “yes” or “no” classifications for cancer recurrence, myocardial infarction, or death). Patients either have such an event or they do not, and the article reports the proportion of patients who develop such events. Consider, for example, a study in which 20% of a control group died but only 15% of those receiving a new treatment died (Table 6-2). How might one express these results?

One possibility is the absolute difference (known as the *absolute risk reduction* [ARR] or *risk difference*) between the proportion who died in the control group (*control group risk* [CGR]) and the proportion who died in the experimental group (*experimental group risk* [EGR]), or  $CGR - EGR = 0.20 - 0.15 = 0.05$ . Another way to express the impact of treatment is as

TABLE 6-2

## Results From a Hypothetical Randomized Trial

Exposure	Outcome, No. of Patients		
	Death	Survival	Total
Treatment (experimental)	15	85	100
Control	20	80	100

Control group risk (CGR):  $20/100 = 20\%$   
 Experimental group risk (EGR):  $15/100 = 15\%$   
 Absolute risk reduction or risk difference:  $CGR - EGR, 20\% - 15\% = 5\%$   
 Relative risk:  $EGR/CGR = (15/100)/(20/100) \times 100\% = 75\%$   
 Relative risk reduction:  $[1 - (EGR/CGR)] \times 100\% = 1 - 75\% = 25\%$

Abbreviations: CGR, control group risk; EGR, experimental group risk.

the RR: the risk of events among patients receiving the new treatment relative to that risk among patients in the control group, or  $EGR/CGR = 0.15/0.20 = 0.75$ .

The most commonly reported measure of dichotomous treatment effects is the complement of the RR, the RRR. It is expressed as a percentage:  $1 - (EGR/CGR) \times 100\% = (1 - 0.75) \times 100\% = 25\%$ . An RRR of 25% means that of those who would have died had they been in the control group, 25% will not die if they receive treatment; the greater the RRR, the more effective the therapy. Investigators may compute the RR during a specified period, as in a *survival analysis*; the relative measure of effect in such a time-to-event analysis is called the *hazard ratio* (see Chapter 8, Does Treatment Lower Risk? Understanding the Results). When people do not specify whether they are talking about RRR or ARR—for instance, “Drug X was 30% effective in reducing the risk of death” or “The efficacy of the vaccine was 92%”—they are almost invariably talking about RRR (see Chapter 8, Does Treatment Lower Risk? Understanding the Results).

## How Precise Was the Estimate of the Treatment Effect?

We can never be sure of the true risk reduction; the best estimate of the true treatment effect is what we observe in a well-designed randomized trial. This estimate is called a *point estimate* to remind us that, although the true value lies somewhere in its neighborhood, it is unlikely to be precisely correct. Investigators often tell us the neighborhood within which the true effect likely lies by calculating CIs, a range of values within which one can be confident the true effect lies.<sup>34</sup>

We usually use the 95% CI (see Chapter 9, Confidence Intervals: Was the Single Study or Meta-analysis Large Enough?). You can consider the 95% CI as defining the range that—assuming the study has low risk of bias—includes the true RRR 95% of the time. The true RRR will generally lie beyond these extremes only 5% of the time, a property of the CI that relates closely to the conventional level of *statistical significance* of  $P < .05$ . We illustrate the use of CIs in the following examples.

### Example 1

If a trial randomized 100 patients each to experimental and control groups, and there were 20 deaths in the control group and 15 deaths in the experimental group, the authors would calculate a point estimate for the RRR of 25% [ $\text{CGR} = 20/100$  or 0.20,  $\text{EGR} = 15/100$  or 0.15, and  $1 - \text{EGR}/\text{CGR} = (1 - 0.75) \times 100 = 25\%$ ]. You might guess, however, that the true RRR might be much smaller or much greater than 25%, based on a difference of only 5 deaths. In fact, you might surmise that the treatment might provide no benefit (an RRR of 0%) or might even do harm (a negative RRR). And you would be right; in fact, these results are consistent with both an RRR of  $-38\%$  (that is, patients given the new treatment might be 38% more likely to die than control patients) and an RRR of nearly

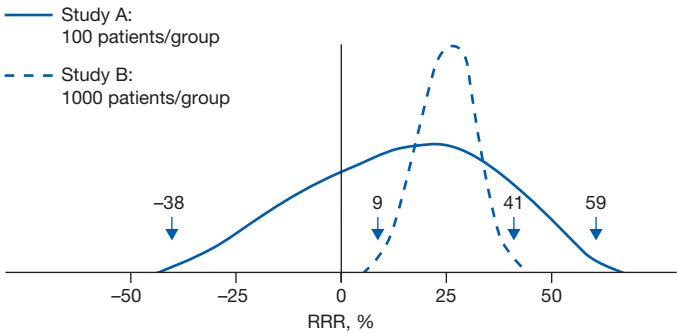
59% (ie, patients subsequently receiving the new treatment might have a risk of dying almost 60% less than those who are not treated). In other words, the 95% CI on this RRR is -38% to 59%, and the trial really has not helped us decide whether or not to offer the new treatment (Figure 6-1).

## Example 2

What if the trial enrolled 1000 patients per group rather than 100 patients per group, and the same event rates were observed as before, so that there were 200 deaths in the

**FIGURE 6-1**

### Confidence Intervals in Trials of Various Sample Size



Abbreviation: RRR, relative risk reduction.

Two studies with the same point estimate, a 25% RRR, but different sample sizes and correspondingly different CIs. The x-axis represents the different possible RRR, and the y-axis represents the likelihood of the true RRR having that particular value. The solid line represents the CI around the first example, in which there were 100 patients per group, and the number of events in the active and control groups were 15 and 20, respectively. The dashed line represents the CI around the second example, in which there were 1000 patients per group, and the number of events in the active and control groups were 150 and 200, respectively.

control group ( $CGR = 200/1000 = 0.20$ ) and 150 deaths in the experimental group ( $EGR = 150/1000 = 0.15$ )? Again, the point estimate of the RRR is 25% ( $1 - EGR/CGR = 1 - [0.15/0.20] \times 100 = 25\%$ ).

In this larger trial, you might think that our confidence that the true reduction in risk is close to 25% is much greater; again, you would be right. The 95% CI on the RRR for this set of results is all on the positive side of 0 and runs from 9% to 41% (Figure 6-1).

These examples show that the larger the sample size and higher the number of outcome events in a trial, the greater our confidence that the true RRR (or any other measure of effect) is close to what we observed. The point estimate—in this case, 25%—is the one value most likely to represent the true RRR. As one considers values farther and farther from the point estimate, they become less and less likely to represent the truth. By the time one crosses the upper or lower boundaries of the 95% CI, the values are unlikely to represent the true RRR. All of this assumes the study is at low risk of bias.

Not all randomized trials have dichotomous outcomes, nor should they. In a study of respiratory muscle training for patients with chronic airflow limitation, one primary outcome measured how far patients could walk in 6 minutes in an enclosed corridor.<sup>35</sup> This 6-minute walk improved from a mean of 406 to 416 m (up 10 m) in the experimental group receiving respiratory muscle training, and 409 to 429 m (up 20 m) in the control group. The point estimate for improvement in the 6-minute walk due to respiratory muscle training therefore was negative, at -10 m (or a 10-m difference in favor of the control group).

Here, too, you should look for the 95% CIs around this difference in changes in exercise capacity and consider their implications. The investigators tell us that the lower boundary of the 95% CI was -26 (ie, the results are consistent

with a difference of 26 m in favor of the control treatment) and the upper boundary was 5 m. Even in the best of circumstances, patients are unlikely to perceive adding 5 m to the 400 recorded at the start of the trial as important, and this result effectively excludes an important benefit of respiratory muscle training as applied in this study.

Having determined the magnitude and precision of the treatment effect, clinicians can turn to the final question of how to apply the article's results to their patients.

## USING THE GUIDE

**U**sing the numbers provided in the article,<sup>4</sup> patients in the ramipril group walked 75 seconds (95% CI, 60-89 seconds) longer without pain than the placebo group and 255 seconds (95% CI, 215-295 seconds) longer overall. The effect of ramipril is convincing given that the 95% CIs are narrow and the lower boundaries are far from showing no effect (ie, 0 seconds). The clinical importance of walking 75 seconds without pain is likely noticeable given that they could walk a mean of 140 seconds without pain at baseline. This finding is consistent with a substantial improvement in a secondary outcome, a measure of *health-related quality of life*, for patients in the ramipril group.

## HOW CAN I APPLY THE RESULTS TO PATIENT CARE?

### Were the Study Patients Similar to the Patient in My Practice?

If the patient before you would have qualified for enrollment in the study, you can apply the results with considerable confidence



or consider the results *generalizable*. Often, your patient has different attributes or characteristics from those enrolled in the trial and would not have met a study's eligibility criteria. Patients may be older or younger, may be sicker or less sick, or may have comorbid disease that would have excluded them from participation in the study.

A study result probably applies even if, for example, adult patients are 2 years too old for enrollment in the study, had more severe disease, had previously been treated with a competing therapy, or had a comorbid condition. A better approach than rigidly applying the study *inclusion* and *exclusion criteria* is to ask whether there is some compelling reason why the results do not apply to the patient. You usually will not find a compelling reason, in which case you can generalize the results to your patient with confidence.

A related issue has to do with the extent to which we can generalize findings from a study using a particular drug to another closely (or not so closely) related agent. The issue of drug *class effects* and how conservative one should be in assuming class effects remains controversial. Generalizing findings of surgical treatment may be even riskier. Randomized trials of carotid endarterectomy, for instance, demonstrate much lower perioperative rates of stroke and death than one might expect in one's own community, which may reflect on either the patients or surgeons (and their relative expertise) selected to participate in randomized trials.<sup>36</sup> An example of how expertise might be considered is provided below.

### Expertise in Procedural Interventions

Unlike pharmacologic interventions in which we expect the intervention to vary minimally between patients, procedural interventions may differ substantially based on the expertise of the physician and the technology available to deliver the intervention.

For example, it is suggested that “off-pump” coronary artery bypass surgery reduces the risk of postoperative complications compared with the traditional “on-pump” technique. When the 2 techniques are compared in a randomized trial, one must be careful interpreting the results because of potential differences in expertise. For example, if surgeons participating in the trial are, on average, less skilled with the off-pump technique, the outcomes of patients in the off-pump group may reflect surgeon inexperience more than the true risks and merits of the technique. Further, surgeons may choose to switch from off-pump to on-pump technique more frequently than they would switch from on-pump to off-pump. This will bias the result toward demonstrating no difference between the techniques. One way of preventing these misleading results is by ensuring only surgeons with sufficient expertise in both on-pump and off-pump techniques are allowed to participate in the trial, as was done in the CABG Off or On Pump Revascularization Study (CORONARY) trial.<sup>37</sup> Another method of preventing this differential expertise bias is to randomize patients to a surgeon with expertise in one technique or to a surgeon with expertise in the alternate technique rather than randomize the patient to a surgeon who will perform either procedure to which the patient is randomized.<sup>38</sup>

A final issue arises when a patient fits the features of a subgroup of patients in the trial report. We encourage you to be skeptical of *subgroup analyses*.<sup>39</sup> The treatment is likely to benefit the subgroup more or less than the other patients only if the difference in the effects of treatment in the subgroups is large and unlikely to occur by chance. Even when these conditions

apply, the results may be misleading, particularly when investigators did not specify their hypotheses before the study began, if they had a large number of hypotheses, or if other studies fail to replicate the finding.<sup>40</sup>

### Were All Patient-Important Outcomes Considered?

Treatments are indicated when they provide important benefits. Demonstrating that a bronchodilator produces small increments in forced expiratory volume in patients with chronic airflow limitation, that a vasodilator improves cardiac output in heart failure patients, or that a lipid-lowering agent improves lipid profiles does not provide sufficient justification for administering these drugs. In these instances, investigators have chosen *substitute outcomes* or *surrogate outcomes* rather than those that patients would consider important. What clinicians and patients require is evidence that the treatments improve outcomes that are important to patients, such as reducing shortness of breath during the activities required for daily living, avoiding hospitalization for heart failure, or decreasing the risk of a major stroke.<sup>41</sup>

Trials of the impact of antiarrhythmic drugs after myocardial infarction illustrate the danger of using substitute outcomes or end points. Because abnormal ventricular depolarizations were associated with a high risk of death and antiarrhythmic drugs demonstrated a reduction in abnormal ventricular depolarizations (the substitute end point), it made sense that they should reduce death. A group of investigators performed randomized trials on 3 agents (encainide, flecainide, and moricizine) that were previously found to be effective in suppressing the substitute end point of abnormal ventricular depolarizations. The investigators had to stop the trials when they discovered that mortality was substantially higher in patients receiving antiarrhythmic treatment than in those receiving placebo.<sup>42,43</sup> Clinicians relying on the substitute end point of arrhythmia suppression would have continued to administer the 3 drugs, to the considerable detriment of their patients.

Even when investigators report favorable effects of treatment on a patient-important outcome, you must consider whether there may be deleterious effects on other outcomes. For instance, cancer chemotherapy may lengthen life but decrease its quality. Randomized trials often fail to adequately document the toxicity or adverse effects of the experimental intervention.<sup>44</sup>

*Composite end points* represent a final dangerous trend in presenting outcomes. Like surrogate outcomes, composite end points are attractive for reducing sample size and decreasing length of follow-up. Unfortunately, they can mislead. For example, a trial that reduced a composite outcome of death, nonfatal myocardial infarction, and admission for an acute coronary syndrome actually demonstrated a trend toward increased mortality with the experimental therapy and convincing effects only on admission for an acute coronary syndrome.<sup>45</sup> The composite outcome would most strongly reflect the treatment effect of the most common of the components, admission for an acute coronary syndrome, even though there is no convincing evidence the treatment reduces the risk of death or myocardial infarction.

Another long-neglected outcome is the resource implications of alternative management strategies. Health care systems face increasing resource constraints that mandate careful attention to *economic analysis*.

## Are the Likely Treatment Benefits Worth the Potential Harm and Costs?

If the results of a study apply to your patient and the outcomes are important to your patient, the next question concerns whether the probable treatment benefits are worth the associated *risks, burden, and resource requirements*. A 25% reduction in the RR of death may sound impressive, but its impact on your patient may nevertheless be minimal. This notion is illustrated by using a concept called *number needed to treat* (NNT), the number of patients who must receive an intervention of therapy during a specific period to prevent 1 adverse outcome or produce 1 positive outcome.<sup>46</sup>

**TABLE 6-3**

**Considerations in the Decision to Treat 2 Patients With Myocardial Infarction With Clopidogrel and Aspirin or Aspirin Alone**

Grade	Risk of Death or MI 1 Year After MI With Aspirin Alone (CER)	Risk With Clopidogrel Plus Aspirin (EGR) (ARR = CGR – EGR)	NNT (100/ARR When ARR Is Expressed as a Percentage)
40-year-old man with small MI	5.3%	4.2% (1.1% or 0.011)	91
70-year-old man with large MI and heart failure	36%	28.8% (7.2% or 0.072)	14

Abbreviations: ARR, absolute risk reduction; CER, control event rate; CGR, control group risk; EGR, experimental group risk; MI, myocardial infarction; NNT, number needed to treat.

The impact of a treatment is related not only to its RRR but also to the risk of the adverse outcome it is designed to prevent. One large trial in myocardial infarction suggests that clopidogrel in addition to aspirin reduces the RR of death from a cardiovascular cause, nonfatal myocardial infarction, or stroke by approximately 20% in comparison to aspirin alone.<sup>47</sup> Table 6-3 considers 2 patients presenting with acute myocardial infarction without elevation of ST segments on their electrocardiograms.

In the first case, a 40-year-old man presents with electrocardiographic findings that suggest an inferior myocardial infarction without ST-segment elevation. You find no signs of heart failure; the patient is in normal sinus rhythm, with a rate of 80/min; and he does not have elevated troponin. This individual's risk of death or recurrent

myocardial infarction in the next year is estimated to be 5.3%. Compared with aspirin alone, clopidogrel in addition to aspirin would reduce this risk by 20% to 4.2%, an ARR of 1.1% (0.011). The inverse of this ARR (ie, 100 divided by the ARR expressed as a percentage) is equal to the number of such patients we would have to treat to prevent 1 event (ie, 1 death, or recurrent myocardial infarction after a mild myocardial infarction in a low-risk patient), the NNT. In this case, we would have to treat approximately 91 such patients to prevent 1 recurrent myocardial infarction or save 1 life ( $100/1.1 = 91$ ). Given the small decrease in the outcome of death, recurrent myocardial infarction, or stroke (most noticeably recurrent myocardial infarction) with clopidogrel, the small increased risk of major bleeding associated with clopidogrel, and its additional cost, many clinicians might prefer aspirin alone in this patient.

In the second case, a 70-year-old man presents with electrocardiographic signs of anterior myocardial infarction with pulmonary edema and cardiogenic shock. His risk of dying or having a recurrent myocardial infarction in the subsequent year is approximately 36%. A 20% RRR of death in such a high-risk patient generates an ARR of 7.2% (0.072), and we would have to treat only 14 such individuals to avert a recurrent myocardial infarction or death ( $100/7.2 = 13.8$ ). Many clinicians would consider clopidogrel in addition to aspirin.

A key element of the decision to start therapy, therefore, is to consider the patient's risk of the event if left untreated. For any given RRR, the higher the probability that a patient will experience an adverse outcome if we do not treat, the more likely the patient will benefit from treatment and the fewer such patients we need to treat to prevent 1 adverse outcome

(see Chapter 8, Does Treatment Lower Risk? Understanding the Results). Knowing the NNT assists clinicians in helping patients weigh the benefits and downsides associated with their management options.

Trading off benefits and risks also requires an accurate assessment of the adverse effects of treatment. Randomized trials with relatively small sample sizes are unsuitable for detecting rare but catastrophic adverse effects of therapy. Clinicians often must look to other sources of information—often characterized by higher risk of bias—to obtain an estimate of the adverse effects of therapy (see Chapter 10, Harm [Observational Studies]).

When determining the optimal treatment choice based on the relative benefits and harms of a therapy, the *values and preferences* of each individual patient must be considered. How best to communicate information to patients and how to incorporate their values into clinical decision making remain areas of active investigation in *evidence-based medicine* (see Chapter 18, Decision Making and the Patient).

## CLINICAL SCENARIO RESOLUTION

**T**he study that we identified found an increase in pain-free and total walking time of patients with peripheral arterial disease treated with ramipril compared with placebo.<sup>4</sup> The authors did not describe any harmful effects of ramipril other than more withdrawals due to cough than placebo-treated patients. This finding may leave some uncertainty as to the net benefits to patients. In particular, there is no mention of kidney failure or hyperkalemia-induced cardiac arrest, the most serious adverse effects associated with ramipril. However, there is a large body of literature on patients with other types of vascular disease that suggests that ramipril, at the dose used in this study, is well tolerated and safe, particularly if clinicians monitor patients periodically for

the precursors to these adverse effects (ie, changes in kidney function or serum potassium).

Your patient is significantly limited by his intermittent claudication. He is similar to patients included in this study. Given the treatment effect on walking time and the observed effect on health-related quality of life, as well as an apparently minimal side effect profile, the study suggests patient-important benefits to taking ramipril.

The patient finds his limited walking ability and the pain he experiences debilitating. He believes that being able to walk for 1 additional minute would be worthwhile. He is, however, under financial stress and is concerned that ramipril costs \$1.20 per pill, or approximately \$450 in the next year. You explain that the investigators' choice of medication leaves some doubt about the best drug to use. The investigators could have chosen lisinopril, an ACE inhibitor with marginal differences from ramipril, which the patient can purchase for approximately one-third the price. Ultimately, implicitly accepting a class effect, the patient chooses the lisinopril.

## References

1. Momsen AH, Jensen MB, Norager CB, Madsen MR, Vestersgaard-Andersen T, Lindholt JS. Drug therapy for improving walking distance in intermittent claudication: a systematic review and meta-analysis of robust randomised controlled studies. *Eur J Vasc Endovasc Surg*. 2009;38(4):463-474.
2. Wong PF, Chong L-Y, Stansby G. Antiplatelet therapy to prevent cardiovascular events and mortality in patients with intermittent claudication. *JAMA*. 2013;309(9):926-927.
3. Jaar BG. ACP Journal Club. Ramipril improved walking times and QOL in peripheral artery disease and intermittent claudication. *Ann Intern Med*. 2013;158(12):JC7.
4. Ahimastos AA, Walker PJ, Askew C, et al. Effect of ramipril on walking times and quality of life among patients with peripheral artery disease and intermittent claudication: a randomized controlled trial. *JAMA*. 2013;309(5):453-460.
5. Hu FB, Bronner L, Willett WC, et al. Fish and omega-3 fatty acid intake and risk of coronary heart disease in women. *JAMA*. 2002;287(14):1815-1821.



6. Kotwal S, Jun M, Sullivan D, Perkovic V, Neal B. Omega 3 fatty acids and cardiovascular outcomes: systematic review and meta-analysis. *Circ Cardiovasc Qual Outcomes*. 2012;5(6):808-818.
7. Bosch J, Gerstein HC, Dagenais GR, et al; ORIGIN Trial Investigators. n-3 fatty acids and cardiovascular outcomes in patients with dysglycemia. *N Engl J Med*. 2012;367(4):309-318.
8. Bjelakovic G, Nikolova D, Simonetti RG, Gluud C. Antioxidant supplements for prevention of gastrointestinal cancers: a systematic review and meta-analysis. *Lancet*. 2004;364(9441):1219-1228.
9. Miller ER III, Pastor-Barriuso R, Dalal D, Riemersma RA, Appel LJ, Guallar E. Meta-analysis: high-dosage vitamin E supplementation may increase all-cause mortality. *Ann Intern Med*. 2005;142(1):37-46.
10. Ikram H, Crozier IG. Xamoterol in severe heart failure. *Lancet*. 1990;336(8713):517-518.
11. Califf RM, Adams KF, McKenna WJ, et al. A randomized controlled trial of eprostenol therapy for severe congestive heart failure: The Flolan International Randomized Survival Trial (FIRST). *Am Heart J*. 1997;134(1):44-54.
12. Hampton JR, van Veldhuisen DJ, Kleber FX, et al; Second Prospective Randomised Study of Ibopamine on Mortality and Efficacy (PRIME II) Investigators. Randomised study of effect of ibopamine on survival in patients with advanced severe heart failure. *Lancet*. 1997;349(9057):971-977.
13. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA*. 1995;273(5):408-412.
14. Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet*. 1998;352(9128):609-613.
15. Balk EM, Bonis PA, Moskowitz H, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA*. 2002;287(22):2973-2982.
16. Hansson L, Lindholm LH, Niskanen L, et al. Effect of angiotensin-converting-enzyme inhibition compared with conventional therapy on cardiovascular morbidity and mortality in hypertension: the Captopril Prevention Project (CAPPP) randomised trial. *Lancet*. 1999;353(9153):611-616.
17. Kaptchuk TJ. Powerful placebo: the dark side of the randomised controlled trial. *Lancet*. 1998;351(9117):1722-1725.
18. Hróbjartsson A, Gøtzsche PC. Is the placebo powerless? an analysis of clinical trials comparing placebo with no treatment. *N Engl J Med*. 2001;344(21):1594-1602.
19. McRae C, Cherin E, Yamazaki TG, et al. Effects of perceived treatment on quality of life and medical outcomes in a double-blind placebo surgery trial. *Arch Gen Psychiatry*. 2004;61(4):412-420.
20. Rana JS, Mannam A, Donnell-Fink L, Gervino EV, Sellke FW, Laham RJ. Longevity of the placebo effect in the therapeutic angiogenesis and laser

- myocardial revascularization trials in patients with coronary heart disease. *Am J Cardiol.* 2005;95(12):1456-1459.
21. Noseworthy JH, Ebers GC, Vandervoort MK, Farquhar RE, Yetisir E, Roberts R. The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. *Neurology.* 1994;44(1):16-20.
  22. Ioannidis JP, Bassett R, Hughes MD, Volberding PA, Sacks HS, Lau J. Predictors and impact of patients lost to follow-up in a long-term randomized trial of immediate versus deferred antiretroviral treatment. *J Acquir Immune Defic Syndr Hum Retrovirol.* 1997;16(1):22-30.
  23. Akl EA, Briel M, You JJ, et al. Potential impact on estimated treatment effects of information lost to follow-up in randomised controlled trials (LOST-IT): systematic review. *BMJ.* 2012;344:e2809.
  24. Montori VM, Devereaux PJ, Adhikari NK, et al. Randomized trials stopped early for benefit: a systematic review. *JAMA.* 2005;294(17):2203-2209.
  25. Greenhalgh RM, Brown LC, Powell JT, Thompson SG, Epstein D, Sculpher MJ; United Kingdom EVAR Trial Investigators. Endovascular versus open repair of abdominal aortic aneurysm. *N Engl J Med.* 2010;362(20):1863-1871.
  26. The Coronary Drug Project Research Group. Influence of adherence to treatment and response of cholesterol on mortality in the coronary drug project. *N Engl J Med.* 1980;303(18):1038-1041.
  27. Asher WL, Harper HW. Effect of human chorionic gonadotrophin on weight loss, hunger, and feeling of well-being. *Am J Clin Nutr.* 1973;26(2):211-218.
  28. Hogarty GE, Goldberg SC. Drug and psychotherapy in the aftercare of schizophrenic patients: one-year relapse rates. *Arch Gen Psychiatry.* 1973;28(1):54-64.
  29. Fuller R, Roth H, Long S. Compliance with disulfiram treatment of alcoholism. *J Chronic Dis.* 1983;36(2):161-170.
  30. Pizzo PA, Robichaud KJ, Edwards BK, Schumaker C, Kramer BS, Johnson A. Oral antibiotic prophylaxis in patients with cancer: a double-blind randomized placebo-controlled trial. *J Pediatr.* 1983;102(1):125-133.
  31. Horwitz RI, Viscoli CM, Berkman L, et al. Treatment adherence and risk of death after a myocardial infarction. *Lancet.* 1990;336(8714):542-545.
  32. Montori VM, Guyatt GH. Intention-to-treat principle. *CMAJ.* 2001;165(10):1339-1341.
  33. Alshurafa M, Briel M, Akl EA, et al. Inconsistent definitions for intention-to-treat in relation to missing outcome data: systematic review of the methods literature. *PLoS One.* 2012;7(11):e49163.
  34. Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. *Br Med J (Clin Res Ed).* 1983;286(6376):1489-1493.
  35. Guyatt G, Keller J, Singer J, Halcrow S, Newhouse M. Controlled trial of respiratory muscle training in chronic airflow limitation. *Thorax.* 1992;47(8):598-602.
  36. Walker MD, Marler JR, Goldstein M, et al; Executive Committee for the Asymptomatic Carotid Atherosclerosis Study. Endarterectomy for asymptomatic carotid artery stenosis. *JAMA.* 1995;273(18):1421-1428.

37. Lamy A, Devereaux PJ, Prabhakaran D, et al; CORONARY Investigators. Off-pump or on-pump coronary-artery bypass grafting at 30 days. *N Engl J Med*. 2012;366(16):1489-1497.
38. Devereaux PJ, Bhandari M, Clarke M, et al. Need for expertise based randomised controlled trials. *BMJ*. 2005;330(7482):88.
39. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med*. 1992;116(1):78-84.
40. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? updating criteria to evaluate the credibility of subgroup analyses. *BMJ*. 2010;340:c117.
41. Guyatt G, Montori V, Devereaux PJ, Schünemann H, Bhandari M. Patients at the center: in our practice, and in our use of language. *ACP J Club*. 2004;140(1):A11-A12.
42. Echt DS, Liebson PR, Mitchell LB, et al. Mortality and morbidity in patients receiving encainide, flecainide, or placebo: The Cardiac Arrhythmia Suppression Trial. *N Engl J Med*. 1991;324(12):781-788.
43. Rogers W, Epstein A, Arciniegas J, et al; The Cardiac Arrhythmia Suppression Trial II Investigators. Effect of the antiarrhythmic agent moricizine on survival after myocardial infarction. *N Engl J Med*. 1992;327(4):227-233.
44. Ioannidis JP, Lau J. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. *JAMA*. 2001;285(4):437-443.
45. Pfisterer M, Buser P, Osswald S, et al; Trial of Invasive versus Medical therapy in Elderly patients (TIME) Investigators. Outcome of elderly patients with chronic symptomatic coronary artery disease with an invasive vs optimized medical treatment strategy: one-year results of the randomized TIME trial. *JAMA*. 2003;289(9):1117-1123.
46. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med*. 1988;318(26):1728-1733.
47. Yusuf S, Zhao F, Mehta SR, Chrolavicius S, Tognoni G, Fox KK; Clopidogrel in Unstable Angina to Prevent Recurrent Events Trial Investigators. Effects of clopidogrel in addition to aspirin in patients with acute coronary syndromes without ST-segment elevation. *N Engl J Med*. 2001;345(7):494-502.

7

# How to Use a Noninferiority Trial

Sohail M. Mulla, Ian A. Scott, Cynthia A. Jackevicius, John J. You, and Gordon Guyatt

## IN THIS CHAPTER

---

### Clinical Scenario

#### Introduction

#### Are the Results Valid?

Did the Investigators Guard Against an Unwarranted Conclusion of Noninferiority?

Did the Investigators Analyze Patients According to the Treatment They Received and to the Groups to Which They Were Assigned?

#### What Are the Results?

#### How Can I Apply the Results to Patient Care?

Are the Likely Advantages of the Experimental Treatment Worth the Potential Harm and Costs?

#### Clinical Scenario Resolution

#### Conclusions

---

## CLINICAL SCENARIO

**Y**ou are an internist seeing a 51-year-old woman with severe osteoarthritis and limited mobility who presents with progressive dyspnea for a 3-day period. She is subjectively in distress, with a pulse of 105/min, a respiratory rate of 28/min, and an arterial oxygen saturation of 85% while breathing room air. Aside from her arthritis, the physical examination findings are unremarkable, and a lower-extremity examination reveals no sign of deep venous thrombosis. A computed tomographic (CT) pulmonary angiogram reveals unequivocal clot in 2 lobar arteries.

Recently, you have been treating patients with deep venous thrombosis without hospital admission using low-molecular-weight heparin (LMWH) administration in the outpatient setting. You are less comfortable not admitting a patient in the more dangerous setting of pulmonary embolism. You recall receiving, from the updating service to which you subscribe (see Chapter 4, Finding Current Best Evidence), a recent *randomized trial* that addressed this issue. Before discussing the issue of inpatient vs outpatient treatment with your patient, you quickly review the article.<sup>1</sup> In doing so, you find that the trial tested for *noninferiority* and you wonder, as you begin to read the methods and results, if there are special issues you should consider when using this article to guide your clinical care.

## INTRODUCTION

---

Traditionally, *randomized clinical trials* (RCTs) have sought to ascertain whether an experimental treatment is superior to standard treatment or *placebo* in improving quality of life or preventing morbid or mortal events—what we will refer to as effectiveness outcomes. In these *superiority trials*, the primary objective is to determine the magnitude of increased

benefit of the experimental intervention over standard therapy on effectiveness outcomes.

Recently, another paradigm has emerged that offers novel experimental treatments not on the basis of superiority in effectiveness outcomes but instead because they reduce harms or other treatment burdens relative to standard treatment. In modern medicine, clinicians are fortunate to have many effective treatments; unfortunately, these treatments are often associated with *harms*, inconvenience, or excessive cost. For these interventions, reducing treatment *burden*, including limitations and inconvenience, becomes a legitimate goal of innovative therapy.

In such instances, a question arises: can clinicians be confident that the experimental treatment's impact on effectiveness outcomes—the prime reason for wanting to prescribe it—is sufficiently close to that of standard treatment that they are comfortable substituting it for the existing standard? In technical terms, is the novel treatment noninferior to the standard treatment?

Noninferiority trials provide an alternative to *equivalence trials*, which endeavor to establish that an experimental treatment is neither better nor worse (beyond a specified margin) than the standard. In contrast, the noninferiority trialist is unconcerned if the experimental treatment is better as long as it is “not much worse.” Perhaps illustrating the limitations of the term, a noninferior treatment may thus be inferior, just not so inferior that it would cause concern. How much worse (ie, how much less effective) clinicians should consider acceptable will depend on the importance of the effectiveness outcome and the magnitude of the reduction in harms or burden achieved by the new treatment.

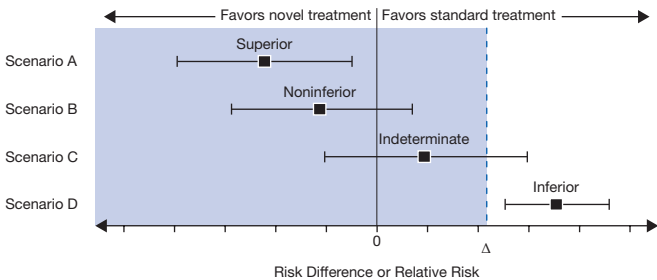
Consider how the concept of “not much worse” plays out for the patient in the previously presented scenario. She may dislike spending time in the hospital and may strongly prefer treatment at home, but there may be risks

she incurs in choosing home management. Perhaps the care she would receive in the hospital would result in a lower risk of recurrent *venous thromboembolism* (VTE) and a lower risk of serious bleeding, which can complicate antithrombotic therapy. Would our patient be willing to incur the additional risk of VTE or serious bleeding that may be associated with home management? If so, what level of increased risk would she be willing to tolerate?

The example illustrates the following point: given that patients will choose the new experimental treatment only if the risks are not much worse than the standard treatment, the critical issue in interpreting noninferiority trials is the choice of an acceptable threshold of “not much worse.” This noninferiority threshold (the dashed line labeled  $\Delta$  in Figure 7-1) is the maximum allowable excess of outcome events that arises

**FIGURE 7-1**

**Possible Outcome Scenarios in Noninferiority Trials**



The dashed line labeled  $\Delta$  represents the noninferiority threshold or the maximum allowable excess of outcome events arising from the experimental treatment compared with the standard treatment. The tinted area represents the noninferiority zone.

from the experimental treatment compared with the standard treatment.

When designing noninferiority trials, investigators set their own thresholds, typically using statistically based criteria. However, there is no universally accepted method for defining an appropriate threshold. It depends on the eye of the beholder. Experts have recommended using sound statistical reasoning and clinical judgment in determining noninferiority thresholds.<sup>2,3</sup> What is sound reasoning for one observer, however, may strike another as misguided.

The US Food and Drug Administration (FDA) has produced draft guidance regarding noninferiority thresholds that has proved highly influential.<sup>4</sup> The logic of the FDA's approach begins by considering the smallest plausible benefit achieved by the existing standard treatment with which the experimental—and it is hoped noninferior—treatment is compared. One establishes the smallest plausible benefit of the existing standard treatment by examining the results of a trial of that treatment against the previous best care or placebo. To establish the smallest plausible benefit, one focuses on the *confidence interval* (CI) around the observed estimate of effect (in technical terms, the CI around the *point estimate*) and, in particular, the boundary of the CI nearest to no effect.

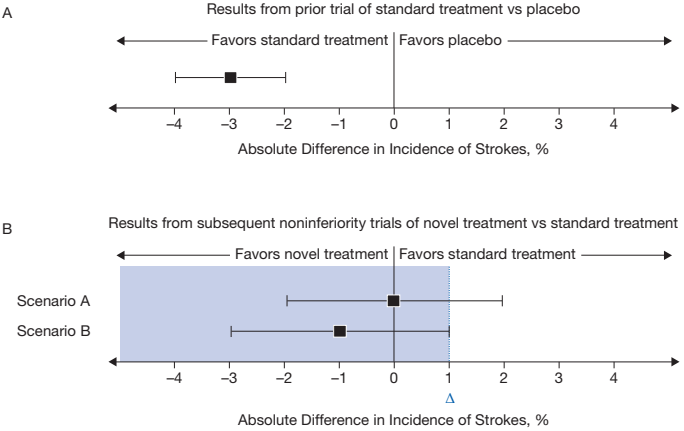
For instance, the point estimate may suggest that the existing standard treatment decreases the absolute incidence of stroke, relative to placebo, by an absolute difference of 3%, with a 95% CI of 2% to 4% (Figure 7-2, top graph). The smallest plausible benefit of the standard drug is then 2%, or 2 fewer strokes for every 100 patients treated.

If the 95% CIs around the difference in strokes in a subsequent trial testing an experimental drug for noninferiority include an increase in strokes of 2% (for instance, a point estimate of no difference, with a CI of a 2% decrease to a 2% increase), the results are consistent with the new drug being no better than placebo (Figure 7-2, scenario A). This is because the absolute benefit of the existing standard may be a reduction



FIGURE 7-2

Setting an Acceptable Noninferiority Threshold



A, Standard treatment decreases the absolute incidence of stroke, relative to placebo, by 3%, with a 95% CI of 2% to 4%. B, The blue dashed line represents the noninferiority zone. In scenario A, the 95% CIs around the difference in strokes between the experimental treatment and the standard treatment include an increase in incidence of strokes by as much as 2% with the experimental treatment, thereby failing to retain 50% of the minimal treatment effect of the standard treatment. In scenario B, the same 95% CIs suggest that the experimental treatment increases the incidence of strokes by no more than 1%, thus successfully preserving at least 50% of the 2% absolute reduction in stroke with the standard treatment.

in strokes of as little as 2%, and those receiving the experimental treatment may have a stroke rate of 2% more than the standard treatment—exactly equivalent to the rate on placebo.

The logic then goes that we should insist on some preservation of the treatment effect. Commonly, drug regulatory authorities stipulate that at least 50% of that minimal treatment effect be preserved. The threshold would, in this example, be 1%; if the experimental treatment increases strokes by no

more than 1% relative to the existing standard, at least 50% of the 2% absolute reduction in stroke has been preserved<sup>5,6</sup> (Figure 7-2, scenario B). Depending on the seriousness of the outcome, some may argue for retaining a greater proportion of benefit, resulting in a more challenging noninferiority margin. We have focused herein on expression of the noninferiority margin in absolute terms; sometimes, the choice of threshold is based on a relative rather than an absolute effect.

## USING THE GUIDE

**S**ometimes, the standard approaches to setting noninferiority thresholds are not applicable, as was the case in the trial of inpatient vs outpatient treatment for pulmonary embolism. Because there are no randomized trials that compare anticoagulation to no anticoagulation in pulmonary embolism, the authors could not use the procedure for setting the noninferiority margin described in the previously presented scenario. As an alternative, they first considered the likelihood of recurrent VTE at 90 days in low-risk inpatients with pulmonary embolism, which they estimated at 0.9%. They then specified a noninferiority margin of 4% (implying that patients would find it acceptable if the rate of recurrent VTE for outpatients would be <4.9%). They justified their choice by saying that it was similar to the noninferiority margins—3% to 5%—set in other trials of different anticoagulant regimens in acute VTE and outpatient vs inpatient treatment for deep venous thrombosis. The authors implicitly chose the same noninferiority margin (4%) for bleeding, although they provide no justification for this choice.

When subsequently reviewing their results, if investigators find that the CI around the estimate for the difference in primary outcome events lies entirely below their chosen noninferiority

threshold, they will claim noninferiority (Figure 7-1, scenario B) or even, in some instances, superiority of the experimental treatment (Figure 7-1, scenario A). If, on the other hand, the CI crosses the threshold, the trial has failed to establish noninferiority (Figure 7-1, scenario C). If the CI lies wholly above the noninferiority threshold, then the experimental treatment is inferior to standard treatment (Figure 7-1, scenario D).

If noninferiority trials choose insufficiently stringent thresholds, they run the risk of concluding noninferiority when, actually, many patients would be unwilling to accept the experimental treatment if they were informed of the largest possible increased risk (ie, decreased effectiveness) associated with its use. If these choices of thresholds go uncontested, wide uptake of experimental treatments could prove detrimental to patients. In interpreting noninferiority thresholds, we will encourage you to use your own judgment rather than accepting that of the investigators, relieving you of the need to decipher what many may experience as obscure statistical reasoning used to define the thresholds.

Although others have explained the rationale and provided criteria for interpreting noninferiority trials,<sup>2,3,5-9</sup> this chapter strives to present a simple and practical approach based on *Users' Guides* principles. We will use contemporary examples to illustrate concepts that can guide optimal clinical practice. In doing so, we follow the 3-step approach of other *Users' Guides* chapters, focusing on issues of validity, interpretation of results, and applicability of results specific to noninferiority trials (Box 7-1).

## ARE THE RESULTS VALID?

---

Limitations of study design of noninferiority trials include issues beyond risk of bias. Thus, in this chapter, we continue to use the term “validity” to address both risk of bias and these additional issues.

## BOX 7-1

## Users' Guides Approach to Evaluating a Noninferiority Trial

### Are the results valid?<sup>a</sup>

Did experimental and standard treatment groups start with the same prognosis?

Was prognostic balance maintained as the trial progressed?

Were the groups prognostically balanced at the completion of the trial?

*Did the investigators guard against an unwarranted conclusion of noninferiority?<sup>b</sup>*

*Was the effect of the standard treatment preserved?*

*Did the investigators analyze patients according to the treatment they received and to the groups to which they were assigned?*

### What are the results?

#### How can I apply the results to patient care?

Were the study patients similar to my patient?

Were all patient-important outcomes considered?

Are the likely advantages of the experimental treatment worth the potential harm and costs?<sup>b</sup>

<sup>a</sup>Limitations of study design of noninferiority trials include issues beyond risk of bias. Thus, in this chapter, we continue to use the term "validity" to address both risk of bias and these additional issues.

<sup>b</sup>Includes issues specific to noninferiority trials.

The question "Are the results valid?" asks to what extent the results are likely to represent an unbiased estimate of effect vs systematic overestimates or underestimates. As with other studies that address disease management questions, noninferiority trials will reduce the risk of bias if they ensure concealed randomization; demonstrate balance of known *prognostic factors*;

*blind* patients, clinicians, and outcome assessors; and ensure complete *follow-up* (see Chapter 6, Therapy [Randomized Trials]). Noninferiority trials are, however, vulnerable to misleading conclusions in ways that superiority trials are not. Although not strictly related to the risk of bias, we have classified the relevant concerns, italicized in Box 7-1, as issues of validity.

## **Did the Investigators Guard Against an Unwarranted Conclusion of Noninferiority?**

### **Was the Effect of the Standard Treatment Preserved?**

One way to achieve apparent noninferiority is to suboptimally administer the standard treatment. Suboptimal treatment can include enrolling patients less likely to be adherent or responsive to standard treatment; enrolling a population at low risk of the effectiveness outcome, particularly if the noninferiority threshold is expressed in absolute terms; reducing treatment intensity or administering treatment by a suboptimal route (eg, orally rather than intravenously); or terminating follow-up before treatment effects are fully manifest. One strategy to assess whether the treatment effect was likely to have been preserved would be to evaluate the extent to which the design and conduct of the study attempted to overcome each of these threats to the standard treatment effect.

Another way to determine whether the effect of standard treatment has been preserved is to compare the event rate in the noninferiority trial with those seen in historical trials that involve the standard treatment. A higher *control event rate* in the standard treatment group in the noninferiority trial compared with the typical rate seen in historical trials would raise the suspicion of suboptimal administration of the standard treatment. Unfortunately, the competing explanation—prognostic differences between the populations enrolled in noninferiority vs historical trials—is also likely. Comparing patient characteristics among the trials could help decide which

of the competing explanations is more likely, but the possibility remains that unmeasured prognostic features are responsible for the observed difference in event rates.

Take, for instance, the trial Rivaroxaban Once Daily Oral Direct Factor Xa Compared With Vitamin K Antagonism for Prevention of Stroke and Embolism in Atrial Fibrillation (ROCKET AF), in which investigators declared rivaroxaban to be noninferior to warfarin in managing patients with atrial fibrillation.<sup>10</sup> Concerns exist about the extent to which the patients treated with warfarin remained within the therapeutic range of anticoagulation throughout this study in comparison with previous RCTs comparing warfarin with placebo. Investigators documented a mean time in therapeutic range (TTR) of 55% in the warfarin group in ROCKET AF—considerably less than rates of approximately 75% (range, 42%-83%) seen in prior studies<sup>11,12</sup> and in contemporary noninferiority trials.<sup>12</sup> Hence, we cannot be confident that the warfarin treatment effect was preserved in ROCKET AF. The apparent noninferiority of rivaroxaban to warfarin may be because the latter was suboptimally administered.<sup>13</sup>

Using the second criterion to determine whether the effect of standard treatment has been preserved, the rate of stroke or systemic embolism in the warfarin group was lower in the ROCKET AF trial than has been seen historically,<sup>14</sup> despite the fact that patients in the ROCKET AF trial were older and had a higher prevalence of hypertension and type 2 diabetes mellitus than those in the previous trials.<sup>11</sup> Thus, control event rates fail to support the suspicion of suboptimal warfarin administration in the control group. The low TTR, nevertheless, remains concerning.

## Did the Investigators Analyze Patients According to the Treatment They Received and to the Groups to Which They Were Assigned?

Another issue has to do with how investigators dealt with patients who were randomized and followed up to the end of the study but who did not take their medication as intended or did not use it at all. The purpose of randomization is to ensure that prognostic factors for the outcome of interest are balanced between treatment groups. It is likely that those who do not adhere to the allocated treatment as set out in the study protocol are prognostically different from those who do.<sup>15</sup>

Investigators may be tempted to include only those individuals who were adherent to study protocol and omit those who were not (often called a *per-protocol analysis*). This is likely, however, to compromise the prognostic balance that randomization created in the first place. Because, more often than not, *nonadherent* patients are prognostically worse than adherent patients, the omission of those who failed to adhere to the experimental treatment is likely to bias results toward an overestimation of treatment benefit in a superiority trial. In contrast, an analyze-as-randomized approach (*intention-to-treat analysis*) analyzes patients in the groups to which they were assigned irrespective of the level of patient adherence. As a result, it yields an unbiased—and typically more conservative—estimate of treatment effectiveness in a superiority trial.<sup>16</sup>

Unfortunately, the analyze-as-randomized approach has serious limitations in the context of noninferiority trials. Picture a noninferiority trial in which the experimental treatment is actually substantially inferior to the current standard. Let us further suppose that, in this trial, many patients in the standard treatment group do not, for whatever reason, adhere to treatment. In the analyze-as-randomized approach, inclusion of these nonadherent patients may result in a substantial underestimate of the benefit of standard treatment and thus cause a misleading inference of noninferiority in comparison with the experimental treatment.

The per-protocol analysis, which focuses only on those who use the treatment more or less as directed, likely introduces prognostic imbalance but can nevertheless provide some reassurance regarding noninferiority. If the results of such an analysis are consistent with those from the analyze-as-randomized approach and if both lie below the noninferiority threshold, our inference regarding noninferiority is strengthened. If, however, there are important differences between the results of the 2 analyses, the inference of noninferiority is weakened.

For example, the Cardiac Insufficiency Bisoprolol Study (CIBIS) III trial addressed the initial use of a  $\beta$ -blocker rather than an angiotensin-converting enzyme (ACE) inhibitor for preventing deaths or hospitalization in patients with heart failure.<sup>17</sup> The investigators set a noninferiority threshold of a 5% absolute increase in the primary *end point* of death or hospitalization with  $\beta$ -blocker use. The as-randomized analysis met their noninferiority threshold: the upper limit of the CI suggested that an increase in death or hospitalization greater than 4.4% with  $\beta$ -blockers was unlikely. In the per-protocol analysis, however, the upper limit of the CI was 5.1%, just above the investigators' chosen threshold. Were one to accept the authors' threshold, the inference of noninferiority is weakened by the results of the per-protocol analysis. Whether one should accept the authors' threshold at all is a point to which we will return.

## USING THE GUIDE

In the pulmonary embolism treatment trial,<sup>1</sup> the investigators randomized 344 patients with acute symptomatic pulmonary embolus at low risk of death to outpatient treatment for 5 or more



days of inpatient treatment. *Allocation concealment* was ensured via a central computer randomization system. Neither patients nor their caregivers were blinded to the allocated treatment, but adjudicators of outcome were. Patients in the treatment and *control* groups were similar with respect to known prognostic factors, including location of the embolus, comorbidity, and clinical findings. Complete follow-up was achieved in all but 5 patients. Although the lack of blinding raises concern, blinding of the outcome assessors provides a safeguard against *risk of bias*.

The crucial issues in optimal administration of the standard intervention in this study are the duration patients in the hospitalized group received LMWH and the TTR during subsequent warfarin treatment. Patients spent a mean of 8.9 days receiving LMWH, as long or longer than the standard in many settings (and thus satisfactory). The TTR was only 52%, which is suboptimal and raises concern. However, the TTR in the outpatient group was also 52%, substantially ameliorating the concern.

The investigators conducted both an analysis-as-randomized and a per-protocol analysis, which excluded patients in the hospitalized group discharged within 24 hours and those in the outpatient group discharged more than 24 hours after randomization. As you will see in the results that we present below, the per-protocol results do not substantially differ from the as-randomized results.

In conclusion, although the trial has some limitations in risk of bias, we would conclude moderate to high credibility of its findings.

## WHAT ARE THE RESULTS?

---

The relevant results of a noninferiority trial focus on the following: (1) the difference between experimental and standard treatment in the effectiveness outcome that is the primary

target of treatment, (2) the harm and burden outcomes that should favor the experimental over the standard treatment, and (3) whether the results provide reassurance that the standard treatment was optimally administered.

## USING THE GUIDE

**F**or pulmonary embolism, the primary effectiveness outcome is reducing recurrent VTE, and the treatment burden (staying in the hospital rather than being treated at home) is easily measured. Another important issue is the incidence of major bleeding, which could be conceptualized as an additional outcome warranting a noninferiority inquiry. Even if outpatient care was noninferior to inpatient care with respect to the primary effectiveness outcome, patients may choose to remain in the hospital if the risks of serious bleeding are substantially higher at home.

For each outcome, we are interested in the point estimate (the best estimate) of the difference in event rates between experimental and standard treatments and its associated CI. The boundaries of the CI represent the range of plausible truth—less likely than the point estimate but still plausible (see Chapter 9, Confidence Intervals: Was the Single Study or Meta-analysis Large Enough?). Herein, we focus on the absolute differences between groups at 90 days. In the as-randomized analysis, recurrent VTE occurred in 1 individual in the outpatient group and none in the inpatient group, a difference of 0.6% or 6 in 1000, with an upper boundary of the 95% CI of 2.7% (27 more VTEs in 1000 outpatients).<sup>1</sup> This result suggests that it is unlikely that the recurrent VTE rate among outpatients is more than 4% (40 in 1000) greater than among inpatients ( $P = .01$ ), the authors' noninferiority threshold.

For serious bleeding, the investigators observed 3 events in the outpatient group and none in the inpatient group (1.8%, or 18 in 1000 more bleeds in outpatients). The upper boundary of the 95% CI is 4.5%, which exceeds the authors' 4% threshold and therefore fails the statistical test of noninferiority ( $P = .09$ ).

The authors also present a per-protocol analysis, the results of which are consistent with the as randomized results. The major bleeding outcome is actually more favorable to outpatient management (a difference of 1.2% favoring inpatient management, with an upper boundary of the 95% CI of 3.8%; the *P* against the 4% threshold is .04).

## HOW CAN I APPLY THE RESULTS TO PATIENT CARE?

---

In applying findings from the medical literature to individual patient care, we suggest asking 3 questions (Box 7-1), of which one—assessing trade-offs between an experimental treatment's likely advantages and potential harm and costs—includes issues specific to noninferiority trials.

### **Are the Likely Advantages of the Experimental Treatment Worth the Potential Harm and Costs?**

Is a particular noninferiority trial simply a failed superiority trial, portrayed to put a happy face on a sad result? When investigators plan their trials, they specify the analysis, and this specification has implications for how results are interpreted. It is the job of the editors to ensure that only trials planned as noninferiority are in fact reported in published articles as noninferiority trials. Unfortunately, editors are not always thorough in performing due diligence in this aspect (and others) of reporting.<sup>18</sup>

The risk that a trial reported as noninferiority may not have been planned as noninferiority again highlights the importance of an independent judgment of the noninferiority threshold. You may be tempted to turn to the authors of a study for guidance on assessing the key inferences from a noninferiority trial: are

the advantages of the experimental treatment worth the risks of loss of effectiveness? In doing so, you are implicitly accepting the authors' noninferiority threshold. For various reasons, investigators may have an incentive to be as lenient as possible with the choice of noninferiority threshold. Thus, accepting that threshold may not serve your patients' best interests.

Consider first the CIBIS-III trial, which investigates the substitution of  $\beta$ -blockers for ACE inhibitors in the initial treatment of heart failure that we used to illustrate the desirability of a per-protocol analysis.<sup>17</sup> The as-randomized and per-protocol results straddled the authors' noninferiority margin of 5%. But is that margin appropriate? The harms or convenience advantages of  $\beta$ -blockers over ACE inhibitors are few, if any. Thus, patients are unlikely to accept starting with  $\beta$ -blockers if it really meant an absolute increase of up to 5% in the end point of death or hospitalization.

Consider next the Post-Operative Radiation Therapy for Endometrial Carcinoma 2 (PORTEC-2) trial, which investigated the effect of vaginal brachytherapy (VBT) vs pelvic external beam radiotherapy (EBRT) on the primary outcome of vaginal recurrence of endometrial carcinoma.<sup>19</sup> The investigators set a noninferiority threshold of a risk difference of 6%—an increase in the primary outcome of 6 events in 100 patients—between the 2 groups at 5 years. After analyzing the data, they declared the VBT regimen to be noninferior to EBRT on the basis that the upper boundary of the CI—an absolute difference of 5%—fell below their threshold. Although patients undergoing VBT report better health-related quality of life than those receiving EBRT,<sup>19</sup> for an outcome as serious as cancer recurrence, we suspect that few patients would be willing to choose the VBT approach if the actual increase was as great as 5%.

The noninferiority threshold implies a trade-off between the advantages of the experimental treatment and the potential loss in effectiveness. Making this trade-off may be a challenging judgment, but it is not fundamentally different from other patient management decisions: they all involve trading off the desirable and undesirable consequences of the alternatives. They therefore involve *value and preference* judgments, and it is the preferences of the individual patient that must drive the decision. When the trade-off between desirable and undesirable consequences is a close one, the best—some would argue the only—way to ensure the chosen course of action is right for the individual is through shared decision making (see Chapter 18, Decision Making and the Patient).

In preparing for shared decision making with your patients, and being cognizant of the limited time you and they may have to spend on this activity, it may be worthwhile to reflect on the values and preferences of your typical patient and the implications for the noninferiority threshold. To gain a better understanding of how your typical patient perceives benefits and risks, you may want to refer to published studies that provide insight into patients' values and preferences.<sup>20</sup>

If, given the benefits and harms of an experimental intervention, you perceive all or virtually all patients would make the same decision, you and your patient may be able to quickly come to a fully satisfactory decision (see Chapter 17, How to Use a Patient Management Recommendation: Clinical Practice Guidelines and Decision Analyses). If, however, the desirable and undesirable consequences are more closely balanced, you will need to have a detailed discussion with your patients.

Considering the most appropriate noninferiority margin will help distinguish between these 2 situations. First, look at the upper boundary of the CI for the primary outcome; then, note the extent to which it exceeds the maximum increase in risk of the primary outcome that your patients would, on average, be willing to accept in exchange for the experimental treatment's reduction in harms or burden.

If the upper boundary is substantially greater than your threshold and very few, if any, of your patients would choose the intervention, decision making may be expeditious. If, however, the upper boundary of the CI is near your threshold—that is, the balance between desirable and undesirable consequences is a close one—ensuring the right decision will involve full exploration of your patients’ views of the trade-off at hand.

## CLINICAL SCENARIO RESOLUTION

**Y**our patient’s clinical profile suggests a relatively low risk of death from pulmonary embolism. She would thus have been eligible for the trial,<sup>1</sup> and its results are directly applicable to her care. Point estimates suggest similar and low risks of recurrent VTE (6 in 1000); the difference in important bleeding is somewhat greater (18 more bleeds per 1000 in the outpatient group). The CIs raise more concern and include an increase in embolism of 2.7% (27 in 1000) and an increase in bleeding of 4.5% (45 in 1000), both within 90 days, in the outpatient care group.

Because their noninferiority margin for VTE has been met, the authors of the pulmonary embolism trial conclude that “[i]n selected low-risk patients with pulmonary embolism, outpatient care can safely and effectively be used in place of inpatient care.”<sup>1</sup> Individuals who, all else being equal, would much prefer home treatment and are ready to focus on the point estimates that suggest that rates of adverse events (at least VTE) are likely similar with outpatient management might agree. On the other hand, risk-averse individuals who perceive the possibility of increased risk of VTE and bleeding with outpatient management as not being worth the benefit of receiving treatment at home would not agree with this conclusion. We believe that there are likely to be a substantial number of such risk-averse individuals. Reliance on the authors’ noninferiority would not serve such patients well.

## CONCLUSIONS

---

Critical appraisal of noninferiority studies closely follows the principles and criteria for assessing any study of experimental management strategies. With respect to validity, assessment of a noninferiority study requires special attention to the optimal use of the standard treatment and to the results of the as-randomized and per-protocol analyses. With respect to the trade-offs between desirable and undesirable consequences in noninferiority trials, close attention to best estimates and CIs around the difference in effectiveness outcomes between experimental and standard treatments is needed. In particular, clinicians should consider whether patients would be willing to accept loss in the effectiveness outcome suggested by the upper boundary of the 95% CI, irrespective of whether this interval lies below or above the investigators' choice of noninferiority threshold.

## References

1. Aujesky D, Roy PM, Verschuren F, et al. Outpatient versus inpatient treatment for patients with acute pulmonary embolism: an international, open-label, randomised, non-inferiority trial. *Lancet*. 2011;378(9785):41-48.
2. Fleming TR. Current issues in non-inferiority trials. *Stat Med*. 2008;27(3):317-332.
3. Kaul S, Diamond GA. Good enough: a primer on the analysis and interpretation of noninferiority trials. *Ann Intern Med*. 2006;145(1):62-69.
4. Temple R, O'Neill R. *Guidance for Industry Non-Inferiority Clinical Trials*. Rockville, MD: Food and Drug Administration, Dept of Health and Human Services; 2010.
5. Le Henanff A, Giraudeau B, Baron G, Ravaud P. Quality of reporting of noninferiority and equivalence randomized trials. *JAMA*. 2006;295(10):1147-1151.
6. Piaggio G, Elbourne DR, Pocock SJ, Evans SJW, Altman DG; CONSORT Group. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. *JAMA*. 2012;308(24):2594-2604.
7. Scott IA. Non-inferiority trials: determining whether alternative treatments are good enough. *Med J Aust*. 2009;190(6):326-330.

8. Gøtzsche PC. Lessons from and cautions about noninferiority and equivalence randomized trials. *JAMA*. 2006;295(10):1172-1174.
9. Schumi J, Wittes JT. Through the looking glass: understanding non-inferiority. *Trials*. 2011;12:106.
10. Patel MR, Mahaffey KW, Garg J, et al; ROCKET AF Investigators. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. *N Engl J Med*. 2011;365(10):883-891.
11. Jackson K, Gersh BJ, Stockbridge N, et al; Duke Clinical Research Institute/American Heart Journal Expert Meeting on Antithrombotic Drug Development for Atrial Fibrillation. Antithrombotic drug development for atrial fibrillation: proceedings, Washington, DC, July 25-27, 2005. *Am Heart J*. 2008;155(5):829-840.
12. Granger CB, Alexander JH, McMurray JJ, et al; ARISTOTLE Committees and Investigators. Apixaban versus warfarin in patients with atrial fibrillation. *N Engl J Med*. 2011;365(11):981-992.
13. Fleming TR, Emerson SS. Evaluating rivaroxaban for nonvalvular atrial fibrillation—regulatory considerations. *N Engl J Med*. 2011;365(17):1557-1559.
14. Hart RG, Benavente O, McBride R, Pearce LA. Antithrombotic therapy to prevent stroke in patients with atrial fibrillation: a meta-analysis. *Ann Intern Med*. 1999;131(7):492-501.
15. Kunz R, Guyatt G. Which patients to include in the analysis? *Transfusion*. 2006;46(6):881-884.
16. Montori VM, Guyatt GH. Intention-to-treat principle. *CMAJ*. 2001;165(10):1339-1341.
17. Willenheimer R, van Veldhuisen DJ, Silke B, et al; CIBIS III Investigators. Effect on survival and hospitalization of initiating treatment for chronic heart failure with bisoprolol followed by enalapril, as compared with the opposite sequence: results of the randomized Cardiac Insufficiency Bisoprolol Study (CIBIS) III. *Circulation*. 2005;112(16):2426-2435.
18. Yank V, Rennie D, Bero LA. Financial ties and concordance between results and conclusions in meta-analyses: retrospective cohort study. *BMJ*. 2007;335(7631):1202-1205.
19. Nout RA, Smit VT, Putter H, et al; PORTEC Study Group. Vaginal brachytherapy versus pelvic external beam radiotherapy for patients with endometrial cancer of high-intermediate risk (PORTEC-2): an open-label, non-inferiority, randomised trial. *Lancet*. 2010;375(9717):816-823.
20. MacLean S, Mulla S, Akl EA, et al; American College of Chest Physicians. Patient values and preferences in decision making for antithrombotic therapy: a systematic review: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians EvidenceBased Clinical Practice Guidelines. *Chest*. 2012;141(2)(suppl):e1S-e23S.



*This page intentionally left blank*

8

# Does Treatment Lower Risk? Understanding the Results

Waleed Alhazzani, Stephen D. Walter, Roman Jaeschke, Deborah J. Cook, and Gordon Guyatt

## IN THIS CHAPTER

---

The 2 × 2 Table

The Risk

The Risk Difference (Absolute Risk Reduction)

The Relative Risk

The Relative Risk Reduction

The Odds Ratio

Relative Risk vs Risk Difference: Why the Fuss?

The Number Needed to Treat

The Number Needed to Harm

Confidence Intervals

Survival Data

Which Measure of Association Is Best?

---

When clinicians consider the results of clinical trials, they are interested in the association between a treatment and an outcome. This chapter will help you understand and interpret study results related to outcomes that are either present or absent (*dichotomous* or binary) for each patient. Such binary outcomes include death, stroke, myocardial infarction, hospitalization, or disease exacerbations. A guide for teaching the *concepts* in this chapter is also available.<sup>1</sup>

## THE 2 × 2 TABLE

Table 8-1 is a 2 × 2 table that captures the information for a dichotomous outcome of a clinical trial.

**TABLE 8-1**

The 2 × 2 Table		
Exposure	Outcome	
	Yes	No
Yes	<i>a</i>	<i>b</i>
No	<i>c</i>	<i>d</i>
Risk with exposure = $a / (a + b)$		
Risk without exposure = $c / (c + d)$		
Odds with exposure = $a / b$		
Odds without exposure = $c / d$		
Relative risk = $\frac{a / (a + b)}{c / (c + d)}$		
Relative risk reduction = $\frac{c / (c + d) - a / (a + b)}{c / (c + d)}$		
Risk difference <sup>a</sup> = $\frac{c}{c + d} - \frac{a}{a + b}$		

(Continued)

TABLE 8-1

**The 2 × 2 Table (Continued)**


---

Number needed to treat = 100 / (risk difference expressed as %)

---

$$\text{Odds ratio} = \frac{a/b}{c/d} = \frac{ad}{cb}$$


---

<sup>a</sup>Also known as the absolute risk reduction.

For instance, during a *randomized trial* that compares mortality rates in patients with bleeding esophageal varices that were controlled by endoscopic ligation or endoscopic sclerotherapy,<sup>2</sup> 18 of 64 participants assigned to ligation died, as did 29 of 65 patients assigned to sclerotherapy (Table 8-2).

TABLE 8-2

**Results From a Randomized Trial of Endoscopic Sclerotherapy Compared With Endoscopic Ligation for Bleeding Esophageal Varices<sup>a</sup>**

Exposure	Outcome		Total
	Death	Survival	
Ligation	18	46	64
Sclerotherapy	29	36	65
Relative risk = (18/64) / (29/65) = 0.63 or 63%			
Relative risk reduction = 1 – 0.63 = 0.37 or 37%			
Risk difference = 0.446 – 0.281 = 0.165 or 16.5%			
Number needed to treat = 100 / 16.5 = 6			
Odds ratio = (18/46) / (29/36) = 0.39 / 0.80 = 0.49 or 49%			

<sup>a</sup>Data from Stiegmann et al.<sup>2</sup>

## THE RISK

---

The simplest measure of occurrence to understand is the *risk* (or *absolute risk*). We often refer to the risk of the adverse outcome in the *control group* as the *baseline risk*, the *control group risk*, or, occasionally, the *control event rate*.

The risk of dying in the ligation group is 28% (18/64 or  $[a/(a + b)]$ ), and the risk of dying in the sclerotherapy group is 45% (29/65 or  $[c/(c + d)]$ ).

## THE RISK DIFFERENCE (ABSOLUTE RISK REDUCTION)

---

One way of comparing 2 risks is by calculating the absolute difference between them. We refer to this difference as the *absolute risk reduction* (ARR) or the *risk difference* (RD). Algebraically, the formula for the RD (the control group risk minus the treatment group risk) is  $[c/(c + d)] - [a/(a + b)]$  (Table 8-1). This measure of effect uses absolute rather than relative terms in looking at the proportion of patients who are spared the adverse outcome.

In our example, the RD is  $0.446 - 0.281$  or 0.165 (ie, an RD of 16.5%).

## THE RELATIVE RISK

---

Another way to compare the risks in the 2 groups is to take their ratio; this is called the *relative risk* or *risk ratio* (RR). The RR tells us the proportion of the original risk (in this case,

the risk of death in patients who received sclerotherapy) that is still present when patients receive the *experimental treatment* (in this case, ligation). From our  $2 \times 2$  table, the formula for this calculation is  $[a/(a + b)] / [c/(c + d)]$  (Table 8-1).

In our example, the RR of dying after receiving initial ligation vs sclerotherapy is 18/64 (the risk in the ligation group) divided by 29/65 (the risk in the sclerotherapy group) or 0.63. In everyday English, we would say that the risk of death with ligation is approximately two-thirds of that with sclerotherapy.

## THE RELATIVE RISK REDUCTION

---

An alternative relative measure of treatment effectiveness is the *relative risk reduction* (RRR), an estimate of the proportion of baseline risk that is removed by the therapy. It may be calculated as  $1 - RR$ . One also can calculate the RRR by dividing the RD (amount of risk removed) by the absolute risk in the control group (Table 8-1).

In our bleeding varices example, where the RR was 0.63, the RRR is thus  $1 - 0.63$  (or 16.5% divided by 44.6%, the risk in the sclerotherapy group); either way, it comes to 0.37. In other words, ligation decreases the risk of death by just more than one-third compared with sclerotherapy.

## THE ODDS RATIO

---

Instead of looking at the risk of an event, we could estimate the odds of having vs not having an event. When considering the

effects of therapy, you usually will not go far wrong if you interpret the *odds ratio* (OR) as equivalent to the RR. The exception is when the risk of having an event is very high—for instance, when more than 40% of control patients experience myocardial infarction or death.

## RELATIVE RISK VS RISK DIFFERENCE: WHY THE FUSS?

---

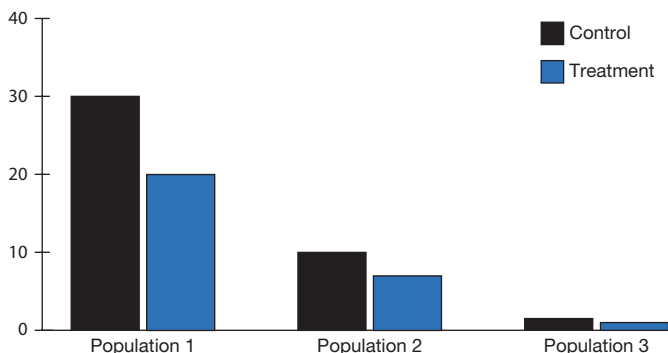
Failing to distinguish between the OR and the RR when interpreting randomized trial results will seldom mislead you; you must, however, distinguish between the RR and the RD. The reason is that the RR is generally far larger than the RD, and presentations of results in the form of RR (or RRR) can convey a misleading message. Furthermore, it is the risk difference in which the patient is ultimately interested. Reducing a patient's risk by 50% sounds impressive. That may, however, represent a reduction in risk from 2% to 1%. The corresponding 1% RD sounds considerably less impressive and in fact conveys the crucial information.

As depicted in [Figure 8-1](#), consider a treatment that is administered to 3 different subpopulations of patients and that, in each case, decreases the risk by one-third (RRR, 0.33; RR, 0.67). When administered to a subpopulation with a 30% risk of dying, treatment reduces the risk to 20%. When administered to a population with a 10% risk of dying, treatment reduces the risk to 6.7%. In the third population, treatment reduces the risk of dying from 1% to 0.67%.

Although treatment reduces the risk of dying by one-third in each population, this piece of information is not adequate to fully capture the impact of treatment. What if the treatment under consideration is a toxic cancer chemotherapeutic drug associated with severe adverse effects in 50% of those to whom it is administered? Under these circumstances, most patients in the lowest risk group in [Figure 8-1](#), whose RD is only 0.3%,

FIGURE 8-1

## Constant Relative Risk With Varying Risk Differences



would likely decline treatment. In the intermediate population, those with an absolute reduction in risk of death of approximately 3%, some might accept the treatment, but many would likely decline. Many in the highest-risk population with an absolute benefit of 10% would likely accept the treatment, but some may not.

We suggest that you consider the RRR in light of your patient's baseline risk. For instance, you might expect an RRR of approximately 25% in vascular events in patients with possible cardiovascular disease with administration of statins. You would view this RRR differently in a 40-year-old woman without hypertension, diabetes mellitus, or a history of smoking with a mildly elevated low-density lipoprotein level (5-year risk of a cardiovascular event of approximately 2%, ARR of approximately 0.5%) and a 70-year-old woman with hypertension and diabetes who smokes (5-year risk of 30%, ARR of 7.5%). All of this assumes a constant RRR across risk groups; fortunately, a more or less constant RRR is usually the case, and we suggest you make that assumption unless there is evidence that suggests it is incorrect.<sup>3-5</sup>



## THE NUMBER NEEDED TO TREAT

---

The impact of treatment also can be expressed by the number of patients you would need to treat to prevent an adverse event, the *number needed to treat* (NNT).<sup>6</sup> Table 8-2 indicates that the risk of dying is 28.1% in the ligation group and 44.6% in the sclerotherapy group, an RD of 16.5%. If treating 100 patients results in avoiding 16.5 events, how many patients do we need to treat to avoid 1 event? The answer: 100 divided by 16.5, or approximately 6, is the NNT.

The NNT calculation always implies a given time of *follow-up* (ie, do we need to treat 50 patients for 1 year or 5 years to prevent an event?). When trials with long follow-ups are analyzed by survival methods, there are a variety of ways of calculating the NNT (see the following subsection, Survival Data). These different methods will, however, rarely lead to results with different clinical implications.<sup>7</sup>

Assuming a constant RRR, the NNT is inversely related to the proportion of patients in the control group who have an adverse event. For instance, if the control group risk doubles, the NNT will decrease by a factor of 2 (ie, be half of what it was). If the risk of an adverse event doubles (eg, if we deal with patients at a higher risk of death than those included in the clinical trial), we need to treat only half as many patients to prevent an adverse event. On the other hand, if the risk decreases by a factor of 4 (patients are younger and have less *comorbidity* than those in the study), we will have to treat 4 times as many people.

The NNT also is inversely related to the RRR. With the same baseline risk, a more effective treatment with twice the RRR will reduce the NNT by half. If the RRR with one treatment is only a quarter of that achieved by an alternative strategy, the NNT will be 4 times greater.

Table 8-3 presents hypothetical data that illustrate these relationships.

**TABLE 8-3****Association Among the Baseline Risk, Relative Risk Reduction, and Number Needed to Treat<sup>a</sup>**

<b>Control Group Risk</b>	<b>Experimental Group Risk</b>	<b>Relative Risk, %</b>	<b>Relative Risk Reduction, %</b>	<b>Risk Difference, %</b>	<b>Number Needed to Treat</b>
0.02 or 2%	0.01 or 1%	50	50	1	100
0.4 or 40%	0.2 or 20%	50	50	20	5
0.04 or 4%	0.02 or 2%	50	50	2	50
0.04 or 4%	0.03 or 3%	75	25	1	100
0.4 or 40%	0.3 or 30%	75	25	10	10
0.01 or 1%	0.005 or 0.5%	50	50	0.5	200

<sup>a</sup>Relative risk = experimental group risk/control group risk; relative risk reduction = 1 – relative risk; risk difference = control group risk – experimental group risk; number needed to treat = 100/risk difference in %.



## THE NUMBER NEEDED TO HARM

---

Clinicians can calculate the *number needed to harm* (NNH) in a similar way. If you expect 5 of 100 patients to become fatigued when taking a  $\beta$ -blocker for a year, of 20 patients you treat, 1 will become tired; therefore, the NNH is 20.

## CONFIDENCE INTERVALS

---

We have presented all of the measures of association of the treatment with ligation vs sclerotherapy as if they represented the true effect. The results of any experiment, however, represent only an estimate of the truth. The true effect of treatment may be somewhat greater—or less—than what we observed. The *confidence interval* (CI) tells us, within the bounds of plausibility (and assuming a low *risk of bias*), how much greater or smaller the true effect is likely to be (see Chapter 9, Confidence Intervals: Was the Single Study or Meta-analysis Large Enough?).

## SURVIVAL DATA

---

Analysis of a  $2 \times 2$  table implies an examination of the data at a specific point in time. This analysis is satisfactory if we are looking for events that occur within relatively short periods and if all patients have the same duration of follow-up. In longer-term studies, however, we are interested not only in the total number of events but also in their timing. For instance, we may focus on whether therapy for patients with a uniformly fatal condition (eg, unresectable lung cancer) delays death.

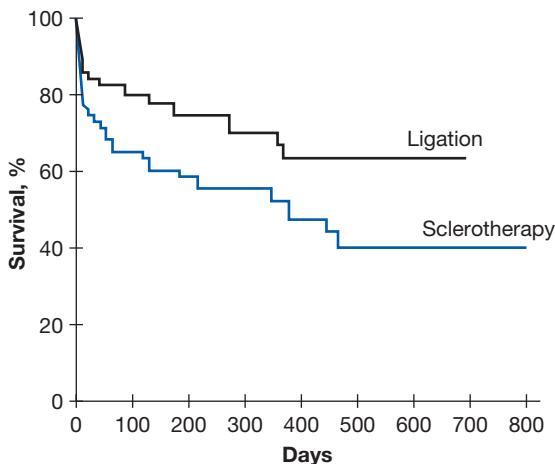
When the timing of events is important, investigators could present the results in the form of several  $2 \times 2$  tables constructed at different points of time after the study began. For example, Table 8-2 represents the situation after the study was

finished. Similar tables could be constructed describing the fate of all patients available for analysis after their enrollment in the trial for 1 week, 1 month, 3 months, or whatever time we chose to examine. The analysis of accumulated data that takes into account the timing of events is called *survival analysis*. Do not infer from the name, however, that the analysis is restricted to deaths; in fact, any dichotomous outcome occurring over time will qualify.

The *survival curve* of a group of patients describes their status at different times after a defined starting point.<sup>8</sup> In Figure 8-2, we show the survival curve from the bleeding varices trial. Because the investigators followed up some patients for a longer time, the survival curve extends beyond the mean follow-up of approximately 10 months. At some point, prediction becomes imprecise because there are few patients remaining to estimate

**FIGURE 8-2**

**Survival Curves for Ligation and Sclerotherapy**



the *probability* of survival. The CIs around the survival curves capture the precision of the estimate.

Even if the true RR, or RRR, is constant throughout the duration of follow-up, the play of chance will ensure that the *point estimates* differ. Ideally then, we would estimate the overall RR by applying an average, weighted for the number of patients available, for the entire survival experience. Statistical methods allow just such an estimate. The probability of events occurring at any point in each group is referred to as the hazard for that group, and the weighted RR during the entire study duration is known as the *hazard ratio*.

A major advantage of using survival analysis is the ability to account for differential length of follow-up. In many trials of a fixed duration, some patients are enrolled early and thus have long follow-up and some later with consequently shorter follow-up. Survival analysis takes into account both those with shorter (by a process called *censoring*) and those with longer follow-up, and all contribute to estimates of hazard and the hazard ratio. Patients are censored at the point at which they are no longer being followed up. Appropriate accounting for those with differential length of follow-up is not possible in  $2 \times 2$  tables that deal only with the number of events.

“Competing risks” is an issue that arises when one event influences the likelihood of another event. The most extreme example is death: if the outcome is stroke, people who die can no longer have a stroke. Competing risks also can arise when there are 2 or more outcome events among living patients (for instance, if a patient has a stroke, the likelihood of a subsequent transient ischemic attack may decrease). Investigators can deal with the problem of competing risks by censoring patients at the time of the “competing” events (death and stroke in the previous examples). The censoring approach, however, has its limitations.<sup>9</sup>

Specifically, the usual assumption is that the censored events are independent of the main outcome of interest, but in practice this assumption may not be correct. In our example, it

is probable that patients who experience myocardial infarction have a higher death rate than those without myocardial infarction, and this would violate the assumption of independence. Investigators also sometimes use censoring for those *lost to follow-up*. This is much more problematic because the censoring assumes that those with shorter follow-up are similar to those with longer follow-up—the only difference, indeed, being length of follow-up. Because loss to follow-up may be associated with a higher or lower likelihood of events (and thus, those lost differ from those who are followed up), the censoring approach does not deal with the risk of bias associated with loss to follow-up.<sup>9</sup>

## WHICH MEASURE OF ASSOCIATION IS BEST?

---

As *evidence-based practitioners*, we must decide which measure of association deserves our focus. Does it matter? The answer is yes. The same results, when presented in different ways, may lead to different treatment decisions.<sup>9-13</sup> For example, Forrow et al<sup>10</sup> found that clinicians were less inclined to treat patients after presentation of trial results as the absolute change in the outcome compared with the relative change in the outcome. In a similar study, Naylor et al<sup>11</sup> found that clinicians rated the effectiveness of an intervention lower when events were presented in absolute terms rather than using RRR. Moreover, clinicians offered lower effectiveness ratings when they viewed results expressed in terms of NNT than when they saw the same data as RRRs or ARRs. The awareness of this phenomenon in the pharmaceutical industry may be the reason for their propensity to present physicians with treatment-associated RRRs.

Patients are as susceptible as clinicians to how results are communicated. In one study, when researchers presented patients with a hypothetical scenario of life-threatening illness, the patients were more likely to choose a treatment described in terms of RRR than in terms of the corresponding ARR.<sup>14</sup> Other investigators found similar results.<sup>15,16</sup>

Considering how our interpretations differ with data presentations, we are best advised to consider all of the data (as either a  $2 \times 2$  table or a survival analysis) and then reflect on both the relative and the absolute figures. As you examine the results, you will find that if you can estimate your patient's baseline risk, knowing how well the treatment works—expressed as an RR or RRR—allows you to estimate the patient's risk with treatment. Considering the RD—the difference between the risk with and without treatment—and its reciprocal, the NNT, in an individual patient will be most useful in guiding the treatment decision.

## References

1. Barratt A, Wyer PC, Hatala R, et al. Tips for learners of evidence-based medicine, 1: relative risk reduction, absolute risk reduction and number needed to treat. *CMAJ*. 2004;171(4:online-1 to online-8):353-358. <http://www.cmaj.ca/cgi/data/171/4/353/DC1/1>. Accessed December 20, 2013.
2. Stiegmann GV, Goff JS, Michaletz-Onody PA, et al. Endoscopic sclerotherapy as compared with endoscopic ligation for bleeding esophageal varices. *N Engl J Med*. 1992;326(23):1527-1532.
3. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Stat Med*. 2002;21(11):1575-1600.
4. Schmid CH, Lau J, McIntosh MW, Cappelleri JC. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. *Stat Med*. 1998;17(17):1923-1942.
5. Furukawa TA, Guyatt GH, Griffith LE. Can we individualize the 'number needed to treat'? an empirical study of summary effect measures in meta-analyses. *Int J Epidemiol*. 2002;31(1):72-76.
6. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med*. 1988;318(26):1728-1733.
7. Barratt AL, Wyer PC, Guyatt G, et al. NNT for studies with long-term follow-up. *CMAJ*. 2005;172(5):613-615.
8. Coldman AJ, Elwood JM. Examining survival data. *CMAJ*. 1979;121(8):1065-1068, 1071.
9. Kleinbaum DG, Klein M. *Survival Analysis: A Self-Learning Text*. New York, NY: Springer; 2012.
10. Forrow L, Taylor WC, Arnold RM. Absolutely relative: how research results are summarized can affect treatment decisions. *Am J Med*. 1992;92(2):121-124.

11. Naylor CD, Chen E, Strauss B. Measured enthusiasm: does the method of reporting trial results alter perceptions of therapeutic effectiveness? *Ann Intern Med.* 1992;117(11):916-921.
12. Hux JE, Levinton CM, Naylor CD. Prescribing propensity: influence of life-expectancy gains and drug costs. *J Gen Intern Med.* 1994;9(4):195-201.
13. Redelmeier DA, Tversky A. Discrepancy between medical decisions for individual patients and for groups. *N Engl J Med.* 1990;322(16):1162-1164.
14. Bobbio M, Demichelis B, Giustetto G. Completeness of reporting trial results: effect on physicians' willingness to prescribe. *Lancet.* 1994;343(8907):1209-1211.
15. Malenka DJ, Baron JA, Johansen S, Wahrenberger JW, Ross JM. The framing effect of relative and absolute risk. *J Gen Intern Med.* 1993;8(10):543-548.
16. McNeil BJ, Pauker SG, Sox HC Jr, Tversky A. On the elicitation of preferences for alternative therapies. *N Engl J Med.* 1982;306(21):1259-1262.



*This page intentionally left blank*

9

# Confidence Intervals: Was the Single Study or Meta-analysis Large Enough?

Gordon Guyatt, Stephen D. Walter,  
Deborah J. Cook, and Roman Jaeschke

## IN THIS CHAPTER

---

**How Should We Treat Patients With Heart Failure?  
A Problem in Interpreting Study Results**

**Solving the Problem: What Are Confidence Intervals?**

**Using Confidence Intervals to Interpret the Results  
of Clinical Trials**

**Negative Trials Often Fail to Exclude an Important  
Benefit**

**Was the Individual Trial or Meta-analysis Large Enough?  
Just Check the Confidence Intervals**

**Conclusion**

---

In discussions of whether trials were large enough, you may have heard people refer to the power of the trial as the authors presented in their sample size calculations. Such discussions are complex and confusing. As we illustrate in this chapter, whether a trial or *meta-analysis* is large enough depends only on the *confidence interval* (CI).

Hypothesis testing, on which sample size calculations are typically based, involves estimating the probability that observed results would have occurred by chance if a *null hypothesis*, which states that there is no difference between a treatment condition and a control condition, were true. Health researchers and medical educators have increasingly recognized the limitations of hypothesis testing<sup>1-5</sup>; consequently, an alternative approach, estimation, is becoming more popular.

## HOW SHOULD WE TREAT PATIENTS WITH HEART FAILURE? A PROBLEM IN INTERPRETING STUDY RESULTS

---

In a *blinded randomized clinical trial* of 804 men with heart failure, investigators compared treatment with enalapril (an angiotensin-converting enzyme [ACE] inhibitor) to treatment with a combination of hydralazine and nitrates.<sup>6</sup> In the *follow-up* period, which ranged from 6 months to 5.7 years, 132 of 403 patients (33%) assigned to receive enalapril died, as did 153 of 401 patients (38%) assigned to receive hydralazine and nitrates. The *P* value associated with the difference in mortality is .11.

Looking at this study as an exercise in hypothesis testing and adopting the usual 5% risk of obtaining a *false-positive* result, we would conclude that chance

remains a plausible explanation for the apparent differences between groups. We would classify this as a *negative study* (ie, we would conclude that no important difference existed between the treatment and *control groups*).

The investigators also conducted an additional analysis that compared the time pattern of the deaths occurring in both groups. This *survival analysis*, which generally is more sensitive than the test of the difference in proportions (see Chapter 8, Does Treatment Lower Risk? Understanding the Results), had a nonsignificant  $P$  value of .08, a result that leads to the same conclusion as the simpler analysis that focused on relative proportions at the end of the study. The authors also tell us that the  $P$  value associated with differences in mortality at 2 years (a point predetermined to be a major *end point* of the trial) was significant at .016.

At this point, one might excuse clinicians who feel a little confused. Ask yourself, is this a *positive trial*, dictating use of an ACE inhibitor instead of the combination of hydralazine and nitrates, or is it a negative study, showing no difference between the 2 regimens and leaving the choice of drugs open?

## SOLVING THE PROBLEM: WHAT ARE CONFIDENCE INTERVALS?

---

How can clinicians deal with the limitations of hypothesis testing and resolve the confusion? The solution involves posing 2 questions: (1) “What is the single value most likely to represent the true difference between experimental and control treatments?” and (2) “Given the observed difference between experimental and control groups, what is the plausible range of

differences within which the true difference might actually lie?" Confidence intervals provide an answer to this second question: they offer a range of values within which it is probable that the true value of a parameter (eg, a mean or a *relative risk*) lies. Before applying CIs to resolve the issue of enalapril vs hydralazine and nitrates in patients with heart failure, we illustrate the use of CIs with a thought experiment.

Imagine a series of 5 trials (of equal duration but different sample sizes) wherein investigators have experimented with treating patients with elevated low-density lipoprotein cholesterol and a previous myocardial infarction (MI) to determine whether a drug (a novel cholesterol-lowering agent) would work better than *placebo* in complementing a statin to *prevent* recurrent MI (Table 9-1). The smallest trial enrolled only 8 patients, and the largest enrolled 2000 patients.

Now imagine that all of the trials showed a *relative risk reduction* (RRR) for the treatment group of 50% (meaning that patients in the drug treatment group were 50% as likely as those in the placebo group to have a stroke). In each trial, how confident can we be about the true value of the RRR? If you were looking at the studies individually, which ones would lead you to recommend the treatment to your patients?

Most clinicians know intuitively that we can be more confident in the results of a larger vs a smaller trial. Why is this? In the absence of *bias* or *systematic error*, one can interpret the trial as providing an estimate of the true effect that would occur if all possible eligible patients had participated. When only a few patients participate, chance may lead to a best estimate of the

TABLE 9-1

**Confidence Intervals Around the Relative Risk Reduction  
for the Hypothetical Results of 5 Successively Larger Trials<sup>a</sup>**

Control Group Risk	Experimental Group Risk	RR, %	RRR, %	Calculated 95% CI Around the RRR, %
2/4	1/4	50	50	-174 to 92
10/20	5/20	50	50	-14 to 79.5
20/40	10/40	50	50	9.5-73.4
50/100	25/100	50	50	26.8-66.4
500/1000	250/1000	50	50	43.5-55.9

Abbreviations: CI, confidence interval; RR, relative risk; RRR, relative risk reduction.

<sup>a</sup>Reproduced from Montori et al,<sup>6</sup> by permission of the publisher. Copyright © 2005, Canadian Medical Association.

*treatment effect*—the *point estimate*—which is far removed from the true value. Confidence intervals provide the range within which such variation is likely to occur. The 95% CIs that we often see in biomedical publications represent the range in which it is very likely that the true effect lies. More precision (narrower CIs) results from larger sample sizes and, consequently, a larger number of events. Statisticians (and clinician-friendly statistical software) can calculate 95% CIs around any estimate of treatment effect.

To gain a better appreciation of CIs, go back to Table 9-1. Consider the first trial, in which 2 of 4 patients receiving the control intervention and 1 of 4 patients receiving the experimental intervention have a stroke. The risk in the experimental group was thus half of that in the control group, giving a relative risk (RR) of 50% and an RRR of 50%.

Would you be ready to recommend this treatment to a patient in view of the substantial RRR? Before you answer this, consider whether it is plausible that, with so few patients in the study, we could have just been lucky in our sample and the true treatment effect could really be a 50% increase in RR. In other words, is it plausible that the true *event rate* in the group that received treatment was 3 of 4 instead of 1 of 4?

Most clinicians answer yes to this question, and they are correct. Indeed, calculation of the CIs tells us that the results of the first trial are consistent with close to a tripling of the death rate in the intervention group.

The second trial, enrolling 40 patients, has results that are still consistent with treatment increasing the rate of deaths by, in relative terms, 17%. The third trial results tell us it is very likely that the treatment is beneficial, but the effect may be small (an RRR of less than 10%). Finally, a trial of 2000 patients with the same rate of events in the treatment and control groups provides confidence that the true effect is close to the 50% RRR we observed.

## USING CONFIDENCE INTERVALS TO INTERPRET THE RESULTS OF CLINICAL TRIALS

---

How do CIs help us understand the results of the trial of vasodilators in patients with heart failure?<sup>7</sup> By the end of the study, the mortality was 33% in the ACE inhibitor

arm and 38% in the hydralazine plus nitrate group, an *absolute difference* of 5% and an RR of 0.86. The 5% absolute difference and the 14% RRR represent our best single estimate of the mortality benefit from using an ACE inhibitor. The 95% CI around the RRR is  $-3.5\%$  to  $29\%$ . Note that when the CI crosses an RR of 1.0, the negative RRR represents a benefit for the comparator—in this case, an RRR of 3.5% for hydralazine.

How can we now interpret the study results? We can conclude that patients offered ACE inhibitors will quite possibly (but far from certainly) die later than patients offered hydralazine and nitrates. The magnitude of the true difference may be either trivial or large, and there remains the possibility of a marginally lower mortality with the hydralazine-nitrate regimen.

Use of the CI avoids the yes/no dichotomy of hypothesis testing. It also obviates the need to argue whether the study result should be considered positive or negative. One can conclude that, all else being equal, an ACE inhibitor is the appropriate choice for patients with heart failure, but our confidence in the estimate of effect on mortality is, at best, moderate. Thus, toxicity, expense, and *evidence* from other studies would all bear on the final treatment decision (see Chapter 17, How to Use a Patient Management Recommendation: Clinical Practice Guidelines and Decision Analyses). Because a number of large randomized trials have now shown a mortality benefit from ACE inhibitors in patients with heart failure,<sup>8</sup> one can confidently recommend this class of agents as the treatment of choice. Another study has suggested that for black patients, the hydralazine-nitrate combination offers additional mortality reduction beyond ACE inhibitors.<sup>9</sup>



## NEGATIVE TRIALS OFTEN FAIL TO EXCLUDE AN IMPORTANT BENEFIT

---

Another example of the use of CIs in interpreting study results comes from a randomized trial of low vs high positive end-expiratory pressure (PEEP) in patients with adult respiratory distress syndrome.<sup>10</sup> Of 273 patients in the low-PEEP group, 24.9% died; of 276 in the high-PEEP group, 27.5% died. The point estimate from these results is a 2.6% *absolute risk increase* in deaths in the high-PEEP group.

This trial of more than 500 patients might appear to exclude any possible benefit from high PEEP. The 95% CI on the absolute difference of 2.6% in favor of low PEEP, however, is 10.0% in favor of low PEEP to 4.7% in favor of high PEEP. Were it true that high PEEP reduces the risk of dying by almost 5%, all patients would want to receive the high-PEEP strategy. This would mean one would need to treat approximately 20 patients to prevent a premature death. One can thus conclude that the trial has not excluded a patient-important benefit and, in that sense, was not large enough. As in this example, negative studies seldom indicate that a treatment is not effective; rather, they fail to demonstrate a benefit.

## WAS THE INDIVIDUAL TRIAL OR META-ANALYSIS LARGE ENOUGH? JUST CHECK THE CONFIDENCE INTERVALS

---

The examples thus far demonstrate the limitations of individual trials that seldom enroll sufficient patients to generate satisfactorily narrow CIs. This illustrates why

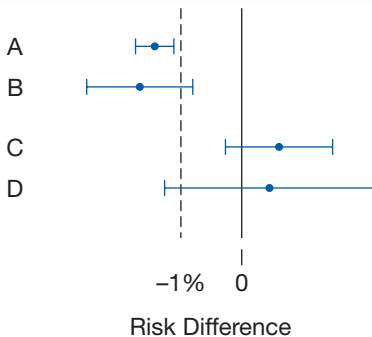
we recommend that, whenever possible, clinicians turn to *systematic reviews* and meta-analyses that pool data from multiple studies and thus achieve narrower CIs than are possible for any single study (see Chapter 4, Finding Current Best Evidence).

As implied in our discussion to this point, CIs provide a way of answering the question, “Was the meta-analysis or individual trial large enough?” In the subsequent discussion, we will focus on meta-analyses. If you are relying on an individual study, however, the principles are identical.

We illustrate the approach in [Figure 9-1](#). In this figure, we present the *pooled estimates* of 4 meta-analyses. The width of CIs from meta-analyses are driven by the number of patients rather than the number of studies. Thus, the narrower CIs (A and C) come from meta-analyses with larger numbers of events and patients, though not necessarily larger numbers of studies.

**FIGURE 9-1**

**When Is a Meta-analysis Sample Size Sufficiently Large?  
Four Hypothetical Meta-analysis Results**



Although most *forest plots* (visual plots of trial results) focus on RRs or *odds ratios*, Figure 9-1 presents the results in absolute terms. Thus, the solid vertical line in the center of the figure represents a *risk difference* (RD) (or *absolute risk reduction*) of 0: the experimental and control groups have the same mortality. Values to the left of the vertical line represent results in which treated groups had a lower mortality than the control groups. Values to the right of the vertical line represent results in which the treated group fared worse and had a higher mortality rate than the control group.

Assume that the treatment carries sufficient toxicity or risk such that, in each case, patients would choose treatment only if the RD were 1% or greater. That is, if the reduction in death rates were greater than 1%, patients would consider it worth enduring the toxic effects and risk of treatment, but if the reduction in event rates were less than 1%, they would not. The dashed line in Figure 9-1 represents this threshold reduction in death rates of 1%.

Now consider the pooled estimate from meta-analysis A: Would you recommend this therapy to your patients if the point estimate represented the truth? What if the upper boundary of the CI (representing the largest plausible effect) represented the truth? What about the lower boundary (representing the smallest plausible effect)?

For all 3 of these questions, the answer is yes, given that 1% is the smallest patient-important difference, and all suggest a benefit of greater than 1%. Thus, the meta-analysis is definitive and provides a strong inference about the treatment decision.

In the case of meta-analysis B, would your patients choose to take the treatment if either the pooled estimate or the upper boundary of the CI represented the true effect? The answer is yes, the patients would because the reduction

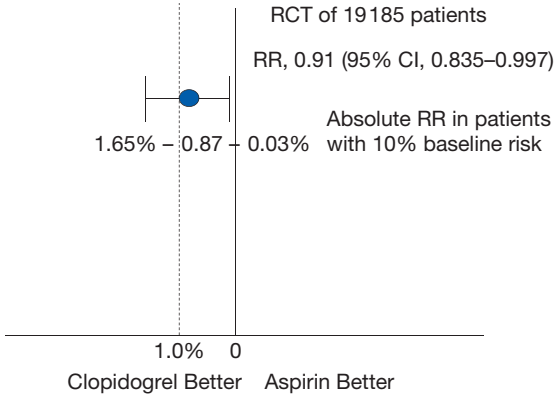
in death rate would be greater than the 1% threshold. What about the lower boundary? The answer here is no because the effect is less than the smallest difference that patients would consider large enough to undergo treatment. Although meta-analysis B reveals a positive result (ie, the CI excludes an effect of 0), the sample size was inadequate and yielded a result that remains compatible with risk reductions below the minimal patient-important difference.

For negative studies, those that fail to exclude a true treatment effect of 0, you should focus on the other end of the CI, that which represents the largest plausible treatment effect consistent with the data. You should consider whether that upper boundary of the CI falls below the smallest difference that patients might consider important. If so, the sample size is adequate, and the meta-analysis is definitive: the treatment benefit is not worth the undesirable consequences (Figure 9-1, meta-analysis D). If the boundary representing the largest plausible effect exceeds the smallest patient-important difference, then the meta-analysis is not definitive and more trials with larger sample sizes are needed (Figure 9-1, meta-analysis C).<sup>6</sup>

Application of the logic we have described can sometimes yield surprising inferences. In a blinded trial in patients with vascular disease, 19 185 patients were randomized to clopidogrel or aspirin (Figure 9-2).<sup>11</sup> Patients receiving clopidogrel experienced a 5.32% annual risk of ischemic stroke, MI, or vascular death vs 5.83% with aspirin, an RRR of 8.7% in favor of clopidogrel (95% CI, 0.3%-16.5%;  $P = .04$ ). Clopidogrel is much more expensive than aspirin. Consider patients with a risk of major vascular events of 10% in the next year (1000 per 10 000). Using the trial's point estimate of the RRR of 8.7%, such patients could expect an absolute reduction in events of 0.87% (8.7% of 10%) or 87 fewer events in 10 000 treated patients.

FIGURE 9-2

**Clopidogrel or Acetylsalicylic Acid for Threatened Vascular Events**



Abbreviations: CI, confidence interval; RCT, randomized clinical trial; RR, relative risk; RRR, relative risk reduction.

Those averse to vascular events may well choose clopidogrel, and were the upper boundary of the CI the true effect (16.5% RRR or, assuming again *baseline risk* of 1000 in 10 000, 165 fewer events in 10 000), they most likely would. If the lower boundary represented the truth, an absolute reduction of only 3 events in 10 000 patients, few, if any, would choose the more expensive drug. Given the different choices at different ends of the CI, we can conclude that the sample size—almost 20 000 patients—was insufficient to provide a definitive answer.

Our logic depends on specifying a threshold benefit below which, given the toxicity, cost, and *burden of treatment*, patients are unlikely to choose to use the intervention. Investigators seldom engage in the discussion of the threshold;

however, if you are to avoid subjecting patients to treatments with marginal benefits and substantial downsides while incorporating their *values and preferences*, you and your patients should do so.

The advent of studies designed to help determine whether we should substitute a treatment that is less expensive, easier to administer, or less toxic for an existing treatment has forced investigators to be explicit about thresholds. In such *noninferiority trials*, we will be ready to make the substitution only if we are sure that the experimental treatment is not substantially less effective than the standard treatment. We deal in detail with the logic of the noninferiority trial in Chapter 7, How to Use a Noninferiority Trial.

## CONCLUSION

---

To decide on your confidence in results, in a positive trial or meta-analysis establishing that the effect of treatment is greater than 0, look to the lower boundary of the CI to determine whether the sample size has been adequate. If this lower boundary—the smallest plausible treatment effect compatible with the data—is greater than the smallest difference that you consider important, the sample size is adequate and the trial or meta-analysis is definitive. If the lower boundary is less than this smallest important difference, the results are nondefinitive and further trials are required.

In a negative trial or meta-analysis, look to the upper boundary of the CI to determine whether the sample size has been adequate. If this upper boundary, the largest treatment effect plausibly compatible with the data, is less than the smallest difference that you consider important, the sample size is adequate and the results are definitively negative. If the upper boundary exceeds the smallest important difference, there may still be an important positive treatment effect, the trial is nondefinitive, and further trials are required.

## Acknowledgment

Portions of this material were previously published in Montori et al.<sup>6</sup>

## References

1. Simon R. Confidence intervals for reporting results of clinical trials. *Ann Intern Med.* 1986;105(3):429-435.
2. Gardner M. *Statistics With Confidence: Confidence Intervals and Statistical Guidelines.* London, England: BMJ Publishing Group; 1989.
3. Bulpitt CJ. Confidence intervals. *Lancet.* 1987;1(8531):494-497.
4. Pocock SJ, Hughes MD. Estimation issues in clinical trials and overviews. *Stat Med.* 1990;9(6):657-671.
5. Braitman LE. Confidence intervals assess both clinical significance and statistical significance. *Ann Intern Med.* 1991;114(6):515-517.
6. Montori VM, Kleinbart J, Newman TB, et al; Evidence-Based Medicine Teaching Tips Working Group. Tips for learners of evidence-based medicine, 2: measures of precision (confidence intervals). *CMAJ.* 2004;171(6):611-615.
7. Cohn JN, Johnson G, Ziesche S, et al. A comparison of enalapril with hydralazine-isosorbide dinitrate in the treatment of chronic congestive heart failure. *N Engl J Med.* 1991;325(5):303-310.
8. Garg R, Yusuf S; Collaborative Group on ACE Inhibitor Trials. Overview of randomized trials of angiotensin-converting enzyme inhibitors on mortality and morbidity in patients with heart failure. *JAMA.* 1995;273(18):1450-1456.
9. Taylor AL, Ziesche S, Yancy C, et al; African-American Heart Failure Trial Investigators. Combination of isosorbide dinitrate and hydralazine in blacks with heart failure. *N Engl J Med.* 2004;351(20):2049-2057.
10. Brower RG, Lanken PN, MacIntyre N, et al; National Heart, Lung, and Blood Institute ARDS Clinical Trials Network. Higher versus lower positive end-expiratory pressures in patients with the acute respiratory distress syndrome. *N Engl J Med.* 2004;351(4):327-336.
11. CAPRIE Steering Committee. A randomised, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE). *Lancet.* 1996;348(9038):1329-1339.

# Harm (Observational Studies)

Mitchell Levine, John P. A. Ioannidis,  
Alfred Theodore Haines, and Gordon Guyatt

## IN THIS CHAPTER

---

### Clinical Scenario

Does Soy Milk (or Soy Formula) Increase the Risk of Developing Peanut Allergy in Children?

### Finding the Evidence

#### How Serious Is the Risk of Bias?

Cohort Studies

In a Cohort Study, Aside From the Exposure of Interest, Did the Exposed and Control Groups Start and Finish With the Same Risk for the Outcome?

Case-Control Studies

In a Case-Control Study, Did the Cases and Control Group Have the Same Risk (Chance) for Being Exposed in the Past?

What Is the Risk of Bias in Cross-sectional Studies?

(continued on following page)



What Is the Risk of Bias in Case Series  
and Case Reports?

How Serious Is the Risk of Bias: Summary

### **What Are the Results?**

How Strong Is the Association Between  
Exposure and Outcome?

How Precise Is the Estimate of the Risk?

### **How Can I Apply the Results to Patient Care?**

Were the Study Patients Similar  
to the Patient in My Practice?

Was Follow-up Sufficiently Long?

Is the Exposure Similar to What Might Occur  
in My Patient?

What Is the Incremental Risk?

Are There Any Benefits That Offset the Risks  
Associated With Exposure?

### **Clinical Scenario Resolution**

---

## CLINICAL SCENARIO

### Does Soy Milk (or Soy Formula) Increase the Risk of Developing Peanut Allergy in Children?

**Y**ou are a general practitioner examining a 29-year-old patient who is 8 months pregnant with her second child. Her first child, who is now 3 years old, had an intolerance to cow's milk as an infant. He was switched to soy formula and then soy milk, which he subsequently tolerated well. At 2 years of age, cow's milk was reintroduced without any problems, and he has been receiving cow's milk since. The mother was planning to start feeding her next child soy formula at birth but heard from a neighbor that it can increase the risk of peanut allergy in her child—a potentially serious and lifelong problem. She asks for your advice on the topic. Because you are not familiar with this issue, you inform the patient that you will examine the *evidence* and discuss your findings with her when she returns for her next prenatal visit in 1 week.

## FINDING THE EVIDENCE

You formulate the relevant question: In infants, is there an association between *exposure* to soy milk and the subsequent development of peanut allergy? You search a point-of-care clinician evidence synthesis tool with the term “peanut allergy.” Under the subtopic “Causes and Risk Factors,” you see that “consumption of soy milk or soy formula” is identified as a possible *risk factor* and a reference is provided. You click on the hypertext link to view the relevant article.<sup>1</sup>

The article describes a *case-control study* that used a geographically defined *cohort* of 13971 preschool children. The investigators identified children with a convincing history of peanut allergy who reacted to a peanut challenge in which they were blind to whether they were being exposed to peanut protein or a “placebo.” They collected detailed information from the children’s parents and from 2 groups of control parents (a random sample from the geographically defined cohort and from a subgroup of children from the cohort who had eczema in the first 6 months of life and whose mothers had a history of eczema).

Box 10-1 presents our usual 3-step approach to using an article about *harm* from the medical literature to guide your practice. You will find these criteria useful for a variety of issues that involve concerns of etiology or risk factors in which a potentially harmful exposure cannot be randomly assigned. These *observational studies* involve using cohort or case-control designs.

## HOW SERIOUS IS THE RISK OF BIAS?

---

Clinicians often encounter patients who face potentially harmful exposures to either medical interventions or environmental agents. These circumstances give rise to common questions: Do cell phones increase the risk of brain tumors? Do vasectomies increase the risk of prostate cancer? Do changes in health care policies (eg, activity-based funding) lead to harmful *health outcomes*? When examining these questions, clinicians and administrators must evaluate the *risk of bias*, the strength of the association between the assumed cause and the adverse outcome, and the relevance to patients in their practice or domain.

In answering any clinical question, our first goal should be to identify whether there is an existing *systematic review* of the topic that can provide a summary of the highest-quality available evidence (see the Summarizing the Evidence section). Interpreting such a *review* requires an understanding of the rules of evidence for individual or *primary studies*, *randomized*

**BOX 10-1****Users' Guides for an Article About Harm****How serious is the risk of bias?**

In a cohort study, aside from the exposure of interest, did the exposed and control groups start and finish with the same risk for the outcome?

Were patients similar for prognostic factors that are known to be associated with the outcome (or did statistical adjustment address the imbalance)?

Were the circumstances and methods for detecting the outcome similar?

Was the follow-up sufficiently complete?

In a case-control study, did the cases and control group have the same risk for being exposed in the past?

Were cases and controls similar with respect to the indication or circumstances that would lead to exposure (or did statistical adjustment address the imbalance)?

Were the circumstances and methods for determining exposure similar for cases and controls?

**What are the results?**

How strong is the association between exposure and outcome?

How precise was the estimate of the risk?

**How can I apply the result to patient care?**

Were the study patients similar to the patient in my practice?

Was follow-up sufficiently long?

Is the exposure similar to what might occur in my patient?

What is the magnitude of the risk?

Are there any benefits that are known to be associated with exposure?

*clinical trials* (RCTs), and observational studies. The tests for judging the risk of bias associated with results of observational studies will help you decide whether exposed and *control groups* (or cases and controls) began and completed the study with sufficient similarities that we can obtain a minimally biased assessment of the influence of exposure on outcome (see Chapter 5, Why Study Results Mislead: Bias and Random Error).

Randomized clinical trials provide less biased estimates of potentially harmful effects than other study designs because randomization is the best way to ensure that groups are balanced with respect to known and unknown determinants of the outcome (see Chapter 6, Therapy [Randomized Trials]). Although investigators conduct RCTs to determine whether therapeutic agents are beneficial, they also should look for harmful effects and may sometimes make surprising discoveries about the adverse effects of the intervention on their primary outcomes.

There are 4 reasons why RCTs may not be helpful for determining whether a putative harmful agent truly has deleterious effects. First, we may consider it unethical to randomize patients to exposures that might result in harmful effects without benefit (eg, smoking).

Second, we are often concerned about rare and serious adverse effects that may become evident only after tens of thousands of patients have consumed a medication for a period of years. For instance, even a very large RCT failed to detect an association between clopidogrel and thrombotic thrombocytopenic purpura,<sup>3</sup> which appeared in a subsequent observational study.<sup>4</sup> Randomized clinical trials that address adverse effects may be feasible for adverse event rates as low as 1%,<sup>5,6</sup> but the RCTs needed to explore harmful events occurring in fewer than 1 in 100 exposed patients are logistically difficult and often prohibitively expensive because of the huge sample size and lengthy *follow-up* required. *Meta-analyses* may be helpful when the event rates are very low.<sup>7</sup> However, availability of large-scale evidence on specific harms in systematic reviews is not common.

For example, in a report of nearly 2000 systematic reviews, only 25 had large-scale data on 4000 or more randomized participants regarding well-defined harms that might be associated with the interventions under study.<sup>8</sup>

Third, RCT duration of follow-up is limited, yet not infrequently we are interested in knowing effects years, or even decades, after the exposure (eg, long-term consequences of chemotherapy in childhood).<sup>9</sup>

Fourth, even when events are sufficiently frequent and occur during a time frame feasible for RCTs to address, study reports often fail to adequately provide information on harm.<sup>10</sup>

Given that clinicians will not find RCTs to answer most questions about harm, they must understand the alternative strategies used to minimize *bias*. This requires a familiarity with observational study designs (Table 10-1).

There are 2 main types of observational studies: cohort and case-control. In a cohort study, investigators identify exposed and nonexposed groups of patients, each a cohort, and then follow them forward in time, monitoring the occurrence of outcomes of interest in an attempt to identify whether there is an association between the exposure and the outcomes. The cohort design is similar to an RCT but without randomization; rather, the determination of whether a patient received the exposure of interest results from the patient's or investigator's preference or from happenstance.

Case-control studies also assess associations between exposures and outcomes. Rare outcomes or those that take a long time to develop can threaten the feasibility not only of RCTs but also of cohort studies. The case-control study provides an alternative design that relies on the initial identification of cases—that is, patients who have already developed the target outcome—and the selection of controls—persons who do not have the outcome of interest. Using case-control designs, investigators assess the relative frequency of previous exposure to the putative harmful agent in the cases and the controls.

**TABLE 10-1****Directions of Inquiry and Key Methodologic Strengths and Weaknesses for Different Study Designs**

<b>Design</b>	<b>Starting Point</b>	<b>Assessment</b>	<b>Strengths</b>	<b>Weaknesses</b>
Randomized clinical trial	Exposure status	Outcome event status	Low susceptibility to bias	Feasibility and generalizability constraints
Cohort	Exposure status	Outcome event status	Feasible when randomization of exposure not possible, generalizability	Susceptible to bias
Case-control	Outcome event status	Exposure status	Overcomes temporal delays and the need for huge sample sizes to accumulate rare events	Susceptible to bias

For example, in addressing the impact of nonsteroidal anti-inflammatory drugs (NSAIDs) on clinically apparent upper gastrointestinal tract hemorrhage, investigators needed a cohort study to deal with the problem of infrequent events. Bleeding among those taking NSAIDs has been reported to occur approximately 1.5 times per 1000 person-years of exposure, in comparison with 1.0 per 1000 person-years in those not taking NSAIDs.<sup>11</sup> Because the event rate in unexposed patients is so low (0.1%), an RCT to study an increase in risk of 50% would require huge numbers of patients (sample size calculations suggested approximately 75 000 patients per group) for adequate power to test the hypothesis that NSAIDs cause the additional bleeding.<sup>12</sup> Such an RCT would not be feasible, but a cohort study, in which the information comes from a large administrative database, would be possible.

## Cohort Studies

Cohort studies may be prospective or retrospective. In prospective cohort studies, the investigator enrolls patients or participants, starts the follow-up, and waits for the outcomes (events of interest) to occur. Such studies may take many years to complete, and thus they are difficult to conduct. An advantage, however, is that the investigators can plan how to monitor patients and collect data.

In retrospective cohort studies, the data regarding both exposures and outcomes have been previously collected; the investigator obtains the data and determines whether participants with and without the outcome of interest have been exposed to the putative causal agent or agents. These studies are easier to perform because they depend on the availability of data



on exposures and outcomes that have already happened. On the other hand, the investigator has less control over the quality and relevance of the available data. In the end, clinicians need not pay too much attention to whether studies are prospective or retrospective but should instead focus on the risk of bias criteria in Box 10-1.

### **In a Cohort Study, Aside From the Exposure of Interest, Did the Exposed and Control Groups Start and Finish With the Same Risk for the Outcome?**

#### **Were Patients Similar for Prognostic Factors That Are Known to Be Associated With the Outcome (or Did Statistical Adjustment Level Address This Imbalance)?**

Cohort studies will yield biased results if the group exposed to the putative harmful agent and the unexposed group begin with additional differences in baseline characteristics that give them a different prognosis (ie, a different risk of the target outcome) and if the analysis fails to deal with this imbalance. For instance, in the association between NSAIDs and the increased risk of upper gastrointestinal tract bleeding, age may be associated with exposure to NSAIDs and gastrointestinal bleeding. In other words, because patients taking NSAIDs will be older and because older patients are more likely to bleed, this variable makes attribution of an increased risk of bleeding to NSAID exposure problematic. When a variable with prognostic power differs in frequency in the exposed and unexposed cohorts, we refer to the situation as *confounding*.

There is no reason that patients who self-select (or who are selected by their physicians) for exposure to a potentially harmful agent should be similar to the nonexposed patients with respect to important determinants of the harmful outcome. Indeed, there are many reasons to expect they will not be similar. Physicians are appropriately reluctant to prescribe medications they perceive will put their patients at risk.

In one study, 24.1% of patients who were given a then-new NSAID, ketoprofen, had received peptic ulcer therapy during the previous 2 years compared with 15.7% of the control population.<sup>13</sup> The likely reason is that the ketoprofen manufacturer succeeded in persuading clinicians that ketoprofen was less likely to cause gastrointestinal bleeding than other agents. A comparison of ketoprofen to other agents would be subject to the risk of finding a spurious increase in bleeding with the new agent (compared with other therapies) because higher-risk patients would have been receiving the ketoprofen. This bias may be referred to as a selection bias or a bias due to confounding by indication.

The prescription of benzodiazepines to elderly patients provides another example of the way that selective physician prescribing practices can lead to a different distribution of risk in patients receiving particular medications, sometimes referred to as the *channeling bias*.<sup>14</sup> Ray et al<sup>15</sup> found an association between long-acting benzodiazepines and risk of falls (*relative risk* [RR], 2.0; 95% *confidence interval* [CI], 1.6-2.5) in data from 1977 to 1979 but not in data from 1984 to 1985 (RR, 1.3; 95% CI, 0.9-1.8). The most plausible explanation for the change is that patients at high risk for falls (those with dementia) selectively received these benzodiazepines during the earlier period. Reports of associations between benzodiazepine use and falls led to greater caution, and the apparent association disappeared when physicians began to avoid using benzodiazepines in those at high risk of falling.

Therefore, investigators must document the characteristics of the exposed and nonexposed participants and either demonstrate their comparability (very unusual in cohort studies) or use statistical techniques to adjust for these differences. Effective adjusted analyses for prognostic factors require the

accurate measurement of those prognostic factors. For prospective cohorts, the investigators may take particular care of the quality of this information. For retrospective databases, however, one has to make use of what is available. Large administrative databases, although providing a sample size that may allow ascertainment of rare events, often have limited quality of data concerning relevant patient characteristics, health care encounters, or diagnoses. For example, in a cross-sectional study designed to measure the accuracy of electronic reporting of care practices compared with manual review, electronic reporting significantly underestimated rates of appropriate asthma medication and pneumococcal vaccination and overestimated rates of cholesterol control in patients with diabetes.<sup>16</sup>

Even if investigators document the comparability of potentially confounding variables in exposed and nonexposed cohorts, and even if they use statistical techniques to adjust for differences, important prognostic factors that the investigators do not know about or have not measured may be unbalanced between the groups and thus may be responsible for differences in outcome. We call this *residual confounding*.

Returning to our earlier example, it may be that the illnesses that require NSAIDs, rather than the NSAIDs themselves, contribute to the increased risk of bleeding. Thus, the strength of inference from a cohort study will always be less than that of a rigorously conducted RCT.

### **Were the Circumstances and Methods for Detecting the Outcome Similar?**

In cohort studies, ascertainment of outcome is the key issue. For example, investigators have reported a 3-fold increase in the risk of malignant melanoma in individuals who work with radioactive materials. One possible explanation for some of the increased risk might be that physicians, concerned about a

possible risk, search more diligently and therefore detect disease that might otherwise go unnoticed (or they may detect disease at an earlier point). This could result in the exposed cohort having an apparent, but spurious, increase in risk—a situation known as *surveillance bias*.<sup>18</sup>

The choice of outcome may partially address this problem. In one cohort study, for example, investigators assessed perinatal outcomes among infants of men exposed to lead and organic solvents in the printing industry by means of a cohort study that assessed all of the men who had been members of the printers' unions in Oslo, Norway.<sup>19</sup> The investigators used job classification to categorize the fathers as being exposed to lead and organic solvents or not exposed to those substances. Investigators' awareness of whether the fathers had been exposed to the lead or solvents might bias their assessment of the infant's outcome for minor birth defects or defects that required special investigative procedures. On the other hand, an outcome such as preterm birth would be unlikely to increase simply as a result of *detection bias* (the tendency to look more carefully for an outcome in one of the comparison groups) because prior knowledge of exposure is unlikely to influence whether an infant is considered preterm or not. The study found that exposure was associated with an 8-fold increase in preterm births but no increase in birth defects, so detection bias was not an issue for the results that were obtained in this study.

### Was the Follow-up Sufficiently Complete?

As we pointed out in Chapter 6, Therapy (Randomized Trials), loss to follow-up can introduce bias because the patients who are lost may have different outcomes from those patients still available for assessment. This is particularly problematic if there are differences in follow-up between the exposed and nonexposed groups.

For example, in a well-executed study,<sup>20</sup> investigators determined the vital status of 1235 of 1261 white men (98%) employed in a chrysotile asbestos textile operation between 1940 and 1975. The RR for lung cancer death over time increased from 1.4 to 18.2 in direct proportion to the cumulative exposure among asbestos workers with at least 15 years since first exposure. In this study, the 2% missing data were unlikely to affect the results, and the loss to follow-up did not threaten the strength of the inference that asbestos exposure caused lung cancer deaths.

## Case-Control Studies

Case-control studies are always retrospective in design. The outcomes (events of interest) have already happened and participants are designated to 1 of 2 groups: those with the outcomes (cases) and those where the outcome is absent (controls). Retrospectively, investigators ascertain prior exposure to putative causal agents. This design entails inherent risks of bias because exposure data require memory and recall or are based on a collection of data that were originally accumulated for purposes other than the intended study.

### **In a Case-Control Study, Did the Cases and Control Group Have the Same Risk (Chance) for Being Exposed in the Past?**

#### **Were Cases and Controls Similar With Respect to the Indication or Circumstances That Would Lead to Exposure (or Did Matching or Statistical Adjustment Address the Imbalance)?**

As with cohort studies, case-control studies are susceptible to unmeasured confounding. For instance, in looking at the association between use of  $\beta$ -agonists and mortality among patients with asthma, investigators need to consider—and

match or adjust for—previous hospitalization and use of other medications to avoid confounding by disease severity. Patients who use more  $\beta$ -agonists may have more severe asthma, and this severity, rather than  $\beta$ -agonist use, may be responsible for increased mortality. As in cohort studies, however, matching and adjustment cannot eliminate the risk of bias, particularly when exposure varies over time. In other words, matching or adjustment for hospitalization or use of other medications may not adequately capture all of the variability in underlying disease severity in asthma. In addition, the adverse lifestyle behaviors of patients with asthma who use large amounts of  $\beta$ -agonists could be the real explanation for the association.

To further illustrate the concern about unmeasured confounding, consider the example of a case-control study that was designed to assess the association between diethylstilbestrol ingestion by pregnant women and the development of vaginal adenocarcinomas in their daughters many years later.<sup>21</sup> An RCT or prospective cohort study designed to test this cause-and-effect relationship would have required at least 20 years from the time when the association was first suspected until the completion of the study. Furthermore, given the infrequency of the disease, an RCT or a cohort study would have required hundreds of thousands of participants. By contrast, using the case-control strategy, the investigators delineated 2 relatively small groups of young women. Those who had the outcome of interest (vaginal adenocarcinoma) were designated as the cases ( $n = 8$ ), and those who did not experience the outcome were designated as the controls ( $n = 32$ ). Then, working backward in time, the investigators determined exposure rates to diethylstilbestrol for the 2 groups. They found a significant association between in

utero diethylstilbestrol exposure and vaginal adenocarcinoma, and they found their answer without a delay of 20 years and by studying only 40 women.

An important consideration in this study would be whether the cases could have been exposed to diethylstilbestrol in any special circumstances that would not have affected women in the control group. In this situation, diethylstilbestrol had been prescribed to women at risk for miscarriages or premature births. Could either of these indications be a confounder? Before the introduction of diethylstilbestrol, vaginal adenocarcinoma in young women was uncommon, but miscarriages and premature birth were common. Thus, it would be unlikely that miscarriages and premature births were directly associated with vaginal adenocarcinoma, and in the absence of such an association, neither could be a confounder.

In another study, investigators used a case-control design relying on computer-record linkages between health insurance data and a drug insurance plan to investigate the possible association between use of  $\beta$ -adrenergic agonists and mortality rates in patients with asthma.<sup>22</sup> The database for the study included 95% of the population of the province of Saskatchewan, Canada. The investigators selected 129 patients who had experienced a fatal or near-fatal asthma attack to serve as cases and used a matching process to select another 655 patients who also had asthma but who had not had a fatal or near-fatal asthma attack to serve as controls.

The tendency of patients with more severe asthma to use more  $\beta$ -adrenergic medications could create a spurious association between drug use and mortality rate. The investigators attempted to control for the confounding effect of disease severity by measuring the number of hospitalizations in the 24 months before death (for the

cases) or before the index date of entry into the study (for the control group) and by using an index of the aggregate use of medications. They found an association between the routine use of large doses of  $\beta$ -adrenergic agonists through metered-dose inhalers and death from asthma (*odds ratio* [OR], 2.6 per canister of inhaler per month; 95% CI, 1.7-3.9), even after correcting for measures of disease severity.

### Were the Circumstances and Methods for Determining Exposure Similar for Cases and Controls?

In case-control studies, ascertainment of the exposure is a key issue. However, if case patients have a better memory for exposure than control patients, the result will be a spurious association.

For example, a case-control study found a 2-fold increase in risk of hip fracture associated with psychotropic drug use.<sup>23</sup> In this study, investigators established drug exposure by examining computerized claim files from the Michigan Medicaid program, a strategy that avoided selective memory of exposure—*recall bias*—and differential probing of cases and controls by an interviewer—*interviewer bias*.

Another example was a case-control study that evaluated whether the use of cell phones was associated with an increased risk of motor vehicle crash.<sup>24</sup> Suppose the investigators had tried to ask people who had a motor vehicle crash and control patients (who were in no crash at the same day and time) whether they were using their cell phone around the time of interest. People who were



in a crash would have been more likely to recall such use because their memory might be heightened by the unfortunate circumstances. This would have led to a spurious association because of differential recall. Alternatively, they might specifically deny the use of a cell phone because of embarrassment or legal concerns, thus obscuring an association. Therefore, the investigators in this study used a computerized database of cell phone use instead of patient recall.<sup>24</sup> Moreover, the investigators used each person in a crash as his or her own control. The time of the crash was matched against corresponding times of the life of the same person when they were driving but when no crash occurred (eg, same time driving to work). This appropriate design established that use of cell phones was associated with an increased risk of having a motor vehicle crash.

Not all studies have access to unbiased information on exposure. For instance, in a case-control study of the association between coffee and pancreatic cancer, the patients with cancer may be more motivated to identify possible explanations for their problem and provide a greater recounting of coffee use.<sup>25</sup> Also, if the interviewers are not blinded to whether a patient is a case or a control patient, the interviewer may probe deeper for exposure information from cases. In this particular study, there were no objective sources of data regarding exposure. Recall or interviewer bias might have explained the apparent association.

As it happened, another bias provided an even more likely explanation for what turned out to be a spurious association. The investigators chose control patients from the practices of the physicians treating the patients with pancreatic cancer. These control patients had a variety of gastrointestinal problems, some of which were exacerbated by coffee

ingestion. The control patients had learned to avoid coffee, which explains the investigators' finding of an association between coffee (which the patients with pancreatic cancer consumed at general population levels) and pancreatic cancer. Subsequent investigations, using more appropriate controls, refuted this association.<sup>26</sup>

In addition to a biased assessment of exposure, *random error* in exposure ascertainment is also possible. In random misclassification, exposed and unexposed patients are misclassified, but the rates of misclassification are similar in cases and controls. Such nondifferential misclassification dilutes any association (ie, the true association will be larger than the observed association). Fortunately, unless the misclassification is extremely large, the reduction in the true association will not be important.

### What Is the Risk of Bias in Cross-sectional Studies?

Like the cohort and the case-control study, the *cross-sectional study* is also an observational study design. Like a cohort study, a cross-sectional study is based on an assembled population of exposed and unexposed participants. However, in the cross-sectional study, the exposure and the existing or prevalent outcome are measured at the same point in time. Accordingly, the direction of association may be difficult to determine. Another important limitation is that the outcome or the threat of experiencing an adverse outcome may have led patients assigned as cases to leave the study, so a measure of association may be biased against the association. However, cross-sectional studies are relatively inexpensive and quick to conduct and may be useful in generating and exploring hypotheses that will be subsequently investigated using other observational designs or RCTs.

## What Is the Risk of Bias in Case Series and Case Reports?

*Case series* (descriptions of a series of patients) and case reports (descriptions of individual patients) do not provide any comparison groups, so it is impossible to determine whether the observed outcome would likely have occurred in the absence of the exposure. Although descriptive studies have been reputed to have significant findings mandating an immediate change in clinician behavior, this is rarely justified, and without availability of evidence from stronger study designs, there are potentially undesirable consequences when actions are taken in response to evidence warranting very low confidence. Recall the consequences of case reports of specific birth defects occurring in association with thalidomide exposure.<sup>27</sup>

Consider the case of the drug Bendectin (a combination of doxylamine, pyridoxine, and dicyclomine used as an antiemetic in pregnancy), whose manufacturer withdrew it from the market as a consequence of case reports suggesting that it was teratogenic.<sup>28</sup> Later, although a number of comparative studies reported the drug's relative safety,<sup>29</sup> they could not eradicate the prevailing litigious atmosphere—which prevented the manufacturer from reintroducing Bendectin. Thus, many pregnant women who might have benefited from the drug's availability were denied the symptomatic relief it could have offered.

For some interventions, registries of adverse events may provide the best possible initial evidence. For example, there are vaccine registries that record adverse events among people who have received the vaccine. These registries may signal problems with a particular adverse event that would be very difficult to capture from prospective studies limited by sample sizes that

were too small. Even retrospective studies might be too difficult to conduct if people who receive the vaccine differ substantially from those who do not and if adjustment or matching cannot deal with the differences. In this situation, investigators might conduct a *before-after study* using the general population before the introduction of the new vaccine occurred. Such comparisons using historical controls are, however, prone to bias because many other factors may have changed in the same period. If, however, changes in the incidence of an adverse event are very large, the signal may be real. An example is the clustering of intussusception cases among children receiving a particular type of rotavirus vaccine,<sup>30</sup> resulting in a decision to withdraw the vaccine. The association was subsequently supported by a case-control study.<sup>31</sup> Eventually, another type of rotavirus vaccine was developed that did not cause this adverse event.

In general, clinicians should not draw conclusions about relationships from case series, but rather, they should recognize that the results may generate questions, or even hypotheses, that clinical investigators can address with studies that have optimal safeguards against risk of bias. When the immediate risk of exposure outweighs the benefits (and outweighs the risk of stopping an exposure), the clinician may have to make a management decision with less than optimal data.

### How Serious Is the Risk of Bias: Summary

Just as it is true for the resolution of questions of therapeutic effectiveness, clinicians should first look to RCTs to resolve issues of harm. They will often be disappointed in the search and must make use of studies of weaker design. Regardless of the design, however, they should look for an appropriate control population. For cohort studies, the control group should have a similar baseline risk of outcome, or investigators should have used statistical techniques to adjust for differences. In case-control studies, the cases and the controls should have had a similar opportunity to have been exposed, so that if a difference

in exposure is observed, one might legitimately conclude that the association could be due to a link between the exposure and the outcome and not due to a confounding factor. Nevertheless, investigators should routinely use statistical techniques to match cases and controls or adjust for differences.

Even when investigators have taken all of the appropriate steps to minimize bias, clinicians should bear in mind that residual differences between groups still may bias the results of observational studies.<sup>32</sup> Because evidence, clinician preferences, and patient *values and preferences* determine the use of interventions in the real world, exposed and unexposed patients are likely to differ in prognostic factors.

## USING THE GUIDE

**R**eturning to our earlier discussion, the study that we retrieved investigating the association between soy milk (or soy formula) and the development of peanut allergy used a case-control design.<sup>1</sup> Those with peanut allergy (cases) appeared to be similar to controls with respect to the indication or circumstances leading to soy exposure, but there were a few potentially important imbalances. In the peanut allergy group, a family history of peanut allergy and an older sibling with a history of milk intolerance were more common and could have biased the likelihood of a subsequent child's being exposed to soy. To avoid confounding, the investigators conducted an *adjusted analysis*.

The methods for determining exposure were similar for cases and controls because the data were collected with the interviewers and parents unaware of the hypothesis that related soy exposure to peanut allergy (thus avoiding interviewer bias and perhaps recall bias). With regard to access to soy, all of the children came from the same geographic region, although this does not ensure that cultural and economic factors that might determine soy access were similar in cases and controls. Overall, protection from risk of bias seemed adequate.

## WHAT ARE THE RESULTS?

---

### How Strong Is the Association Between Exposure and Outcome?

We describe options that can be used for expressing an association between an exposure and an outcome—the *risk ratio* or RR and the OR—in other chapters of this book (see Chapter 8, Does Treatment Lower Risk? Understanding the Results).

For example, in a cohort study that assessed in-hospital mortality after noncardiac surgery in male veterans, 23 of 289 patients with a history of hypertension died compared with 3 of 185 patients without the condition.<sup>34</sup> The RR for mortality in hypertensive patients compared with normotensive patients (23/289 and 3/185, respectively) was 4.9 (95% CI, 1.5-16.1). The RR tells us that death after noncardiac surgery occurs almost 5 times more often in patients with hypertension than in normotensive patients.

The estimate of the RR depends on the availability of samples of exposed and unexposed patients, where the proportion of the patients with the outcome of interest can be determined. The RR is not applicable to case-control studies in which the number of cases and controls—and, therefore, the proportion of individuals with the outcome—is chosen by the investigator. For case-control studies, instead of using a ratio of risks, the RR, we use a ratio of odds, the OR, specifically the odds of a case patient being exposed divided by the odds of a control patient being exposed. Unless the risk of the outcome in the relevant population is high (20% or more), you can think of the OR as providing a good estimate of the much easier to conceptualize RR.

## How Precise Is the Estimate of the Risk?

Clinicians can evaluate the precision of the estimate of risk by examining the CI around that estimate (see Chapter 9, Confidence Intervals: Was the Single Study or Meta-analysis Large Enough?). In a study in which investigators have found an association between an exposure and an adverse outcome, the lower limit of the estimate of RR associated with the adverse exposure provides an estimate of the lowest possible magnitude of the association. Alternatively, in a negative study (in which the results are not *statistically significant*), the upper boundary of the CI around the RR tells the clinician just how big an adverse effect still may be present, despite the failure to find a statistically significant association.

### USING THE GUIDE

The investigators calculated an OR of 2.6 (95% CI, 1.3-5.2) for the risk of peanut allergy in those exposed to soy vs those not exposed.<sup>1</sup> These results were adjusted for skin manifestations of allergy (ie, atopy). The consumption of soy by the infants was independently associated with peanut allergy and could not be explained as a dietary response to other atopic conditions. It nevertheless remains possible that the association with soy was confounded by other, unknown factors. Unfortunately, the investigators did not evaluate the possibility of a dose-response relationship for soy exposure and the development of peanut allergy.

## HOW CAN I APPLY THE RESULTS TO PATIENT CARE?

### Were the Study Patients Similar to the Patient in My Practice?

If possible biases in a study are not sufficient to dismiss the study out of hand, you should consider the extent to which the results

might apply to the patient in your practice. Would your patient have met the eligibility criteria? Is your patient similar to those described in the study with respect to potentially important factors, such as patient characteristics or medical history? If not, is the biology of the harmful exposure likely to be different for the patient for whom you are providing care?

### **Was Follow-up Sufficiently Long?**

Studies can be pristine in terms of avoiding bias but of limited use if patients are not followed up for a sufficiently long period. That is, they may provide an unbiased estimate of the effect of an exposure during the short term, but the time frame in which we are interested is a substantially longer period. For example, most cancers take a decade or longer to develop from the original assault at the biologic level to the clinically detected malignant tumor. If the question is whether a specific exposure, say to an industrial chemical, is related to a subsequent cancer, one would not expect cancers detected in the first few years to reflect any of the effect of the exposure under question.

### **Is the Exposure Similar to What Might Occur in My Patient?**

Clinicians should ask whether there are important differences in the exposures of a study (eg, dose and duration) between their patient and the patients in the study. For example, it should be clear that the risk of thrombophlebitis associated with oral contraceptive use described in the 1970s may not be applicable to the patient in the 21st century because of the lower estrogen dose in oral contraceptives currently used. Another example of questionable applicability comes from a study that found that workers employed in a chrysotile asbestos textile operation between 1940 and 1975 had an increased risk of lung cancer death, a risk that increased from 1.4 to 18.2 in direct relation to cumulative exposure among asbestos workers with at least 15 years since first exposure.<sup>18</sup> The study does not provide



trustworthy information regarding what might be the risks associated with only brief or intermittent exposure to asbestos (eg, a person working for a few months in an office located in a building subsequently found to have abnormally high asbestos levels).

### What Is the Incremental Risk?

The RR and OR do not tell us how frequently the problem occurs; they tell us only that the observed effect occurs more or less often in the exposed group vs the unexposed group. Even when we observe a large and statistically significant relative difference between the 2 groups, the results still may not be important if the adverse event is rare. Thus, we need a method for assessing the absolute impact of the exposure. In our discussion of therapy (see Chapter 6, Therapy [Randomized Trials], and Chapter 8, Does Treatment Lower Risk? Understanding the Results), we describe how to calculate the *risk difference* and the number of patients whom clinicians must treat to prevent an adverse event (*number needed to treat*). When the issue is harm, we can use data from a randomized trial or cohort study (but not a case-control study) in a similar way to calculate the number of patients who would have to be exposed to result in an additional harmful event. However, this calculation requires knowledge of the *absolute risk* in unexposed individuals in our population.

For example, during a mean of 10 months of follow-up, investigators conducting the Cardiac Arrhythmia Suppression Trial, an RCT of antiarrhythmic agents,<sup>35</sup> found that the mortality rate was 3.0% for placebo-treated patients and 7.7% for those treated with either encainide or flecainide. The absolute risk increase was 4.7%, the reciprocal of which ( $100/4.7$ ) tells us that, on average, for

every 21 patients treated with encainide or flecainide for approximately a year, there would be 1 excess death.

This contrasts with our example of the association between NSAIDs and upper gastrointestinal tract bleeding. Of 2000 unexposed patients in that study, 2 will have a bleeding episode each year. Of 2000 patients taking NSAIDs, 3 will have such an episode each year. Thus, if we treat 2000 patients with NSAIDs, we can expect a single additional bleeding event.<sup>11</sup>

### **Are There Any Benefits That Offset the Risks Associated With Exposure?**

Even after evaluating the evidence that an exposure is harmful and establishing that the results are potentially applicable to the patient in your practice, determining subsequent actions may not be simple. In addition to considering the magnitude of the risk, one must consider the adverse consequences of reducing or eliminating exposure to the harmful agent (ie, the magnitude of any potential benefit that patients will no longer receive).

Clinical decision making is simple when harmful consequences are unacceptable and benefit is absent. For example, because the evidence of increased mortality from encainide and flecainide came from an RCT with low risk of bias,<sup>35</sup> we can be at least moderately confident of a substantial increase in risk of death. Because treating only 21 people would result in an excess death, it is no wonder that clinicians quickly curtailed their use of these antiarrhythmic agents when the study results became available.

The clinical decision is also made easier when an acceptable alternative for avoiding the risk is available. Even if the evidence warrants low confidence, the availability of an alternative substance can result in a clear decision.

## CLINICAL SCENARIO RESOLUTION

**Y**ou determine that the patient's unborn child, once he or she reaches early childhood, would likely fulfill the eligibility criteria in the study. Also relevant to the clinical scenario, but perhaps unknown, is whether the soy products discussed in the study are similar to the ones that the patient is considering using. With regard to the magnitude of risk, the prevalence of peanut allergy is approximately 4 per 1000 children. An approximate calculation would suggest that, if the OR with exposure to soy is truly 2.6, 10 children per 1000 would be affected by peanut allergy, an additional 6 children in every 1000. In other words, the number of children needed to be exposed to soy to result in 1 additional child having peanut allergy is 167 ( $1000/6$ ). Finally, there are no data regarding the negative consequences of withholding soy formula or soy milk products, and the use of these products would clearly be dependent on how severe and sustained an intolerance to cow's milk was in a particular child.

To decide on your course of action, you proceed through 3 steps of using the medical literature to guide your clinical practice. First, you consider the risk of bias in the study before you. Adjustments of known confounders did not diminish the association between soy exposure as a neonate and the development of peanut allergy. Also, the design of the study provides adequate safeguards against recall or interviewer bias. You conclude that, with the obvious limitations of the observational design (generally only warranting low confidence in estimates of effect), the study is at low risk of bias.

Turning to the results, you note a moderate association between soy exposure and the development of peanut allergy (moderate typically being considered ORs greater than 2 and less than 5) that is strong enough, despite the limitations of the observational design, that it leaves you moderately confident of an association between exposure to soy and peanut allergy.

The lower boundary of the CI (1.3) and the uncertainty around the baseline risk estimate of 4 per 10000 children lead you to conclude that you have only low confidence in your estimate of the incremental harm of peanut allergy of 6 in 1000.

You proceed to the third step, and consider the implications of the study results for your patient. The study would appear to apply to a future child of your patient. Although the best estimate of the absolute increase in risk is only 6 in 1000, and you have only low confidence in this estimate, the consequences of peanut allergy can be a serious health threat to a patient and quite disruptive for a family because of the required precautions and food restrictions. You discuss the situation with the mother, who elects to start feedings with milk products. Together you agree that, given the limited confidence in estimates and the small absolute risk, should the child appear to have distressing milk allergy, she will probably switch to soy.

## References

1. Lack G, Fox D, Northstone K, Golding J; Avon Longitudinal Study of Parents and Children Study Team. Factors associated with the development of peanut allergy in childhood. *N Engl J Med.* 2003;348(11):977-985.
2. CAPRIE Steering Committee. A randomised, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE). *Lancet.* 1996;348(9038):1329-1339.
3. Bennett CL, Connors JM, Carwile JM, et al. Thrombotic thrombocytopenic purpura associated with clopidogrel. *N Engl J Med.* 2000;342(24):1773-1777.
4. Silverstein FE, Graham DY, Senior JR, et al. Misoprostol reduces serious gastrointestinal complications in patients with rheumatoid arthritis receiving nonsteroidal anti-inflammatory drugs. A randomized, double-blind, placebo-controlled trial. *Ann Intern Med.* 1995;123(4):241-249.
5. Bombardier C, Laine L, Reicin A, et al; VIGOR Study Group. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. *N Engl J Med.* 2000;343(21):1520-1528.
6. Langman MJ, Jensen DM, Watson DJ, et al. Adverse upper gastrointestinal effects of rofecoxib compared with NSAIDs. *JAMA.* 1999;282(20):1929-1933.

7. Papanikolaou PN, Ioannidis JP. Availability of large-scale evidence on specific harms from systematic reviews of randomized trials. *Am J Med.* 2004;117(8):582-589.
8. Geenen MM, Cardous-Ubbink MC, Kremer LC, et al. Medical assessment of adverse health outcomes in long-term survivors of childhood cancer. *JAMA.* 2007;297(24):2705-2715.
9. Ioannidis JP, Haidich AB, Pappa M, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA.* 2001;286(7):821-830.
10. Carson JL, Strom BL, Soper KA, West SL, Morse ML. The association of nonsteroidal anti-inflammatory drugs with upper gastrointestinal tract bleeding. *Arch Intern Med.* 1987;147(1):85-88.
11. Walter SD. Determination of significant relative risks and optimal sampling procedures in prospective and retrospective comparative studies of various sizes. *Am J Epidemiol.* 1977;105(4):387-397.
12. Leufkens HG, Urquhart J, Stricker BH, Bakker A, Petri H. Channelling of controlled release formulation of ketoprofen (Oscorel) in patients with history of gastrointestinal problems. *J Epidemiol Community Health.* 1992;46(4):428-432.
13. Joseph KS. The evolution of clinical practice and time trends in drug effects. *J Clin Epidemiol.* 1994;47(6):593-598.
14. Ray WA, Griffin MR, Downey W. Benzodiazepines of long and short elimination half-life and the risk of hip fracture. *JAMA.* 1989;262(23):3303-3307.
15. Kern LM, Malhotra S, Barrón Y, et al. Accuracy of electronically reported "meaningful use" clinical quality measures: a cross-sectional study. *Ann Intern Med.* 2013;158(2):77-83.
16. Hiatt RA, Fireman B. The possible effect of increased surveillance on the incidence of malignant melanoma. *Prev Med.* 1986;15(6):652-660.
17. Kristensen P, Irgens LM, Daltveit AK, Andersen A. Perinatal outcome among children of men exposed to lead and organic solvents in the printing industry. *Am J Epidemiol.* 1993;137(2):134-144.
18. Dement JM, Harris RL Jr, Symons MJ, Shy CM. Exposures and mortality among chrysotile asbestos workers. Part II: mortality. *Am J Ind Med.* 1983;4(3):421-433.
19. Herbst AL, Ulfelder H, Poskanzer DC. Adenocarcinoma of the vagina. Association of maternal stilbestrol therapy with tumor appearance in young women. *N Engl J Med.* 1971;284(15):878-881.
20. Spitzer WO, Suissa S, Ernst P, et al. The use of beta-agonists and the risk of death and near death from asthma. *N Engl J Med.* 1992;326(8):501-506.
21. Ray WA, Griffin MR, Schaffner W, Baugh DK, Melton LJ III. Psychotropic drug use and the risk of hip fracture. *N Engl J Med.* 1987;316(7):363-369.
22. Redelmeier DA, Tibshirani RJ. Association between cellular-telephone calls and motor vehicle collisions. *N Engl J Med.* 1997;336(7):453-458.

23. MacMahon B, Yen S, Trichopoulos D, Warren K, Nardi G. Coffee and cancer of the pancreas. *N Engl J Med*. 1981;304(11):630-633.
24. Baghurst PA, McMichael AJ, Slavotinek AH, Baghurst KI, Boyle P, Walker AM. A case-control study of diet and cancer of the pancreas. *Am J Epidemiol*. 1991;134(2):167-179.
25. Lenz W. Epidemiology of congenital malformations. *Ann N Y Acad Sci*. 1965;123:228-236.
26. Soverchia G, Perri PF. [2 cases of malformations of a limb in infants of mothers treated with an antiemetic in a very early phase of pregnancy]. *Pediatr Med Chir*. 1981;3(1):97-99.
27. Holmes LB. Teratogen update: bendectin. *Teratology*. 1983;27(2):277-281.
28. Centers for Disease Control and Prevention (CDC). Intussusception among recipients of rotavirus vaccine--United States, 1998-1999. *MMWR Morb Mortal Wkly Rep*. 1999;48(27):577-581.
29. Murphy TV, Gargiullo PM, Massoudi MS, et al; Rotavirus Intussusception Investigation Team. Intussusception among infants given an oral rotavirus vaccine. *N Engl J Med*. 2001;344(8):564-572.
30. Kellermann AL, Rivara FP, Rushforth NB, et al. Gun ownership as a risk factor for homicide in the home. *N Engl J Med*. 1993;329(15):1084-1091.
31. Browner WS, Li J, Mangano DT; The Study of Perioperative Ischemia Research Group. In-hospital and long-term mortality in male veterans following noncardiac surgery. *JAMA*. 1992;268(2):228-232.
32. Echt DS, Liebson PR, Mitchell LB, et al. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The Cardiac Arrhythmia Suppression Trial. *N Engl J Med*. 1991;324(12):781-788.

*This page intentionally left blank*

11

# The Process of Diagnosis

W. Scott Richardson and Mark C. Wilson

## IN THIS CHAPTER

---

**Clinical Scenarios**

**Two Complementary Approaches to Diagnosis**

**Clusters of Findings Define Clinical Problems**

**Clinicians Select a Small List of Diagnostic Possibilities**

**Estimating the Pretest Probability Facilitates the  
Diagnostic Process**

**New Information Generates Posttest Probabilities**

**The Relation Between Posttest Probabilities and  
Threshold Probabilities Determines Clinical Action**

**Conclusion**

---



## CLINICAL SCENARIOS

**C**onsider the following diagnostic situations:

1. A 43-year-old woman presents with a painful cluster of vesicles grouped in the T3 dermatome of her left thorax, which you recognize as shingles from reactivation of herpes zoster.
2. A 78-year-old man returns to your office for follow-up of hypertension. He has lost 10 kg since his last visit 6 months ago. He describes reduced appetite but otherwise has no localizing symptoms. You recall that his wife died a year ago and consider depression as a likely explanation, yet his age and exposure history (ie, smoking) suggest other possibilities.

## TWO COMPLEMENTARY APPROACHES TO DIAGNOSIS

---

The first case in the opening scenarios illustrates the rapid, non-analytic approach that expert diagnosticians use to recognize disorders they have seen many times before (ie, pattern recognition) and that is particularly relevant to the diagnostic properties of aspects of the physical examination.<sup>1-6</sup> The second case illustrates a more challenging circumstance in which simple pattern recognition fails, so expert diagnosticians slow down and toggle to a more analytic mode of diagnostic thinking.<sup>7,8</sup> This includes the probabilistic approach to clinical diagnosis that uses *evidence* from clinical research—the focus of this chapter (Figure 11-1). Using this probabilistic analytic approach, expert diagnosticians generate a list of potential diagnoses, estimate the *probability* associated with each, and conduct investigations, the results of which increase or decrease the probabilities, until they believe they have found the best answer to fit the patient's illness.<sup>9-14</sup>

FIGURE 11-1

### Pattern Recognition vs Probabilistic Diagnostic Reasoning

Pattern recognition	Probabilistic diagnostic reasoning
See it and recognize disorder	Clinical assessment generates pretest probability
↓	↓
Compare posttest probability with thresholds (usually pattern recognition implies probability near 100% and so above threshold)	New information generates posttest probability (may be iterative)
	↓
	Compare posttest probability with thresholds

Applying the probabilistic approach requires knowledge of human anatomy, pathophysiology, and the taxonomy of disease.<sup>11,12,14</sup> Evidence from clinical research represents another form of knowledge required for optimal diagnostic reasoning.<sup>15-17</sup> This chapter describes how evidence from clinical research can facilitate the probabilistic mode of diagnosis.

## CLUSTERS OF FINDINGS DEFINE CLINICAL PROBLEMS

Using the probabilistic mode, clinicians begin with the medical interview and physical examination, which they use to identify individual findings as potential clues. For instance, in the second scenario, the clinician noted a 10-kg weight loss in 6 months that is associated with anorexia but without localizing symptoms. Experienced clinicians often group findings into meaningful clusters, summarized in brief phrases about the symptom, body location, or organ system involved, such as “involuntary weight loss with anorexia.” These clusters, often termed “clinical problems,” represent the starting point for the probabilistic approach to *differential diagnosis*.<sup>11</sup>

## CLINICIANS SELECT A SMALL LIST OF DIAGNOSTIC POSSIBILITIES

---

When considering a patient's differential diagnosis, clinicians must decide which disorders to pursue. If they considered all known causes to be equally likely and tested for them all simultaneously (the "possibilistic" list), unnecessary testing would result. Instead, experienced clinicians are selective, considering those disorders that are more likely (a probabilistic list), more serious if left undiagnosed and untreated (a prognostic list), or more responsive to treatment (a pragmatic list). Wisely selecting an individual patient's prioritized differential diagnosis involves all 3 of these considerations (probabilistic, prognostic, and pragmatic).

One can label the best explanation for the patient's problem as the leading hypothesis or working diagnosis. In the second scenario, the clinician suspected depression as the most likely cause of the patient's anorexia and weight loss. A few (usually 1-5) other diagnoses may be worth considering at the initial evaluation because of their likelihood, seriousness if left undiagnosed and untreated, or responsiveness to treatment. In the case of unexplained weight loss, the man's age raises the specter of neoplasm, and in particular, his past smoking suggests the possibility of lung cancer.

Additional causes of the problem may be too unlikely to consider at the initial diagnostic evaluation but could arise subsequently if the initial hypotheses are later disproved. Most clinicians considering the 78-year-old man with weight loss would not select a disease that causes malabsorption as their initial differential diagnosis but might turn to this hypothesis if investigation ultimately excludes depression and cancer.

## ESTIMATING THE PRETEST PROBABILITY FACILITATES THE DIAGNOSTIC PROCESS

---

Having assembled a short list of plausible target disorders to be investigated—the differential diagnosis for this patient—clinicians

rank these conditions. The probabilistic approach to diagnosis encourages clinicians to estimate the probability of each target condition on the short list, the *pretest probability* (Figure 11-1).<sup>18</sup> The sum of the probabilities for all candidate diagnoses should equal 1.

How can the clinician estimate these pretest probabilities? One method is implicit, drawing on memories of previous cases with the same clinical problem(s) and using the frequency of disorders found in those previous patients to guide estimates of pretest probability for the current patient. Often, though, memory is imperfect, and we are excessively influenced by particularly vivid or recent experiences and by previous inferences, and we put insufficient weight on new evidence. Furthermore, our experience with a given clinical problem may be limited. All of these factors leave the probabilities arising from clinicians' intuition subject to *bias* and *random error*.<sup>19-21</sup>

A complementary approach uses evidence from research to guide pretest probability estimates. In one type of relevant research, patients with the same clinical problem undergo thorough diagnostic evaluation, yielding a set of frequencies of the underlying diagnoses made, which clinicians can use to estimate the initial pretest probability. Another category of relevant research generates *clinical decision rules* or *prediction rules*. Patients with a defined clinical problem undergo diagnostic evaluation, and investigators use statistical methods to identify clinical and diagnostic test features that segregate patients into subgroups with different probabilities of a target condition.

## NEW INFORMATION GENERATES POSTTEST PROBABILITIES

---

Clinical diagnosis is a dynamic process. As new information arrives, it may increase or decrease the probability of a target condition or diagnosis.<sup>6</sup> For instance, in the older man with involuntary weight loss, the presence of a recent major life event (his wife's death) raises the likelihood that depression

is the cause, whereas the absence of localizing gut symptoms decreases the probability of an intestinal disorder. *Likelihood ratios* capture the extent to which new pieces of information revise probabilities (see Chapter 12, Diagnostic Tests).

Although intuitive estimates based on experience may, at times, serve clinicians well in interpreting test results, confidence in the extent to which a result increased or decreased probabilities requires systematic research. This research can take several forms, most notably individual *primary studies* of test accuracy (see Chapter 12, Diagnostic Tests) and *systematic reviews* of these test accuracy studies (see Chapter 14, The Process of a Systematic Review and Meta-analysis). Once these research results have been appraised for *risk of bias* and applicability, the discriminatory power of the clinical findings or test results can be collected into reference resources useful for each clinical discipline.<sup>22,23</sup>

## THE RELATION BETWEEN POSTTEST PROBABILITIES AND THRESHOLD PROBABILITIES DETERMINES CLINICAL ACTION

---

After the test result generates the *posttest probability*, one can compare this new probability to thresholds (Figure 11-2).<sup>24-26</sup> If the posttest probability is equal to 1, the diagnosis would be absolutely certain. Short of certainty, as the posttest probability approaches 1, the diagnosis becomes more and more likely and reaches a threshold of probability above which the clinician would recommend starting treatment for the disorder (the *treatment threshold*) (Figure 11-2). These thresholds apply to both pattern recognition and probabilistic or *Bayesian diagnostic reasoning* (Figure 11-1). For instance, consider the first scenario, the patient who presents with a painful eruption of grouped vesicles in the distribution of a single dermatome. In



TABLE 11-1

## Factors That Affect the Location of Test Thresholds

Factor	Lowers Test Threshold	Raises Test Threshold
Test safety	Low or zero-risk test	Higher risk (eg, invasive)
Test cost	Low-cost test	Higher cost
Test acceptability to patient	High acceptability	Lower acceptability
Prognosis of target disorder	Serious if left undiagnosed	Less serious if missed
Effectiveness of treatment	Treatment is effective	Treatment is less effective
Availability of treatment	Treatment is available	Treatment is not available

What determines these test and treatment thresholds? They are a function of the properties of the test, the disease prognosis, and the nature of the treatment (Tables 11-1 and 11-2). For the test threshold, the safer and less costly the testing strategy, the more serious the condition if left undiagnosed, and the more effective and safe the available treatment is, the lower we would place the test threshold. On the other hand, the less safe or more costly the test strategy, the less serious the condition if undiagnosed, and the less secure we are about the effectiveness and safety of treatment, the higher we would place the test threshold.

Consider, for instance, ordering a troponin test for suspected acute coronary syndrome. The condition, if present, can lead to serious consequences (such as fatal arrhythmias), and the test is inexpensive and noninvasive. This is the reason one sees emergency department physicians ordering the test for patients with even a very low probability of acute coronary syndrome: they have set a very low diagnostic threshold.

TABLE 11-2

**Factors That Affect the Location of Treatment Thresholds**

<b>Factor</b>	<b>Lowers Treatment Threshold</b>	<b>Raises Treatment Threshold</b>
Safety of next test	Higher risk from tests	Low or zero risk from tests
Costs of next test	More expensive tests	Lower costs of tests
Prognosis of target disorder	Serious if left untreated	Less serious if untreated
Effectiveness of treatment	Highly effective treatment	Less effective treatment
Safety of treatment	Low risk from treatment	Higher risk from treatment
Availability of treatment	Treatment readily available	Treatment is less available

Contrast this with pulmonary angiography for suspected pulmonary embolism. Although the condition is serious, the test is invasive and may be complicated. As a result, if after tests such as Doppler compression ultrasonography and ventilation-perfusion scanning or helical computed tomography they are left with a low probability of pulmonary embolism, clinicians may choose to monitor closely. The test threshold is higher because of the invasiveness and risks of the test.

For the treatment threshold, the safer and the less expensive our next test, the more benign the prognosis of the illness, and the higher the costs or greater the adverse effects of the treatment options, the higher we would place the threshold, requiring greater diagnostic certainty before exposing our patients to treatment. On the other hand, the more invasive and less safe the next test needed, the more ominous the prognosis, and the safer and less costly the proposed treatment, the lower we would place the treatment threshold because proceeding with treatment may



be preferable to increasing diagnostic certainty. For instance, consider patients presenting with suspected malignant tumors. In general, before treating, clinicians will subject such patients to invasive diagnostic tests associated with possible serious complications. The reason is that the treatment—surgery, radiation, or chemotherapy—is itself associated with morbidity or even mortality. Thus, clinicians set the treatment threshold very high.

Contrast this with a patient who presents with symptoms of heartburn and acid reflux. Even if symptoms are atypical, clinicians may be ready to prescribe a proton pump inhibitor for symptom relief rather than subject the patient to endoscopy. The lower treatment threshold is a function of the relatively benign nature of the treatment in relation to the invasiveness of the next test.

## CONCLUSION

---

In this chapter, we outlined the probabilistic tradition of analytic diagnostic reasoning and identified how different types of clinical research evidence can inform our diagnostic decisions and actions. The next chapters highlight particular aspects of the diagnostic process.

## References

1. Elstein AS, Shulman L, Sprafka S. *Medical Problem Solving: An Analysis of Clinical Reasoning*. Cambridge, MA: Harvard University Press; 1978.
2. Schmidt HG, Norman GR, Boshuizen HP. A cognitive perspective on medical expertise: theory and implication. *Acad Med*. 1990;65(10):611-621.
3. Eva KW. What every teacher needs to know about clinical reasoning. *Med Educ*. 2005;39(1):98-106.
4. Norman GR, Brooks LR. The non-analytical basis of clinical reasoning. *Adv Health Sci Educ Theory Pract*. 1997;2(2):173-184.
5. Norman GR. The epistemology of clinical reasoning: perspectives from philosophy, psychology, and neuroscience. *Acad Med*. 2000;75(10)(suppl):S127-S135.
6. Sackett DL. A primer on the precision and accuracy of the clinical examination. In: Simel DL, Rennie D, eds. *The Rational Clinical Examination: Evidence-Based*

- Clinical Diagnosis*. New York, NY: McGraw-Hill; 2009. <http://www.jamaevidence.com/content/3474001>. Accessed January 7, 2014.
7. Moulton CA, Regehr G, Mylopoulos M, MacRae HM. Slowing down when you should: a new model of expert judgment. *Acad Med*. 2007;82(10) (suppl):S109-S116.
  8. Croskerry P. A universal model of diagnostic reasoning. *Acad Med*. 2009; 84(8):1022-1028.
  9. Barrows HS, Pickell GC. *Developing Clinical Problem Solving Skills: A Guide to More Effective Diagnosis and Treatment*. New York, NY: WW Norton; 1991.
  10. Kassirer JP, Wong JB, Kopelman RI. *Learning Clinical Reasoning*. 2nd ed. Baltimore, MD: Williams & Wilkins; 2009.
  11. Barondess JA, Carpenter CCJ, eds. *Differential Diagnosis*. Philadelphia, PA: Lea & Febiger; 1994.
  12. Bordage G. Elaborated knowledge: a key to successful diagnostic thinking. *Acad Med*. 1994;69(11):883-885.
  13. Glass RD. *Diagnosis: A Brief Introduction*. Melbourne, Australia: Oxford University Press; 1996.
  14. Cox K. *Doctor and Patient: Exploring Clinical Thinking*. Sydney, Australia: UNSW Press; 1999.
  15. Kassirer JP. Diagnostic reasoning. *Ann Intern Med*. 1989;110(11):893-900.
  16. Richardson WS. Integrating evidence into clinical diagnosis. In: Montori VM, ed. *Evidence-Based Endocrinology*. Totowa, NJ: Humana Press; 2006:69-89.
  17. Richardson WS. We should overcome the barriers to evidence-based clinical diagnosis! *J Clin Epidemiol*. 2007;60(3):217-227.
  18. Sox HC Jr, Higgins MC, Owens DK, eds. *Medical Decision Making*. Chichester, UK: Wiley-Blackwell; 2013.
  19. Richardson WS. Where do pretest probabilities come from? [editorial, EBM Note]. *Evid Based Med*. 1999;4:68-69.
  20. Richardson WS, Glasziou P, Polashenski WA, Wilson MC. A new arrival: evidence about differential diagnosis. *ACP J Club*. 2000;133(3):A11-A12.
  21. Richardson WS. Five uneasy pieces about pre-test probability. *J Gen Intern Med*. 2002;17(11):882-883.
  22. Fletcher RH, Fletcher SW, Fletcher GS. *Clinical Epidemiology: The Essentials*. 5th ed. Philadelphia, PA: Wolters-Kluwer/Lippincott Williams & Wilkins; 2012.
  23. Straus SE, Glasziou P, Richardson WS, Haynes RB, eds. *Evidence-Based Medicine: How to Practice and Teach It*. 4th ed. Edinburgh, UK: Elsevier/Churchill-Livingstone; 2011.
  24. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med*. 1980;302(20):1109-1117.
  25. Gross R. *Making Medical Decisions: An Approach to Clinical Decision Making for Practicing Physicians*. Philadelphia, PA: ACP Publications; 1999.
  26. Hunink M, Glasziou P, eds. *Decision Making in Health and Medicine: Integrating Evidence and Values*. Cambridge, England: Cambridge University Press; 2001.

*This page intentionally left blank*

# Diagnostic Tests

Toshi A. Furukawa, Sharon E. Straus,  
Heiner C. Bucher, Thomas Agoritsas,  
and Gordon Guyatt

## IN THIS CHAPTER

---

### Introduction

### Clinical Scenario

### Finding the Evidence

### How Serious Is the Risk of Bias?

Did Participating Patients Constitute a Representative Sample of Those Presenting With a Diagnostic Dilemma?

Did the Investigators Compare the Test to an Appropriate, Independent Reference Standard?

Were Those Interpreting the Test and Reference Standard Blind to the Other Result?

Did Investigators Perform the Same Reference Standard in All Patients Regardless of the Results of the Test Under Investigation?

### What Are the Results?

What Likelihood Ratios Were Associated With the Range of Possible Test Results?

(continued on following page)

### **How Can I Apply the Results to Patient Care?**

Will the Reproducibility of the Test Result and its Interpretation Be Satisfactory in My Clinical Setting?

Are the Study Results Applicable to the Patients in My Practice?

Will the Test Results Change My Management Strategy?

Will Patients Be Better Off as a Result of the Test?

### **Clinical Scenario Resolution**

---

## INTRODUCTION

---

In the previous chapter (Chapter 11, The Process of Diagnosis), we explained the process of diagnosis, the way diagnostic test results move clinicians across the *test threshold* and the *therapeutic threshold*, and how to use studies to help obtain an accurate *pretest probability*. In this chapter, we explain how to use an article that addresses the ability of a diagnostic test to move clinicians toward the extremely high (ruling in) and extremely low (ruling out) *posttest probabilities* they seek.

### CLINICAL SCENARIO

#### How Can We Identify Dementia Quickly and Accurately?

**Y**ou are a busy primary care practitioner with a large proportion of elderly patients in your practice. Earlier in the day, you saw a 70-year-old woman who lives alone and has been managing well. On this visit, she informed you of a long-standing problem, joint pain in her lower extremities. During the visit, you get the impression that, as you put it to yourself, “she isn’t quite all there,” although you find it hard to specify further. On specific questioning about memory and function, she acknowledges that her memory is not what it used to be but otherwise denies problems. Pressed for time, you deal with the osteoarthritis and move on to the next patient.

That evening, you ponder the problem of making a quick assessment of your elderly patients when the possibility of cognitive impairment occurs to you. The Mini-Mental State Examination (MMSE), with which you are familiar, takes too long. You wonder if there are any brief instruments that allow a reasonably accurate rapid diagnosis of cognitive impairment to help you identify patients who need more extensive investigation.

## FINDING THE EVIDENCE

---

You formulate the clinical question, “In older patients with suspected cognitive impairment, what is the accuracy of a brief *screening* tool to identify patients who need more extensive investigation for possible dementia?” To conduct a rapid and specific search, you access the PubMed Clinical Queries page (see Chapter 4, Finding Current Best Evidence). Typing in the search terms “identify dementia brief MMSE,” you select “diagnosis” as the clinical study category and “narrow” as the scope of the filter. This search strategy yields 8 citations.

You survey the abstracts, looking for articles that focus on patients with suspected dementia and report accuracy similar to your previous standard, the MMSE. An article that reports results for an instrument named Six-Item Screener (SIS) meets both criteria.<sup>1</sup> You retrieve the full-text article electronically and start to read it, hoping its methods and results will justify using the instrument in your office.

## HOW SERIOUS IS THE RISK OF BIAS?

---

Box 12-1 summarizes our *Users' Guides* for assessing the *risk of bias*, examining the results, and determining the applicability of a study reporting on the accuracy of a diagnostic test.

### **Did Participating Patients Constitute a Representative Sample of Those Presenting With a Diagnostic Dilemma?**

A diagnostic test is useful only if it distinguishes among conditions or disorders that might otherwise be confused. Although most tests can differentiate healthy persons from severely affected

## BOX 12-1

### Users' Guide for an Article About Interpreting Diagnostic Test Results

#### How serious is the risk of bias?

Did participating patients constitute a representative sample of those presenting with a diagnostic dilemma?

Did investigators compare the test to an appropriate, independent reference standard?

Were those interpreting the test and reference standard blind to the other result?

Did all patients receive the same reference standard irrespective of the test results?

#### What are the results?

What likelihood ratios were associated with the range of possible test results?

#### How can I apply the results to patient care?

Will the reproducibility of the test results and their interpretation be satisfactory in my clinical setting?

Are the study results applicable to the patients in my practice?

Will the test results change my management strategy?

Will patients be better off as a result of the test?

ones, this ability will not help us in clinical practice. Studies that confine themselves to florid cases vs asymptomatic healthy volunteers are unhelpful because, when the diagnosis is obvious, we do not need a diagnostic test. Only a study that closely resembles clinical practice and includes patients with mild, early manifestations of the *target condition* can establish a test's true value.



We label studies with unrepresentative patient selection as suffering from *spectrum bias*. There are 3 empirical studies that have systematically examined for various sources of *bias* in studies of diagnostic tests.<sup>2-4</sup> All 3 studies documented bias associated with unrepresentative patient selection.

The story of carcinoembryonic antigen (CEA) testing in patients with colorectal cancer reveals how choosing the wrong spectrum of patients can dash the hopes raised with the introduction of a diagnostic test. A study found that CEA was elevated in 35 of 36 people with known advanced cancer of the colon or rectum. The investigators found much lower levels in healthy people, pregnant women, or patients with a variety of other conditions.<sup>5</sup> The results suggested that CEA might be useful in diagnosing colorectal cancer or even in screening for the disease. In subsequent studies of patients with less advanced stages of colorectal cancer (and, therefore, lower disease severity) and patients with other cancers or other gastrointestinal disorders (and, therefore, different but potentially confused disorders), the accuracy of CEA testing as a diagnostic tool plummeted. Clinicians appropriately abandoned CEA measurement for new cancer diagnosis and screening.

Enrolling *target-positive* patients (those with the underlying condition of interest; in our scenario, people with dementia) and *target-negative* patients (those without the target condition) from separate populations results in overestimates of the diagnostic test's power. This *case-control design* (where cases are known to be target positive and controls are known to be target negative) of a diagnostic test may be likened to a phase 2 efficacy trial: if it fails (ie, the test fails to discriminate target-positive

from target-negative patients), the test is hopeless; if it succeeds, it cannot guarantee real-world effectiveness.

Even if investigators enroll target-positive and target-negative patients from the same population, nonconsecutive patient sampling and retrospective data collection may inflate estimates of diagnostic test performances.

### **Did the Investigators Compare the Test to an Appropriate, Independent Reference Standard?**

The accuracy of a diagnostic test is best determined by comparing it to the “truth.” Readers must assure themselves that investigators have applied an appropriate *reference, criterion, or gold standard* (such as biopsy, surgery, autopsy, or long-term *follow-up* without treatment) to every patient who undergoes the test under investigation.

One way a study can go wrong is if the test that is being evaluated is part of the reference standard. The incorporation of the test into the reference standard is likely to inflate the estimate of the test’s diagnostic power. Thus, clinicians should insist on independence as one criterion for a satisfactory reference standard.

For instance, consider a study that evaluated the utility of abdominojugular reflux for the diagnosis of congestive heart failure. Unfortunately, this study used clinical and radiographic criteria that included the abdominojugular reflux as the reference test.<sup>6</sup> Another example comes from a study evaluating screening instruments for depression in terminally ill people. The authors claimed perfect performance (*sensitivity* of 1.0 and *specificity* of 1.0) for a single question (“Are you depressed?”) to detect depression. Their diagnostic criteria included 9 questions of which one was, “Are you depressed?”<sup>7</sup>

In reading articles about diagnostic tests, if you cannot accept the reference standard (within reason; after all, nothing is perfect), then the article is unlikely to provide trustworthy results.<sup>3</sup>

### **Were Those Interpreting the Test and Reference Standard Blind to the Other Result?**

If you accept the reference standard, the next question is whether the interpreters of the test and reference standard were unaware of the results of the other investigation (*blind* assessment).

Consider how, once clinicians see a pulmonary nodule on a computed tomogram (CT), they can see the previously undetected lesion on the chest radiograph or, once they learn the results of an echocardiogram, they hear a previously inaudible cardiac murmur.

The more likely that knowledge of the reference standard result can influence the interpretation of a test, the greater the importance of independent interpretation. Similarly, the more susceptible the reference standard is to changes in interpretation as a result of knowledge of the test being evaluated, the more important the blinding of the reference standard interpreter. The empirical study of Lijmer et al<sup>2</sup> found bias associated with unblinded assessments, although the magnitude was small.

### **Did Investigators Perform the Same Reference Standard in All Patients Regardless of the Results of the Test Under Investigation?**

The properties of a diagnostic test will be distorted if its results influence whether patients undergo confirmation by the reference standard (*verification*<sup>8,9</sup> or *work-up bias*).<sup>10,11</sup> This can occur in 2 ways.

First, only a selected sample of patients who underwent the index test may be verified by the reference standard. For example, patients with suspected coronary artery disease whose exercise test results are positive may be more likely to undergo coronary angiography (the reference standard) than those whose exercise

test results are negative. This type of verification bias is known as *partial verification bias*.

Second, results of the index test may be verified by different reference standards. Use of different reference tests for positive and negative results is known as *differential verification bias*.

Verification bias proved a problem for the Prospective Investigation of Pulmonary Embolism Diagnosis (PIOPED) study that evaluated the utility of ventilation perfusion scanning in the diagnosis of pulmonary embolism. Patients whose ventilation perfusion scan results were interpreted as “normal/near normal” and “low *probability*” were less likely to undergo pulmonary angiography (69%) than those with more positive ventilation perfusion scans (92%). This is not surprising because clinicians might be reluctant to subject patients with a low probability of pulmonary embolism to the *risks* of angiography.<sup>12</sup>

Most articles would stop here, and readers would have to conclude that the magnitude of the bias resulting from different proportions of patients with high-probability and low-probability ventilation perfusion scans undergoing adequate angiography is uncertain but perhaps large. The PIOPED investigators, however, applied a second reference standard to the 150 patients with low-probability or normal or near-normal scans who did not undergo angiography (136 patients) or in whom angiogram interpretation was uncertain (14 patients). They judged such patients to be free of pulmonary embolism if they did well without treatment. Accordingly, they followed up all such patients for 1 year without treating them with anti-coagulant drugs. No patient developed clinically evident pulmonary embolism during follow-up, allowing us to conclude that patient-important pulmonary embolism

(if we define patient-important pulmonary embolism as requiring anticoagulation therapy to prevent subsequent adverse events) was not present at the time they underwent ventilation perfusion scanning. Thus, the PIOPED study achieved the goal of applying a reference standard assessment to all patients but failed to apply the same standard to all.

## USING THE GUIDE

**T**he study of a brief diagnostic test for cognitive impairment included 2 cohorts. One was a *random sample* of black persons 65 years and older in the general population; the other, a *consecutive sample* of unscreened patients referred by family, caregivers, or health care professionals for cognitive evaluation at the Alzheimer Disease Center. In the former group, the authors included all patients with a high suspicion of dementia on a detailed screening test and a random sample of those with moderate and low suspicion. The investigators faced diagnostic uncertainty in both populations. The populations are not perfect: the former included individuals without any suspicion of dementia, and the latter had already passed an initial screen at the primary care level (indeed, whether to refer for full geriatric assessment is one of the questions you are trying to resolve for the patient who triggered your search for evidence). Fortunately, test properties proved similar in the 2 populations, considerably lessening your concern.

All patients received the SIS, which asks the patient to remember 3 words (apple, table, penny), then to say the day of the week, month, and year, and finally to recall the 3 words without prompts. The number of errors provides a result with a range of 0 to 6.

For the reference standard diagnosis of dementia, patients had to satisfy both *Diagnostic and Statistical Manual of Mental Disorders* (Third Edition Revised) and *International Statistical Classification of Diseases, 10th Revision (ICD-10)* criteria, based on an assessment by a geriatric psychiatrist or a neurologist that included history, physical and neurologic examination, a complete neuropsychological test battery that included the MMSE and 5 other tests, and an interview with a relative of the participant.

Although you are satisfied with this reference standard, the published article leaves you unsure whether those making the SIS and the reference diagnosis were blind to the other result. To resolve the question, you email the first author and ask for clarification. A couple of emails later, you have learned that “research assistants who had been trained and tested” administered the neuropsychological battery. On the other hand, “a consensus team composed of a geriatric psychiatrist, and social psychologist, a geriatrician, and a neuropsychologist” made the reference standard diagnoses. The author reports, “There were open discussions of the case and they had access to the entire medical record including results of neuropsychological testing at their disposal.” The 6 items included in the SIS are derived from the MMSE but “were not pulled out as a separate instrument in the consensus team conference.”

Thus, although there was no blinding, you suspect that this did not create important bias and are therefore ready to consider its results.

## WHAT ARE THE RESULTS?

---

### What Likelihood Ratios Were Associated With the Range of Possible Test Results?

In deciding how to interpret diagnostic tests results, we will consider their ability to change our initial estimate of the likelihood

the patient has the target condition (we call this the pretest probability) to a more accurate estimate (we call this the posttest probability of the target disorder). The *likelihood ratio* (LR) for a particular test result moves us from the pretest probability to a posttest probability.

Put yourself back in the shoes of the primary care physician in the scenario and consider 2 patients with suspected cognitive impairment with clear consciousness. The first is the 70-year-old woman in the clinical scenario who seems to be managing rather well but has a specific issue that her memory is not what it used to be.

The other is an 85-year-old woman, another long-standing patient, who arrives accompanied, for the first time, by her son. The concerned son tells you that she has, on one of her usual morning walks, lost her way. A neighbor happened to catch her a few miles away from home and notified him of the incident. On visiting his mother's house, he was surprised to find her room a mess. However, in your office she greets you politely and protests that she was just having a bad day and does not think the incident warrants any fuss (at which point, the son looks to the ceiling in frustrated disbelief). Your clinical hunches about the probability of dementia for these 2 people (ie, their pretest probabilities) are quite different. In the first woman, the probability is relatively low, perhaps 20%; in the second, relatively high, perhaps 70%.

The results of a formal screening test (eg, the SIS) will not tell us definitively whether dementia is present. Rather the results modify the pretest probability of that condition, yielding a new posttest probability. The direction and magnitude of this change from pretest to posttest probability are determined by the test's properties, and the property of most value is the LR.

We will use the results of the study by Callahan et al to illustrate the usefulness of LRs. [Table 12-1](#) presents the distribution of the SIS scores in the cohort of patients from the study by Callahan et al.

**TABLE 12-1****Six-Item Screener (SIS) Scores in Patients With and Without Dementia and Corresponding Likelihood Ratios**

SIS Score	Dementia	No Dementia	Likelihood Ratio
6	105	2	47
5	64	2	28
4	64	8	7.1
3	45	16	2.5
2	31	35	0.79
1	25	80	0.28
0	11	163	0.06
Total	345	306	

How likely is a test result of 6 among people who have dementia? Table 12-1 indicates that 105 of 345 people (30.4%) with the condition made 6 errors. We can also see that of 306 people without dementia, 2 (0.65%) made 6 errors. How likely is this test result (ie, making 6 errors) in someone with dementia as opposed to someone without?

Determining this requires us to look at the ratio of the 2 likelihoods that we have just calculated ( $30.4/0.65$ ) and equals 47. In other words, the test result of 6 is 47 times as likely to occur in a patient with as opposed to without dementia.

In a similar fashion, we can calculate the LR associated with a test result of each score. For example, the LR for the test score of 5 is  $(64/345)/(2/306) = 28$ . Table 12-1 provides the LR for each possible SIS score.

How can we interpret LRs? Likelihood ratios indicate the extent to which a given diagnostic test result will raise or lower the pretest probability of the target disorder. An LR of 1 tells us that the posttest probability is exactly the same as the pretest

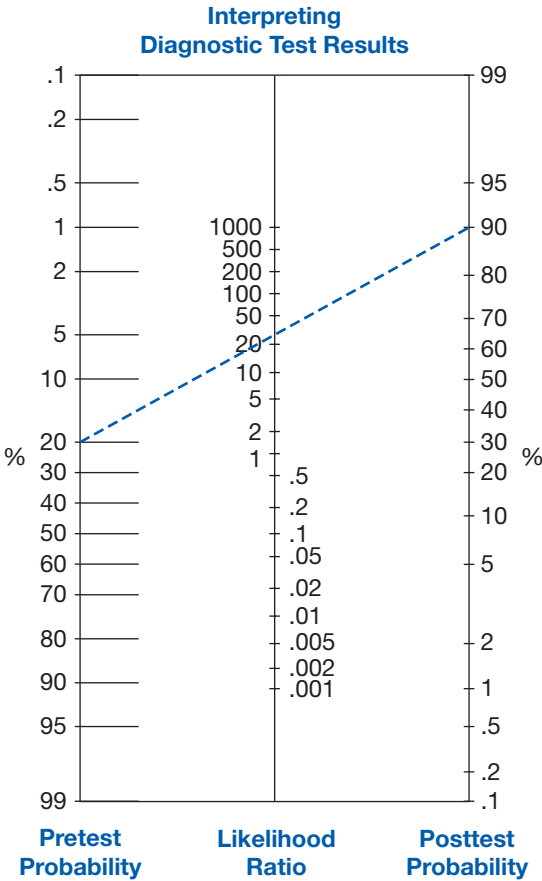


probability. Likelihood ratios greater than 1.0 increase the probability that the target disorder is present; the higher the LR, the greater the increase. Conversely, LRs less than 1.0 decrease the probability of the target disorder, and the smaller the LR, the greater the decrease in probability.

How big is a “big” LR, and how small is a “small” one? Use of LRs in your day-to-day practice will lead to your own sense of their interpretation, but consider the following a rough guide: LRs greater than 10 or less than 0.1 generate large and often conclusive changes from pretest to posttest probability, LRs of 5 to 10 and 0.1 to 0.2 generate moderate shifts in pretest to posttest probability, LRs of 2 to 5 and 0.5 to 0.2 generate small (but sometimes important) changes in probability, and LRs of 1 to 2 and 0.5 to 1 alter probability to a small (and rarely important) degree.

Having determined the magnitude and significance of LRs, how do we use them to go from pretest to posttest probability? One way is to convert pretest probability to odds, multiply the result by the LR, and convert the consequent posttest odds into a posttest probability. If you wonder why the conversion to odds is necessary, consider the fact that LRs compare the likelihood of a test result between patients with and without a target disease (corresponding to the odds of that disease). The calculation is complicated, but there are now several Internet pages and smartphone applications that do this for you ([http://meta.cche.net/clint/templates/calculators/lr\\_nomogram.asp](http://meta.cche.net/clint/templates/calculators/lr_nomogram.asp) and <http://www.cebm.net/nomogram.asp> or <http://medcalc3000.com> and <https://itunes.apple.com/app/twobytwo/id436532323?mt=8>).

When you do not have access to them, one strategy is to use the *nomogram* proposed by Fagan<sup>13</sup> (Figure 12-1), which does all of the conversions and allows an easy transition from pretest to posttest probability. The left-hand column of this nomogram represents the pretest probability, the middle column represents the LR, and the right-hand column represents the posttest probability. You obtain the posttest probability by anchoring a ruler at the pretest probability and rotating it until it lines up with the LR for the observed test result.

**FIGURE 12-1****Likelihood Ratio Nomogram**

Copyright © 1975 Massachusetts Medical Society. All rights reserved. Reproduced from Fagan,<sup>13</sup> with permission from the Massachusetts Medical Society.

Recall the elderly woman from the opening scenario with suspected dementia. We have decided that the probability of this patient having the condition is approximately 20%. Suppose that the patient made 5 errors on the SIS. Anchoring a ruler at her pretest probability of 20% and aligning it with the LR of 28 associated with the test result of 5, you can get a posttest probability of approximately 90%.

The pretest probability is an estimate. Although the literature dealing with differential diagnosis can sometimes help us in establishing the pretest probability, we know of no such study that will complement our intuition in arriving at a pretest probability when the suspicion of dementia arises. Although our intuition does not allow precise estimates of pretest probability, we can deal with residual uncertainty by examining the implications of a plausible range of pretest probabilities.

For example, if the pretest probability in this case is as low as 10% or as high as 30%, using the nomogram, we will get the posttest probability of approximately 80% and above 90%. [Table 12-2](#) tabulates the posttest probabilities corresponding with each possible SIS score for the 65-year-old woman in the clinical scenario.

We can repeat this exercise for our second patient, the 85-year-old woman who had lost her way. You estimate that her history and presentation are compatible with a 70% probability of dementia. Using our nomogram ([Figure 12-1](#)), the posttest probability with an SIS score of 6 or 5 is almost 100%; with an SIS score of 4, it is 94%; with an SIS score of 3, it is 85%; and so on. The pretest probability (with a range of possible pretest probabilities of 60% to 80%), LRs, and posttest probabilities associated with each of these possible SIS scores are presented in [Table 12-3](#).

Having learned to use LRs, you may be curious about where to find easy access to the LRs of the tests you use regularly in your own practice. The Rational Clinical Examination<sup>14</sup> is a series of *systematic reviews* of the diagnostic properties of the history

**TABLE 12-2****Pretest Probabilities, Likelihood Ratios of the Six-Item Screener, and Posttest Probabilities in the 70-Year-Old Woman With Moderate Suspicion of Dementia**

<b>Pretest Probability, % (Range)<sup>a</sup></b>	<b>SIS Score (LR)</b>	<b>Posttest Probability, % (Range)<sup>a</sup></b>
20 (10-30)	6 (47)	92 (84-95)
	5 (28)	88 (76-92)
	4 (7.1)	64 (44-75)
	3 (2.5)	38 (22-52)
	2 (0.79)	16 (8-25)
	1 (0.28)	7 (3-11)
	0 (0.06)	1 (1-3)

Abbreviations: LR, likelihood ratio; SIS, Six-Item Screener.

<sup>a</sup>The values in parentheses represent a plausible range of pretest probabilities. That is, although the best guess as to the pretest probability is 20%, values of 10% to 30% would also be reasonable estimates.

and physical examination that have been published in *JAMA* (an updated database is available on the JAMAevidence homepage at <http://jamaevidence.com/resource/523>). Additional examples of LRs are accumulated on the JAMAevidence website (<http://www.jamaevidence.com>).

### **Dichotomizing Continuous Test Scores: Sensitivity, Specificity, and Likelihood Ratios**

Readers who have followed the discussion to this point will understand the essentials of interpretation of diagnostic tests. In part because they remain in wide use, it is also helpful to understand 2 other terms in the lexicon of diagnostic testing: sensitivity and specificity. Many articles that address diagnostic tests report a  $2 \times 2$  table and its associated sensitivity and

TABLE 12-3

**Pretest Probabilities, Likelihood Ratios of the Six-Item Screener, and Posttest Probabilities in the 85-Year-Old Woman With High Suspicion of Dementia**

Pretest Probability, % (Range) <sup>a</sup>	SIS Score (LR)	Posttest Probability, % (Range) <sup>a</sup>
70 (60-80)	6 (47)	99 (99-99)
	5 (28)	98 (98-99)
	4 (7.1)	94 (91-97)
	3 (2.5)	85 (79-76)
	2 (0.79)	65 (54-76)
	1 (0.28)	40 (30-53)
	0 (0.06)	12 (8-19)

Abbreviations: LR, likelihood ratio; SIS, Six-Item Screener.

<sup>a</sup>The values in parentheses represent a plausible range of pretest probabilities. That is, although the best guess as to the pretest probability is 20%, values of 60% to 80% would also be reasonable estimates.

specificity, as in Table 12-4, and to go along with it a figure that depicts the overall power of the diagnostic test (called a *receiver operating characteristic curve*).

Sensitivity is the proportion of people with a positive test result among those with the target condition. Specificity is the proportion of people with a negative test result among those without the target condition.

The study by Callahan et al<sup>1</sup> recommends a cutoff of 3 or more errors for the diagnosis of dementia. Table 12-5 provides the breakdown of the cohort of referred patients according to this cutoff.

When we set the cutoff of 3 or more, the SIS has a sensitivity of 0.81 (278/345) and a specificity of 0.91 (278/306). We can also calculate the LRs, exactly as we did in Table 12-1. The LR for an SIS score of 3 or greater is therefore  $(278/345)/(28/306) = 8.8$ , and the LR for an SIS score less than 3 is  $(67/345)/(278/306) = 0.21$ . The LR for a positive test result is often denoted as LR+ and that for a negative test result as LR-.

TABLE 12-4

**Comparison of the Results of a Diagnostic Test With the Results of Reference Standard Using a  $2 \times 2$  Table**

Test Results	Reference Standard	
	Disease Present	Disease Absent
Test result positive	True positive (TP)	False positive (FP)
Test result negative	False negative (FN)	True negative (TN)

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}$$

Likelihood Ratio for

$$\begin{aligned} \text{Positive Test Result (LR+)} &= \frac{\text{Sensitivity}}{1 - \text{Specificity}} = \frac{\text{TP rate}}{\text{FP rate}} \\ &= \frac{\text{TP}/(\text{TP} + \text{FN})}{\text{FP}/(\text{FP} + \text{TN})} \end{aligned}$$

Likelihood Ratio for

$$\begin{aligned} \text{Negative Test Result (LR-)} &= \frac{1 - \text{Sensitivity}}{\text{Specificity}} = \frac{\text{FN rate}}{\text{TN rate}} \\ &= \frac{\text{FN}/(\text{TP} + \text{FN})}{\text{TN}/(\text{FP} + \text{TN})} \end{aligned}$$

Let us now try to resolve our clinical scenario using this dichotomized  $2 \times 2$  table. We had supposed that the pretest probability for the woman in the opening scenario was 20% and she had made 5 errors. Because the SIS score of 5 is associated here with an LR+ of 8.8, using Fagan's nomogram, we arrive at the posttest probability of approximately 70%, a figure considerably lower than the 90% that we had arrived at when we had a specific LR for 5 errors. This is because the dichotomized LR+ for SIS scores of 3 or more pooled strata for SIS scores of 3, 4, 5, and 6, and the resultant LR is thus diluted by the adjacent strata.

TABLE 12-5

**Comparison of the Results of a Diagnostic Test (Six-Item Screener) With the Results of Reference Standard (Consensus *DSM-IV* and *ICD-10* Diagnosis) Using the Recommended Cutoff**

SIS Score	Dementia	No Dementia
≥3	278	28
<3	67	278
Total	345	306

Abbreviations: *DSM-IV*, *Diagnostic and Statistical Manual of Mental Disorders* (Fourth Edition); *ICD-10*, *International Classification of Diseases, 10th Revision*; SIS, Six-Item Screener.

Although the difference between 70% and 90% may not dictate change in management strategies for the case in the clinical scenario, this will not always be the case. Consider a third patient, an elderly gentleman with a pretest probability of 50% of dementia who has surprised us by not making a single error on the SIS. With the dichotomous LR+/LR− approach (or, for that matter, with the sensitivity and specificity approach because these are mathematically equivalent and interchangeable), you combine the pretest probability of 50% with the LR− of 0.21 and arrive at the posttest probability of approximately 20%, very likely necessitating further neuropsychological and other examinations. The true posttest probability for this man when we apply the LR associated with a score of 0 from Table 12-1 (0.06) is only approximately 5%. With this posttest probability, you (and the patient and his family) can feel relieved and, at least for the time being, be spared further testing.

In summary, use of multiple cuts or thresholds (sometimes referred to as multilevel LRs or stratum-specific LRs) has 2 key advantages over the sensitivity and specificity approach. First, for a test that produces continuous scores or a number of categories (which many tests in medicine do, notably many laboratory

tests), use of multiple thresholds retains as much information as possible. Second, knowing the LR of a particular test result, one can use a simple nomogram to move from the pretest to the posttest probability that is linked to your own patient.

## USING THE GUIDE

**T**hus far, we have established that the results are likely true for the people who were included in the study, and we have calculated the multilevel LRs associated with each possible score of the test. We have indicated how the results could be applied to our patient (although we do not yet know the patient's score and have not decided how to proceed when we do).

## HOW CAN I APPLY THE RESULTS TO PATIENT CARE?

---

### **Will the Reproducibility of the Test Result and Its Interpretation Be Satisfactory in My Clinical Setting?**

The value of any test depends on its ability to yield the same result when reapplied to stable patients. Poor *reproducibility* can result from problems with the test itself (eg, variations in reagents in radioimmunoassay kits for determining hormone levels) or from its interpretation (eg, the extent of ST-segment elevation on an electrocardiogram). You can easily confirm this when you recall the clinical disagreements that arise when you and one or more colleagues examine the same electrocardiogram, ultrasonogram, or CT (even when all of you are experts).

Ideally, an article about a diagnostic test will address the reproducibility of the test results using a measure that corrects



for agreement by chance, especially for issues that involve interpretation or judgment.

If the reported reproducibility of a test in the study setting is mediocre and disagreement between observers is common, and yet the test still discriminates well between those with and without the target condition, the test is likely to be very useful. Under these circumstances, there is a good chance that the test can be readily applied to your clinical setting.

If reproducibility of a diagnostic test is very high, either the test is simple and unambiguous or those interpreting the results are highly skilled. If the latter applies, less skilled interpreters in your own clinical setting may not do as well. You will either need to obtain appropriate training (or ensure that those interpreting the test in your setting have that training) or look for an easier and more robust test.

### **Are the Study Results Applicable to the Patients in My Practice?**

Test properties may change with a different mix of disease severity or with a different distribution of competing conditions. When patients with the target disorder all have severe disease, LRs will move away from a value of 1.0 (ie, sensitivity increases). If patients are all mildly affected, LRs move toward a value of 1.0 (ie, sensitivity decreases). If patients without the target disorder have competing conditions that mimic the test results seen in patients who have the target disorder, the LRs will move closer to 1.0, and the test will appear less useful (ie, specificity decreases). In a different clinical setting in which fewer of the disease-free patients have these competing conditions, the LRs will move away from 1.0, and the test will appear more useful (ie, specificity increases). Differing prevalence in your setting may alert you to the possibility that the spectrum of target-positive and target-negative patients could differ in your practice.<sup>15</sup>

Investigators have reported the phenomenon of differing test properties in different subpopulations for exercise

electrocardiography in the diagnosis of coronary artery disease. The more severe the coronary artery disease, the larger the LRs of abnormal exercise electrocardiograph results for angiographic narrowing of the coronary arteries.<sup>16</sup> Another example comes from the diagnosis of venous thromboembolism, where compression ultrasonography for proximal-vein thrombosis has proved more accurate in symptomatic outpatients than in asymptomatic postoperative patients.<sup>17</sup>

Sometimes, a test fails in just the patients one hopes it will best serve. The LR of a negative dipstick test result for the rapid diagnosis of urinary tract infection is approximately 0.2 in patients with clear symptoms and thus a high probability of urinary tract infection but is higher than 0.5 in those with low probability,<sup>18</sup> rendering it of little help in ruling out infection in the latter situation.

If you practice in a setting similar to that of the study and if the patient under consideration meets all of the study eligibility criteria, you can be confident that the results are applicable. If not, you must make a judgment. As with therapeutic interventions, you should ask whether there are compelling reasons why the results should not be applied to the patients in your practice, either because of the severity of disease in those patients or because the mix of competing conditions is so different that generalization is unwarranted. You may resolve the issue of *generalizability* if you can find a systematic review that summarizes the results of a number of studies.<sup>19</sup>

### **Will the Test Results Change My Management Strategy?**

It is useful, when making and communicating management decisions, to link them explicitly to the probability of the target disorder. For any target disorder there are probabilities below which a clinician would dismiss a diagnosis and order no further tests: the test threshold. Similarly, there are probabilities above which a clinician would consider the diagnosis confirmed and would stop testing and initiate treatment (ie, the treatment

threshold). When the probability of the target disorder lies between the test and treatment thresholds, further testing is mandated (see Chapter 11, The Process of Diagnosis).

If most patients have test results with LRs near 1.0, test results will seldom move us across the test or treatment threshold. Thus, the usefulness of a diagnostic test is strongly influenced by the proportion of patients suspected of having the target disorder whose test results have very high or very low LRs. Among the patients suspected of having dementia, a review of Table 12-1 allows us to determine the proportion of patients with extreme results ( $LR > 10$  or  $< 0.1$ ). The proportion can be calculated as  $(105 + 2 + 64 + 2 + 11 + 163) / (345 + 306)$  or  $347 / 651 = 53\%$ . The SIS is likely to move the posttest probability in a decisive manner in half of the patients suspected of having dementia and examined—a very impressive proportion and better than for most of our diagnostic tests.

A final comment has to do with the use of sequential tests. The LR approach fits in particularly well in thinking about the diagnostic pathway. Each item of history—or each finding on physical examination—represents a diagnostic test in itself. We can use one test to get a certain posttest probability that can be further increased or decreased by using another, subsequent test. In general, we can also use laboratory tests or imaging procedures in the same way. If 2 tests are very closely related, however, application of the second test may provide little or no additional information, and the sequential application of LRs will yield misleading results. For example, once one has the results of the most powerful laboratory test for iron deficiency, serum ferritin, additional tests, such as serum iron or transferrin saturation, add no further useful information.<sup>20</sup> Once one has conducted an SIS, additional information from the MMSE is likely to be minimal.

*Clinical prediction rules* deal with the lack of independence of a series of tests and provide the clinician with a way of combining their results. For instance, in patients with suspected pulmonary embolism, one could use a rule that incorporates

leg symptoms, heart rate, hemoptysis, and other aspects of the history and physical examination to accurately classify patients with suspected pulmonary embolism as being characterized by high, medium, and low probability.<sup>21</sup>

### Will Patients Be Better Off as a Result of the Test?

The ultimate criterion for the usefulness of a diagnostic test is whether the benefits that accrue to patients are greater than the associated risks.<sup>22</sup> How can we establish the benefits and risks of applying a diagnostic test? The answer lies in thinking of a diagnostic test as a therapeutic maneuver (see Chapter 6, Therapy [Randomized Trials]). Establishing whether a test does more good than harm will involve (1) randomizing patients to a diagnostic strategy that includes the test under investigation and a management schedule linked to it, or to one in which the test is not available, and (2) following up patients in both groups forward in time to determine the frequency of *patient-important outcomes*.

When is demonstrating accuracy sufficient to mandate the use of a test and when does one require a *randomized clinical trial*? The value of an accurate test will be undisputed when the target disorder is dangerous if left undiagnosed, if the test has acceptable risks, and if effective treatment exists. This is the case for the CT-angiogram for suspected pulmonary embolism. A high probability or normal or near-normal results of the CT-angiogram may well eliminate the need for further investigation and may result in anticoagulant agents being appropriately given or appropriately withheld (with either course of action having a substantial positive influence on patient outcome).

Sometimes, a test may be completely benign, represent a low resource investment, be evidently accurate, and clearly lead to useful changes in management. Such is the case for use of the SIS in patients with suspected dementia, when test results may dictate reassurance or extensive investigation and ultimately planning for a tragic deteriorating course.

In other clinical situations, tests may be accurate and management may even change as a result of their application, but their effect on patient outcome may be far less certain. Consider one of the issues we raised in our discussion of framing clinical questions (see Chapter 3, What Is the Question?). There, we considered a patient with apparently resectable non-small cell carcinoma of the lung and wondered whether the clinician should order a positron emission tomogram (PET)-CT and base further management on the results or use alternative diagnostic strategies. For this question, knowledge of the accuracy of CT is insufficient. A randomized trial of PET-CT-directed management or an alternative strategy for all patients is warranted. Other examples include catheterization of the right side of the heart for critically ill patients with uncertain hemodynamic status and bronchoalveolar lavage for critically ill patients with possible pulmonary infection. For these tests, randomized trials have helped elucidate optimal management strategies.

## CLINICAL SCENARIO RESOLUTION

**A**lthough the study itself does not report reproducibility, its scoring is simple and straightforward because you need only count the number of errors made to 6 questions. The SIS does not require any props or visual cues and is therefore unobtrusive, easy to administer, and takes only 1 to 2 minutes to complete (compared with 5 to 10 minutes for the MMSE). Although you note that trained research staff administered the SIS, the appendix of the article gives a detailed, word-by-word instruction on how to administer the SIS. You believe that you too can administer this scale reliably.

The patient in the clinical scenario is an older woman who was able to come to your clinic by herself but appeared no longer as lucid as she used to be. The Alzheimer Disease Center cohort in the study we had been examining in this chapter consists of

people suspected of having dementia by their caregivers and brought to a tertiary care center directly. Their test characteristics were reported to be similar to those observed in the general population cohort, that is, in a sample with less severe presentations. You decide that there is no compelling reason that the study results would not apply to your patient.

You invite your patient back to the office for a follow-up visit and administer the SIS. The result is a score of 4, which, given your pretest probability of 20%, increases the probability to more than 60%. After hearing that you are concerned about her memory and possibly about her function, she agrees to a referral to a geriatrician for more extensive investigation.

## References

1. Callahan CM, Unverzagt FW, Hui SL, Perkins AJ, Hendrie HC. Six-item screener to identify cognitive impairment among potential subjects for clinical research. *Med Care*. 2002;40(9):771-781.
2. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*. 1999;282(11):1061-1066.
3. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ*. 2006;174(4):469-476.
4. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med*. 2004;140(3):189-202.
5. Thomson DM, Krupey J, Freedman SO, Gold P. The radioimmunoassay of circulating carcinoembryonic antigen of the human digestive system. *Proc Natl Acad Sci U S A*. 1969;64(1):161-167.
6. Marantz PR, Kaplan MC, Alderman MH. Clinical diagnosis of congestive heart failure in patients with acute dyspnea. *Chest*. 1990;97(4):776-781.
7. Chochinov HM, Wilson KG, Enns M, Lander S. "Are you depressed?" screening for depression in the terminally ill. *Am J Psychiatry*. 1997;154(5):674-676.
8. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*. 1983;39(1):207-215.

9. Gray R, Begg CB, Greenes RA. Construction of receiver operating characteristic curves when disease verification is subject to selection bias. *Med Decis Making*. 1984;4(2):151-164.
10. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med*. 1978;299(17):926-930.
11. Choi BC. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. *J Clin Epidemiol*. 1992;45(6):581-586.
12. PLOPED Investigators. Value of the ventilation/perfusion scan in acute pulmonary embolism: results of the Prospective Investigation of Pulmonary Embolism Diagnosis (PIOPED). *JAMA*. 1990;263(20):2753-2759.
13. Fagan TJ. Letter: Nomogram for Bayes theorem. *N Engl J Med*. 1975;293(5):257.
14. Sackett DL, Rennie D. The science of the art of the clinical examination. *JAMA*. 1992;267(19):2650-2652.
15. Leeflang MM, Rutjes AW, Reitsma JB, Hooft L, Bossuyt PM. Variation of a test's sensitivity and specificity with disease prevalence. *CMAJ*. 2013;185(11):E537-E544.
16. Hlatky MA, Pryor DB, Harrell FE Jr, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. *Am J Med*. 1984;77(1):64-71.
17. Ginsberg JS, Caco CC, Brill-Edwards PA, et al. Venous thrombosis in patients who have undergone major hip or knee surgery: detection with compression US and impedance plethysmography. *Radiology*. 1991;181(3):651-654.
18. Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Ann Intern Med*. 1992;117(2):135-140.
19. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM; Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. *Ann Intern Med*. 2008;149(12):889-897.
20. Guyatt GH, Oxman AD, Ali M, Willan A, McIlroy W, Patterson C. Laboratory diagnosis of iron-deficiency anemia: an overview. *J Gen Intern Med*. 1992;7(2):145-153.
21. van Belle A, Büller HR, Huisman MV, et al; Christopher Study Investigators. Effectiveness of managing suspected pulmonary embolism using an algorithm combining clinical probability, D-dimer testing, and computed tomography. *JAMA*. 2006;295(2):172-179.
22. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. *CMAJ*. 1986;134(6):587-594.

# Prognosis

Adrienne G. Randolph, Deborah J. Cook,  
and Gordon Guyatt

## IN THIS CHAPTER

---

### **Clinical Scenario**

#### **Finding the Evidence**

#### **Why and How We Measure Prognosis**

#### **How Serious Is the Risk of Bias?**

Was the Sample of Patients in a Study Representative?

Were the Patients Classified Into Prognostically  
Similar Groups?

Was Study Follow-up Sufficiently Complete?

Were Study Outcome Criteria Objective and Unbiased?

#### **What Are the Results?**

How Likely Are the Outcomes Over Time?

How Precise Are the Estimates of Likelihood?

#### **How Can I Apply the Results to Patient Care?**

Were the Study Patients and Their Management Similar  
to Those in My Practice?

(continued on following page)



Was Follow-up Sufficiently Long?

Can I Use the Study Results in the Management  
of Patients in My Practice?

**Clinical Scenario Resolution**

---

## CLINICAL SCENARIO

You are a pediatrician expecting to see an infant who was born at 26 weeks' gestation tomorrow for her first outpatient clinic visit at 4 months after birth. You know the family well because you care for their older child who was born at 35 weeks' gestation and is now a healthy 3-year-old girl. This infant had a prolonged stay in the neonatal intensive care unit but required relatively minimal respiratory support during her first 3 weeks of life. The neonatologist told you that the infant did extremely well, experiencing none of the complications that often occur in extremely preterm infants. He also informs you that he warned the family, "Your baby is at risk for long-term neurocognitive and motor complications related to being born so prematurely. Although some babies born this prematurely grow up to lead normal lives, many have minor disabilities, and there is a nontrivial chance that your baby could develop moderate to severe disabilities." You have 5 other children in your pediatric practice born at less than 27 weeks of gestation; all of them have major neurodevelopmental problems. On the basis of your professional experience, you wonder if the neonatologist has presented the family with an overly optimistic outlook. You decide to check out the *evidence* for yourself.

## FINDING THE EVIDENCE

---

You use your clinic's free Internet connection to access MEDLINE at the National Library of Medicine website via PubMed. To find the appropriate search terms for your population of interest, you first type "premature" in the *Medical Subject Headings* database and find that there is a term called "Infant, Extremely Premature" defined as a human

infant born before 28 weeks' gestation. You select it and click on the related link for Clinical Queries. Under Clinical Study Categories, you choose the search filter "Prognosis" and limit the scope to "Narrow." This retrieves 31 clinical studies and 5 potential reviews. You first look for a *systematic review* but do not find one that is relevant for evaluating outcomes across multiple extremely premature infant cohorts. However, the second *primary study* in the search results seems promising: Neurodevelopmental Outcome of Extremely Preterm Infants at 2.5 Years After Active Perinatal Care in Sweden.<sup>1</sup> This study reports the cognitive, language, and motor development of a prospective cohort of a *consecutive sample* of extremely preterm infants born before 27 weeks' gestation in Sweden between 2004 and 2007.<sup>1</sup>

## WHY AND HOW WE MEASURE PROGNOSIS

---

Clinicians help patients in 3 broad ways: diagnosing or ruling out medical and health-related problems, administering treatment that does more good than *harm*, and giving them an indication of what the future is likely to hold. Clinicians require studies of *prognosis*—those examining the possible outcomes of a disease and the probability with which they can be expected to occur—to achieve the second and third goals.

Knowledge about prognosis can help clinicians make the right treatment decisions. If a patient is likely to improve without intervention, clinicians should not recommend treatments, particularly those that are expensive or potentially toxic. If a patient is at low *risk* of adverse outcomes, even beneficial treatments may not be worthwhile. On the other hand, some patients will experience poor outcomes regardless of which treatments are offered by the clinician. Whatever the treatment possibilities,

by understanding prognosis and presenting the expected future course of a patient's illness, clinicians can offer reassurance and hope or preparation for long-term disability or death.

To estimate a patient's prognosis, we examine outcomes in groups of patients with a similar clinical presentation. We may then refine our prognosis by looking at subgroups defined by demographic variables, such as age, and by *comorbidity* and decide in which subgroup the patient belongs. When these variables or factors influence which patients do better or worse, we call them *prognostic factors*.

In this chapter, we focus on how to use articles that may contain trustworthy prognostic information that clinicians will find useful for counseling patients (Box 13-1).

## BOX 13-1

### Users' Guides to an Article About Prognosis

#### How serious is the risk of bias?

- Was the sample of patients representative?
- Were the patients classified into prognostically homogeneous groups?
- Was follow-up sufficiently complete?
- Were outcome criteria objective and unbiased?

#### What are the results?

- How likely are the outcomes over time?
- How precise are the estimates of likelihood?

#### How can I apply the results to patient care?

- Were the study patients and their management similar to those in my practice?
- Was follow-up sufficiently long?
- Can I use the results in the management of patients in my practice?

## HOW SERIOUS IS THE RISK OF BIAS?

---

### Was the Sample of Patients in a Study Representative?

*Bias* has to do with systematic differences from the truth. A prognostic study is biased if it systematically overestimates or underestimates the likelihood of adverse outcomes in the group of patients under study. When a study sample is systematically different from the population of interest and is biased because patients will have a better or worse prognosis than those in the population of interest, we label the sample as unrepresentative.

How can you recognize an unrepresentative sample? First, determine whether patients pass through some sort of filter before entering the study. If they do, the result is likely a sample that is systematically different from the underlying population of interest. One such filter is the sequence of referrals that leads patients from primary to tertiary centers. Tertiary centers often care for patients with rare and unusual disorders or increased illness severity. Research describing the outcomes of patients in tertiary centers may not be applicable to the general patient with the disorder in the community (otherwise known as *referral bias*).

As an example, chronic hepatitis caused by infection with hepatitis C virus (HCV) can, after many years, lead to liver fibrosis, cirrhosis, and even hepatocellular carcinoma. Researchers have found that rates of progression to cirrhosis as diagnosed by liver biopsy can vary markedly, depending on how the patients are recruited.<sup>2</sup> For a group of patients coming from the same demographic areas or health care settings, the mean estimated 20-year probability of progression to cirrhosis from their initial liver biopsy varied from 6% to 12% to 23%, depending, respectively, on whether the patients were recruited from a population-based posttransfusion HCV surveillance registry, referrals to general hospitals, or a tertiary referral center.<sup>2</sup> Those in a tertiary referral cohort may have other *risk factors* that predispose them to develop cirrhosis at higher rates than other patients.

## Were the Patients Classified Into Prognostically Similar Groups?

Prognostic studies are most useful if individual members of the entire group of study participants are similar enough that the outcome of the group is applicable to each participant. This will be true only if patients are at a similar well-described point in their disease process. The point in the clinical course need not be early, but it does need to be consistent. For instance, studies that evaluate the prognosis of patients with spinal cord injury could focus on in-hospital mortality right after the acute injury, patient outcomes after initial transfer to a rehabilitation center, or the ability of a group of patients to cope independently from the point of discharge to home.

After ensuring that patients were at the same disease stage, you must consider other factors that might influence patient outcome. If factors such as age or disease severity influence prognosis, then providing a single prognosis for young and old and those with mild and severe disease will be misleading for each of these subgroups. For instance, a study that evaluated the outcomes of 8509 patients with traumatic brain injury<sup>3</sup> found that for each increase in age equal to the interquartile range of the patients (24 years), there was approximately double the risk of an unfavorable outcome (death or severe or moderate disability) at 6 months after injury. Patients with a more severe initial neurologic presentation as indicated by bilateral or unilateral absence of pupillary reactivity and those with no response or extensor response on the motor activity subcategory of the Glasgow Coma Scale also had markedly higher risk of having an unfavorable outcome. The percentage of patients having an unfavorable outcome at 6 months increased from 35% to 59% to 77% for patients in whom both pupils, 1 pupil, or neither pupil, respectively, was reactive on initial evaluation. Providing an overall intermediate prognosis across the entire study group (a 48% chance of an unfavorable outcome) to the family of a 20-year-old man who presented with reactive pupils could profoundly mislead them.

Not only must investigators consider all important prognostic factors, they must also consider prognostic factors in relation to one another. If sickness but not age truly determines outcome, and sicker patients tend to be older, investigators who fail to simultaneously consider age and severity of illness may mistakenly conclude that age is an important prognostic factor. For example, investigators in the Framingham study examined risk factors for stroke.<sup>4</sup> They reported that the rate of stroke in patients with atrial fibrillation and rheumatic heart disease was 41 per 1000 person-years, which was similar to the rate for patients with atrial fibrillation but without rheumatic heart disease. Patients with rheumatic heart disease were, however, much younger than those who did not have rheumatic heart disease. To properly understand the influence of rheumatic heart disease, investigators in these circumstances must consider separately the relative risk of stroke in young people with and without rheumatic disease and the risk of stroke in elderly people with and without rheumatic disease. We call this separate consideration an *adjusted analysis*. Once adjustments were made for age, the investigators found that the rate of stroke was 6-fold greater in patients with rheumatic heart disease and atrial fibrillation than in patients with atrial fibrillation who did not have rheumatic heart disease.

If a large number of variables have a major effect on prognosis, investigators should use statistical techniques, such as *regression analysis*, to determine the most powerful predictors. Such an analysis may lead to a *clinical decision rule* that guides clinicians in simultaneously considering all of the important prognostic factors.

How can you decide whether the groups are sufficiently similar with respect to their risk? On the basis of your clinical experience and your understanding of the biology of the condition under study, can you think of factors that the investigators have neglected that are likely to define subgroups with very different prognoses? To the extent that the answer is yes, the risk of bias increases.

## Was Study Follow-up Sufficiently Complete?

Investigators who lose track of a large number of patients increase the risk of bias associated with their prognostic study. The reason is that those who are followed up may have systematically higher or lower risk than those not followed up. As the number of patients who do not return for *follow-up* increases, the risk of bias also increases.

How many patients *lost to follow-up* is too many? The answer depends on the association between the proportion of patients who are lost and the proportion of patients who have had the adverse outcome of interest—the larger the number of patients whose fate is unknown relative to the number who have had the adverse event, the greater the risk of bias. For instance, let us assume that 30% of a particularly high-risk group (such as elderly patients with diabetes) have had an adverse outcome (such as cardiovascular death) during long-term follow-up in a study. If 10% of the patients have been lost to follow-up, the true rate of patients who had died may be as low as approximately 27% or as high as 37%. Across this range, the clinical implications would not change appreciably, and the loss to follow-up does not increase the risk of bias of the study. However, in a much lower-risk patient sample (otherwise healthy middle-aged patients, for instance), the observed event rate may be 1%. In this case, if we assumed that all 10% of the patients lost to follow-up had died, the event rate of 11% might have very different implications.

A large loss to follow-up constitutes a more serious risk of bias when the patients who are lost may be different from those who are easier to find. In one study, for example, investigators managed to follow up 180 of 186 patients treated for neurosis.<sup>5</sup> Of the 180 patients successfully followed up, 60% were easily traced. The death rate in these patients was 3%. The other 40% of the 180 were more difficult to find. The death rate in these patients was 27%.



## Were Study Outcome Criteria Objective and Unbiased?

Outcome events may be objective and easily measured (eg, death), require some judgment (eg, myocardial infarction), or require considerable judgment and effort to measure (eg, disability, quality of life). Investigators should clearly specify and define their *target outcomes* and, whenever possible, base their criteria on objective measures.

The study of children with brain injury in a prolonged unconscious state provides a good example of the challenges involved in measuring outcome.<sup>6</sup> The study investigators found that children's family members frequently interpreted their interactions with the children with unfounded optimism. The investigators therefore required that family members' reports of development of a social response in the affected children be verified by study personnel.

### USING THE GUIDE

**R**eturning to our opening clinical scenario, the investigators who evaluated the outcome of extremely premature infants<sup>1</sup> captured the outcome of all infants born at less than 27 weeks' gestation in Sweden in a setting in which active perinatal care was available. This included easy and free access to care, a low threshold to provide life support at delivery, and transfer of extremely premature infants to specialized units in tertiary care centers. Because this is a population-based sample, it is likely to be representative and free of referral bias. The infants were classified into prognostic groups based on their gestational age at birth, which is known to be a strong prognostic factor. Of 707 extremely premature infants who were born alive, 497 (70%) were still alive at 1 year of age. Neurodevelopmental outcomes were assessed at 2.5 years of age in 456 (92%) of these infants. The most common nonmortality reason for loss to follow-up was an error in the identity number assigned at birth.

Trained psychologists evaluated cognitive, language, and motor development using the Bayley Scales of Infant and Toddler Development (Bayley-III). Because the Bayley-III had not been standardized in Sweden, the investigators included a matched *control group* randomly selected from the Swedish Medical Birth Registry. Visual and hearing impairment and the development of cerebral palsy were evaluated by pediatric ophthalmologists and pediatric neurologists, respectively. Because the follow-up assessments were performed as part of clinical care, outcome assessors were not blinded to the fact that the patient was born extremely prematurely. Although knowledge of birth status could bias the assessment of future clinical outcomes, you are reassured that many of the assessments use standardized, objective criteria.

## WHAT ARE THE RESULTS?

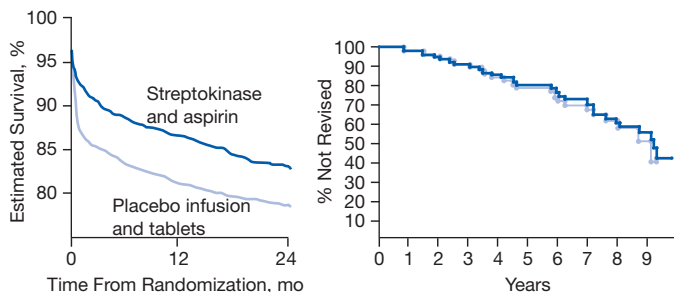
---

### How Likely Are the Outcomes Over Time?

Results from studies of prognosis or risk are often reported as the proportion or percentage of patients with a certain outcome (eg, death, inability to walk, dependence on dialysis) after a certain period of time elapses (eg, 28 days, 3 months, 12 months, 5 years). A more informative way to depict these results is a *survival curve*, which is a graph of the number of events over time (or conversely, the chance of being free of these events over time) (see Chapter 8, Does Treatment Lower Risk? Understanding the Results). The events must be categorized as dichotomous variables, yes or no (eg, death, stroke, recurrence of cancer), and investigators must know the time at which the events occur. [Figure 13-1](#) shows 2 survival curves: one of survival after a myocardial infarction<sup>7</sup> and the other of the need

FIGURE 13-1

## Survival Curves



Left, Survival after myocardial infarction. Right, Results of hip replacement surgery: percentage of patients who survived without needing a new procedure (revision) after their initial hip replacement.

Reproduced from *The Lancet*,<sup>7</sup> Copyright ©1988, with permission from Elsevier (left). Reproduced from Dorey and Amstutz,<sup>8</sup> with permission from the *Journal of Bone and Joint Surgery* (right).

for revision surgery after hip replacement surgery.<sup>8</sup> The chance of dying after a myocardial infarction is highest shortly after the event (reflected by an initially steep downward slope of the curve, which then becomes flat), whereas few hip replacements require revision until much later (this curve, by contrast, starts out flat and then steepens).

### How Precise Are the Estimates of Likelihood?

The more precise the estimate of prognosis a study provides, the more useful it is to us. Usually, authors report the risks of adverse outcomes with their associated 95% *confidence intervals* (CIs). If the study is unbiased, the 95% CI defines the range of risks within which it is highly likely that the true risk lies (see Chapter 9, Confidence Intervals: Was the Single Study or Meta-analysis Large Enough?). For example, a study of the

prognosis of patients with dementia provided a 95% CI around the 49% estimate of survival at 5 years after presentation (ie, 39%-58%).<sup>9</sup> In most survival curves, the earlier follow-up periods usually include results from more patients than do the later periods (owing to losses to follow-up and because patients are not enrolled in the study at the same time), which means that the survival curves are usually more precise in the earlier periods, indicated by narrower confidence bands.

The number of patients evaluated and the number of events influence our confidence in the results. Table 13-1 reveals that in the case of extreme results (all or no patients have the outcome), confidence limits remain quite wide until the number of patients included is approximately 40 to 50, and they do

**TABLE 13-1****95% Confidence Limits on Extreme Results<sup>a</sup>**

<b>If the denominator is:</b>	<b>And the % is 0%, the true % could be as high as:</b>	<b>And the % is 100%, the true % could be as low as:</b>
1	95%	5%
2	78%	22%
3	63%	37%
4	53%	46%
5	45%	55%
10	26%	74%
25	11%	89%
50	6%	94%
100	3%	97%
150	2%	98%
300	1%	99%

<sup>a</sup>Adapted from Sackett et al.<sup>10</sup>

not narrow until the numbers reach the hundreds.<sup>10</sup> When the numerator is 0 or 1 and there are at least 30 patients in the sample, a simple equation called “the rule of 3s” can be applied, where  $100 \times 3$  divided by the number of patients estimates the upper limit of the 95% CI.<sup>11</sup>

## USING THE GUIDE

Of the 456 children in the study who survived to 2.5 years of age whose neurodevelopment could be assessed, 42% were classified as healthy, 31% had mild disabilities, and 27% had moderate or severe disabilities.<sup>1</sup> However, the proportion of children with mild or no disabilities increased from 40% at 22 weeks to 83% at 26 weeks. Boys were at higher risk of moderate to severe disabilities (31%) than girls (23%). Figure 13-2 shows the mean Bayley-III composite cognitive, language, and motor scores at 2.5 years of corrected age according to gestational age at birth and compared with the control group. You are pleased to see in Figure 13-2 that the 99% confidence limits for the infants born at 26 completed weeks of pregnancy are fairly tight, in part because the number of patients in that group is the largest ( $n = 148$ ), compared with the very wide CIs in the group of infants born at 23 weeks' gestation ( $n = 37$ ).

## HOW CAN I APPLY THE RESULTS TO PATIENT CARE?

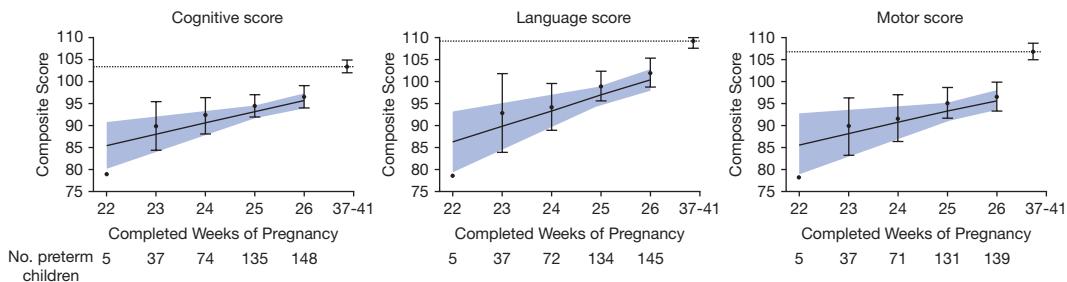
### Were the Study Patients and Their Management Similar to Those in My Practice?

Authors should describe the study patients explicitly and in sufficient detail that you can make a comparison with your



**FIGURE 13-2**

**Mean Bayley Scales of Infant and Toddler Development Composite Cognitive, Language, and Motor Scores at 2.5 Years of Corrected Age for Extremely Preterm Children by Gestational Age at Birth and for the Term Control Group<sup>1</sup>**



The diagonal line indicates the mean of the controls and the vertical bars represent the 99% confidence intervals (CIs) of the mean values. The regression lines with 99% CIs for respective scores of children in the preterm group are based on these equations in which GA indicates gestational age in completed weeks: cognitive score =  $83.12 + (GA - 21) \times 2.517$ ,  $P < .001$ ; language score =  $82.78 + (GA - 21) \times 3.551$ ,  $P < .001$ ; and motor score =  $83.24 + (GA - 21) \times 2.523$ ,  $P = .001$ .

Reproduced from Serenius et al,<sup>1</sup> *JAMA* 2013.



patients. One factor sometimes neglected in prognostic studies that could strongly influence outcome is therapy. Therapeutic strategies often vary markedly among institutions and change over time as new treatments become available or old treatments regain popularity. In fact, investigators studying the cohort of extremely premature infants later reported major differences in 1-year mortality across health care regions in Sweden due to variation in perinatal practices.<sup>12</sup>

### **Was Follow-up Sufficiently Long?**

Because the presence of illness often precedes the development of an outcome event by a long period, investigators must follow up patients for a period long enough to detect the outcomes of interest. For example, recurrence in some women with early breast cancer can occur many years after initial diagnosis and treatment.<sup>13</sup> A prognostic study may provide an unbiased assessment of outcome during a short period if it meets the risk of bias criteria in Box 13-1, but it may be of little use if a patient is interested in prognosis during a long period.

### **Can I Use the Study Results in the Management of Patients in My Practice?**

Prognostic data often provide the basis for sensible decisions about therapy. Even if the prognostic result does not help with selection of the appropriate therapy, it can help you in counseling a concerned patient or relative. Some conditions, such as asymptomatic hiatal hernia or asymptomatic colonic diverticulae, have such a good overall prognosis that they have been termed nondisease.<sup>14</sup> On the other hand, a result of uniformly bad prognosis could provide a clinician with a starting place for a discussion with a patient and family, leading to counseling about end-of-life care.

## CLINICAL SCENARIO RESOLUTION

**T**he active perinatal and follow-up care described in the study of extremely premature infants appears similar to the excellent prenatal care and neonatal intensive care that your patient received. Assuming that the same level of intensive follow-up care is provided, you conclude that the study is likely to provide a good estimate of the prognosis of the child under your care. As a pediatrician and parent, you know that many cognitive issues are not detected until a child begins elementary school, where learning issues sometimes reveal themselves. Although it would be ideal if the cohort were followed up to 6 years of age, you realize that more patients would likely be lost to follow-up. One issue that bothered you is that many of the children died before reaching 2.5 years of age. With further investigation, you are reassured that approximately 84% of those born at 25 to 26 weeks' gestation survived to 2.5 years of age; most of the deaths occurred in those born at 22 to 23 weeks' gestation (approximately 38% survived to 2.5 years). You conclude that you agree with the neonatologist. At this point, this female infant has a nontrivial chance of having moderate to severe disabilities, a substantial chance of having minor disabilities, and even some chance of developing normally without neurocognitive disability.

## References

1. Serenius F, Källén K, Blennow M, et al; EXPRESS Group. Neurodevelopmental outcome in extremely preterm infants at 2.5 years after active perinatal care in Sweden. *JAMA*. 2013;309(17):1810-1820.
2. Sweeting MJ, De Angelis D, Neal KR, et al; Trent HCV Study Group; HCV National Register Steering Group. Estimated progression rates in three United Kingdom hepatitis C cohorts differed according to method of recruitment. *J Clin Epidemiol*. 2006;59(2):144-152.



3. Steyerberg EW, Mushkudiani N, Perel P, et al. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. *PLoS Med.* 2008;5(8):e165, discussion e165.
4. Wolf PA, Dawber TR, Thomas HE Jr, Kannel WB. Epidemiologic assessment of chronic atrial fibrillation and risk of stroke: the Framingham study. *Neurology.* 1978;28(10):973-977.
5. Sims AC. Importance of a high tracing-rate in long-term medical follow-up studies. *Lancet.* 1973;2(7826):433-435.
6. Kriel RL, Krach LE, Jones-Saete C. Outcome of children with prolonged unconsciousness and vegetative states. *Pediatr Neurol.* 1993;9(5):362-368.
7. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. *Lancet.* 1988;2(8607):349-360.
8. Dorey F, Amstutz HC. The validity of survivorship analysis in total joint arthroplasty. *J Bone Joint Surg Am.* 1989;71(4):544-548.
9. Walsh JS, Welch HG, Larson EB. Survival of outpatients with Alzheimer-type dementia. *Ann Intern Med.* 1990;113(6):429-434.
10. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical Medicine.* 2nd ed. Toronto, Ontario: Little Brown & Company; 1991.
11. Hanley JA, Lippman-Hand A. If nothing goes wrong, is everything all right? interpreting zero numerators. *JAMA.* 1983;249(13):1743-1745.
12. Serenius F, Sjörs G, Blennow M, et al; EXPRESS study group. EXPRESS study shows significant regional differences in 1-year outcome of extremely preterm infants in Sweden. *Acta Paediatr.* 2014;103(1):27-37.
13. Early Breast Cancer Trialists' Collaborative Group. Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy. 133 randomised trials involving 31,000 recurrences and 24,000 deaths among 75,000 women. *Lancet.* 1992;339(8784):1-15.
14. Meador CK. The art and science of nondisease. *N Engl J Med.* 1965; 272:92-95.

# The Process of a Systematic Review and Meta-analysis

M. Hassan Murad, Roman Jaeschke, PJ Devereaux, Kameshwar Prasad, Alonso Carrasco-Labra, Thomas Agoritsas, Deborah J. Cook, and Gordon Guyatt

## IN THIS CHAPTER

---

### Clinical Scenario

Should Patients Undergoing Noncardiac Surgery Receive  $\beta$ -Blockers?

### Finding the Evidence

#### Systematic Reviews and Meta-analysis: An Introduction

Definitions

Why Seek Systematic Reviews?

A Synopsis of the Process of a Systematic Review and Meta-analysis

Judging the Credibility of the Effect Estimates

(continued on following page)

### **Was the Process Credible?**

Did the Review Explicitly Address a Sensible Clinical Question?

Was the Search for Relevant Studies Detailed and Exhaustive?

Was the Risk of Bias of the Primary Studies Assessed?

Did the Review Address Possible Explanations of Between-Study Differences in Results?

Did the Review Present Results That Are Ready for Clinical Application?

Were Selection and Assessments of Studies Reproducible?

Did the Review Address Confidence in Effect Estimates?

### **Clinical Scenario Resolution**

---

## CLINICAL SCENARIO

### Should Patients Undergoing Noncardiac Surgery Receive $\beta$ -Blockers?

**Y**ou receive a request for consultation from a general surgeon regarding the perioperative management of a 66-year-old man undergoing hip replacement surgery in 2 days. The patient has a history of type 2 diabetes and hypertension and is a smoker. He has no history of heart disease. The patient's blood pressure is 135/80 mm Hg. Because the patient has multiple *risk factors* for heart disease, you are considering whether he should be treated perioperatively with  $\beta$ -blockers to reduce the risk of death, non-fatal myocardial infarction, and other vascular complications.

## FINDING THE EVIDENCE

Being aware that a large amount of literature exists on this controversial topic, you decide to conduct a search that will provide you with an accurate and rapid overview of current best *evidence*. Because the question is about therapy, you are particularly interested in finding a recent *systematic review* and *meta-analysis of randomized clinical trials* (RCTs) that deal with this topic. Using the free *federated search engine* ACCESSSS (<http://plus.mcmaster.ca/accesssss>; see Chapter 4, Finding Current Best Evidence), you enter these search terms: beta blockers, perioperative, and mortality.

Starting with the summaries at the top of your search output, you locate 2 relevant preappraised summaries on the “management of cardiac risk for noncardiac surgery.” Both summaries cite the results of a large systematic review

and meta-analysis published in 2008,<sup>1</sup> along with references to current US and European *clinical practice guidelines*. However, you notice that the last updates of these chapters date back 4 to 6 months ago. You therefore look further down in your search output to check preappraised research (see Chapter 4, Finding Current Best Evidence) and rapidly identify a more recently published systematic review and meta-analysis addressing your question and that was highly rated for relevance and newsworthiness by clinicians from 4 specialties.<sup>2</sup> You download the full text of the article reporting this meta-analysis.

## SYSTEMATIC REVIEWS AND META-ANALYSIS: AN INTRODUCTION

---

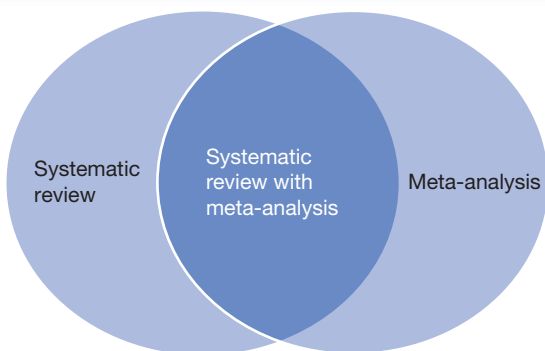
### Definitions

A systematic review is a summary of research that addresses a focused clinical question in a systematic, reproducible manner. Systematic reviews can provide estimates of therapeutic efficacy, *prognosis*, and diagnostic test accuracy and can summarize the evidence for questions of “how” and “why” addressed by *qualitative research* studies. Although we will refer to other sorts of questions, this chapter focuses on systematic reviews that address the effect of therapeutic interventions or harmful *exposures* on *patient-important outcomes*.

A systematic review is often accompanied by a meta-analysis (a statistical pooling or aggregation of results from different studies) to provide a single best estimate of effect. The pooling of studies increases precision (ie, narrows the *confidence intervals* [CIs]), and the single best effect estimate generated

FIGURE 14-1

## The Overlap of Study Designs: Systematic Review and Meta-analysis



facilitates clinical decision making. Therefore, you may see a published systematic review in which the authors chose not to do a meta-analysis, and you may see a meta-analysis conducted without a systematic review (ie, studies were combined statistically but were not selected following a comprehensive, explicit, and reproducible approach) (Figure 14-1). Most useful clinically will be a well-performed systematic review—the methods for which we describe in this chapter—with an accompanying meta-analysis.

In contrast to systematic reviews, traditional *narrative reviews* typically address multiple aspects of the disease (eg, etiology, diagnosis, prognosis, or management), have no explicit criteria for selecting the included studies, do not include systematic assessments of the *risk of bias* associated with *primary studies*, and do not provide quantitative best estimates or rate the confidence in these estimates. The traditional narrative review articles are useful for obtaining a broad overview of a clinical condition but may not provide a reliable and unbiased answer to a focused clinical question.

## Why Seek Systematic Reviews?

When searching for evidence to answer a clinical question, it is preferable to seek a systematic review, especially one that includes a meta-analysis, rather than looking for the best individual study or studies. The reasons include the following:

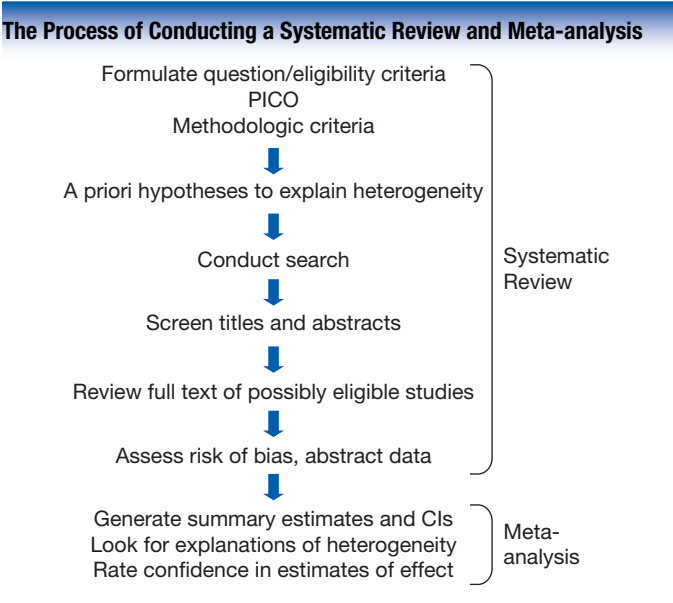
1. Single studies are liable to be unrepresentative of the total body of evidence, and their results may therefore be misleading.
2. Collecting and appraising a number of studies take time you probably do not have.
3. A systematic review is often accompanied by a meta-analysis to provide the best estimate of effect that increases precision and facilitates clinical decision making.
4. If the systematic review is performed well, it will likely provide all of the relevant evidence with an assessment of the best estimates of effect and the confidence they warrant.
5. Systematic reviews include a greater range of patients than any single study, potentially enhancing your confidence in applying the results to the patient before you.

## A Synopsis of the Process of a Systematic Review and Meta-analysis

In applying the *Users' Guides*, you will find it useful to have a clear understanding of the process of conducting a systematic review and meta-analysis. [Figure 14-2](#) shows how the process begins with the definition of the question, which is synonymous with specifying eligibility criteria for deciding which studies to include in a review. These criteria define the population, the exposures or interventions, and the outcomes of interest. A systematic review also may restrict studies to those that minimize the risk of bias. For example, systematic reviews that address a question of therapy often will include only RCTs.

Having specified their selection criteria, reviewers will conduct a comprehensive search of the literature in all relevant

**FIGURE 14-2**



In a systematic review without meta-analysis, the step of generating summary estimates and confidence intervals is not applicable. If the systematic review includes a meta-analysis and presents estimates of effect from individual studies, seeking explanation for heterogeneity and rating confidence in estimates is possible.

Abbreviations: CI, confidence interval; PICO, Patient, Intervention, Comparison, Outcome.

medical databases, which typically yields a large number of potentially relevant titles and abstracts. They then apply the selection criteria to the titles and abstracts, arriving at a smaller number of articles that they retrieve. Once again, the reviewers apply the selection criteria, this time to the complete reports.

Having completed the culling process, the reviewers assess the risk of bias of the individual studies and abstract data from each study. Finally, they summarize the results, including, if appropriate, a quantitative synthesis or meta-analysis. The



meta-analysis provides *pooled estimates* (ie, combined estimates) of the effect on each of the outcomes of interest, along with the associated CIs. Meta-analyses frequently include an examination of the differences in effect estimates across included studies in an attempt to explain differences in results (exploring *heterogeneity*). If based on previously specified hypotheses about possible differences in patients, interventions, or outcomes that may explain differences in results, such explorations become more credible.

### Judging the Credibility of the Effect Estimates

When applying the results of a systematic review to patient care, you can look for estimates of effect. A systematic review without a meta-analysis typically presents results from individual studies; the meta-analysis adds a single pooled (combined) estimate of effect, with an associated CI, for each relevant outcome. Pooled estimates could be for therapy outcomes (eg, death, myocardial infarction, quality of life, late catastrophic adverse effects), estimates of the properties of diagnostic tests (eg, *likelihood ratios*), or estimates of patients' likely outcomes (eg, prognosis). Clinicians need to know the extent to which they can trust these estimates.

Two fundamental problems can undermine this trust. One is the extent to which systematic review authors have applied rigorous methods in conducting their review. We refer to this as the *credibility* of the review.<sup>3</sup> By credibility, we mean the extent to which the design and conduct of the review are likely to have protected against misleading results.<sup>4</sup> As you will see, credibility may be undermined by eligibility criteria that are inappropriate or not specified, the conduct of an inadequate search, and the omission of risk of bias assessments of individual studies (see [Box 14-1](#) for issues to be considered in the credibility of the review process; these issues are applicable to any systematic review, with or without a meta-analysis).

A highly credible review—one that has adhered to methodologic standards—may nevertheless leave us with only very low confidence in estimates of effect. Common reasons for this

**BOX 14-1****Users' Guides for Credibility of the Systematic Review Process**

Did the review explicitly address a sensible clinical question?

Was the search for relevant studies exhaustive?

Was the risk of bias of the primary studies assessed?

Did the review address possible explanations of between-study differences in results?

Did the review present results that are ready for clinical application?

Were selection and assessments of studies reproducible?

Did the review address confidence in effect estimates?

include the following: the individual studies may be plagued by high risk of bias and inconsistent results, even the pooled (combined) sample sizes may be small and the results may be imprecise, and the patients enrolled in the studies may differ in important ways from those in whom we are interested. This chapter deals with credibility assessment of the review process; the next chapter (Chapter 15, Understanding and Applying the Results of a Systematic Review and Meta-analysis) will guide you in deciding how much confidence we can place on estimates of effect in the presence of a credible review process.

**WAS THE PROCESS CREDIBLE?****Did the Review Explicitly Address a Sensible Clinical Question?**

A systematic review has, relative to a traditional narrative review, a narrow focus and addresses a specific question that—for questions of therapy or harm—is defined by particular patients,

interventions, comparisons, and outcomes. When review authors conduct a meta-analysis, the issue of how narrow or wide is the scope of the question becomes particularly important. Let us look at these hypothetical examples of 4 meta-analyses with varying scope:

1. A meta-analysis that pooled results from all modalities of cancer therapy for all types of cancer to generate a single estimate of the effect on mortality.
2. A meta-analysis that pooled the results of the effect of all doses of all antiplatelet agents (including aspirin, sulfapyrazone, dipyridamole, ticlopidine, and clopidogrel) on major thrombotic events (including myocardial infarctions, strokes, and acute arterial insufficiency in the lower extremities).
3. A meta-analysis that pooled the results of the effect of all doses of all antiplatelet agents on mortality in patients with clinically manifest atherosclerosis (whether in the heart, brain, or lower extremities).
4. A meta-analysis that pooled the results of the effect of a wide range of aspirin doses to *prevent* thrombotic stroke in patients presenting with a transient ischemic attack (TIA) due to carotid artery disease.

Clinicians will clearly be uncomfortable with the first meta-analysis, which addresses all treatments for all cancers. Clinicians are unlikely to find the second and third meta-analyses on antiplatelet agents in major thrombotic events and mortality useful because they remain too broad. In contrast, most clinicians may be comfortable with the fourth, more focused meta-analysis of aspirin and thrombotic stroke, although they may express concerns about pooling across a wide range of aspirin doses.

What makes a meta-analysis too broad or too narrow? When deciding whether the question posed in the meta-analysis is sensible, clinicians need to ask themselves whether the underlying biology is such that they would anticipate more or less the same

**BOX 14-2****Were Eligibility Criteria for Inclusion in the Systematic Review Appropriate?**

Are results likely to be similar across the range of included patients (eg, older and younger, sicker and less sick)?

Are results likely to be similar across the range of studied interventions or exposures (eg, for therapy, higher dose or lower dose; for diagnosis, test results interpreted by experts or nonexperts)?

Are results likely to be similar across the range of ways the outcome was measured (eg, shorter or longer follow-up)?

*treatment effect* across the range of patients included (Box 14-2). They should ask a parallel question about the other components of the study question: Is the underlying biology such that, across the range of interventions and outcomes studied, they expect more or less the same treatment effect? Clinicians also can construct a similar set of questions for other areas of clinical inquiry. For example, across the range of patients, ways of testing, and *reference* or *gold standard* for diagnosis, does one expect more or less the same likelihood ratios associated with studies that examine a diagnostic test<sup>5</sup> (see Chapter 12, Diagnostic Tests)?

Clinicians reject a meta-analysis that pools data across all modes of cancer therapy for all types of cancer because they know that some cancer treatments are effective in certain cancers, whereas others are not effective. Combining the results of these studies would yield an estimate of effect that would make little sense or be misleading for most of the interventions. Clinicians who reject the meta-analysis on all antiplatelet agents and mortality in patients with atherosclerosis would argue that the biologic variation in antiplatelet agents is likely to lead to important differences in treatment effect. Furthermore, they may contend that there are important differences in the biology

of atherosclerosis in the vessels of the heart, brain and neck, and legs. Those who would endorse this meta-analysis would argue for the similar underlying biology of antiplatelet agents—and atherosclerosis in different parts of the body—and thus anticipate a similar magnitude of treatment effects.

For the last, more focused review, most clinicians would accept that the biology of aspirin action is likely to be similar in patients whose TIA reflected right-sided or left-sided brain ischemia, in patients older than 75 years and in younger patients, in men and women, across different aspirin doses, during periods of *follow-up* ranging from 1 to 5 years, and in patients with stroke who have been identified by the attending physician and those identified by a team of experts. The similar biology is likely to result in a similar magnitude of treatment effect, which explains the comfort of the meta-analysis authors with combining studies of aspirin in patients who have had a TIA.

The clinician's task is to decide whether, across the range of patients, interventions or exposures, and outcomes, it is plausible that the intervention will have a similar effect. This judgment is possible only if the review authors have provided a precise statement of what range of patients, exposures, and outcomes they decided to include; in other words, the explicit eligibility criteria for their review.

In addition, systematic review authors must specify the criteria for study inclusion related to the risk of bias. Generally, these should be similar to the most important criteria used to evaluate the risk of bias in primary studies<sup>6</sup> (Box 14-3). Explicit eligibility criteria not only facilitate the decision regarding whether the question was sensible but also make it less likely that the authors will preferentially include or exclude studies that support their own previous conclusions or beliefs.

Clinicians may legitimately ask, even within a relatively narrowly defined question, whether they can be confident that results will be similar across patients, interventions, and outcome measurement. Referring to the question of aspirin use by patients with a TIA, the effect could conceivably differ in those

**BOX 14-3****Guides for Selecting Articles That Are Most Likely to Provide Results at Lower Risk of Bias**

Therapy	Were patients randomized?
	Was follow-up complete?
Diagnosis	Was the patient sample representative of those with the disorder?
	Was the diagnosis verified using credible criteria that were independent of the items of medical history, physical examination, laboratory tests, or imaging procedures under study?
Harm	Did the investigators find similarity in all known determinants of outcome or adjust for differences in the analysis?
	Was follow-up sufficiently complete?
Prognosis	Was there a representative sample of patients?
	Was follow-up sufficiently complete?

with more or less severe underlying atherosclerosis, across aspirin doses, or during short-term and long-term follow-up. Thus, at the time of examining the results, we need to ask whether the assumption with which we started proved accurate: was the effect the same across patients, interventions, and outcomes? We return to this issue in the next chapter (see Chapter 15, Understanding and Applying the Results of a Systematic Review and Meta-analysis).

### **Was the Search for Relevant Studies Detailed and Exhaustive?**

Systematic reviews are at risk of presenting misleading results if they fail to secure a complete, or at least representative, sample

of the available eligible studies. To achieve this objective, reviewers search bibliographic databases. For most clinical questions, searching a single database is insufficient and can lead to missing important studies. Therefore, searching MEDLINE, EMBASE, and the Cochrane Central Register of Controlled Trials is recommended for most clinical questions.<sup>7</sup> Searching other databases may be required, depending on the nature of the review question. The systematic review authors check the reference lists of the articles they retrieve and seek personal contact with experts in the area. It also may be important to examine recently published abstracts presented at scientific meetings and to look at less frequently used databases, including those that summarize doctoral theses and databases of ongoing trials held by pharmaceutical companies or databases of ongoing registered trials.

Another important source of unpublished studies is the US Food and Drug Administration (FDA) reviews of new drug applications. A study that evaluated the risk of dyspepsia associated with the use of nonsteroidal anti-inflammatory drugs found that searching FDA records yielded 11 trials, of which only 1 was published.<sup>8</sup> Another study of FDA reports found that they included numerous unpublished studies, and the findings of these studies can appreciably alter the estimates of effect.<sup>9</sup> Unless the authors of systematic reviews tell us what they did to locate the studies, it is difficult to know how likely it is that relevant studies were missed.

*Reporting bias* occurs in a number of forms, the most familiar of which is the failure to report or publish studies with negative results. This *publication bias* may result in misleading results of systematic reviews that fail to include unpublished studies.<sup>10,11</sup>

If authors include unpublished studies in a review, they should try to obtain full reports, and they should use the same criteria to appraise the risk of bias of both published and unpublished studies. There is a variety of techniques available to explore the possibility of publication bias, but none of them are fully satisfactory. Systematic reviews based on a small number

of studies with limited total sample sizes are particularly susceptible to publication bias, especially if most or all of the studies have been sponsored by a commercial entity with a vested interest in the results.

Another increasingly recognized form of reporting bias occurs when investigators measure a number of outcomes but report only those that favor the *experimental intervention* or those that favor the intervention most strongly (*selective outcome reporting bias*). If reviewers report that they have successfully contacted authors of primary studies and were assured of the full disclosure of results, concern about reporting bias decreases.

Reviewers may go even farther than simply contacting the authors of primary studies. They may recruit these investigators as collaborators in their review, and in the process, they may obtain individual patient records. Such *individual patient data meta-analysis* can facilitate powerful analyses (addressing issues such as true *intention-to-treat* analyses and informed *subgroup analyses*), which may strengthen the inferences from a systematic review.

### **Was the Risk of Bias of the Primary Studies Assessed?**

Even if a systematic review includes only RCTs, knowing the extent to which each individual trial used safeguards against bias is important. Differences in study methods might explain important differences among the results.<sup>12</sup> For example, less rigorous studies sometimes overestimate the effectiveness of therapeutic and preventive interventions.<sup>13</sup> Even if the results of different studies are consistent, determining their risk of bias is still important. Consistent results are less compelling if they come from studies with a high risk of bias than if they come from studies with a low risk of bias.

Consistent results from *observational studies* putatively addressing treatment issues also should raise concern. Clinicians may systematically select patients with a good prognosis to receive therapy, and this pattern of practice may be consistent over time and geographic setting. There are many examples



of observational studies that found misleading results subsequently contradicted by large RCTs. For example, considerable preclinical and epidemiologic evidence suggested that antioxidant vitamins reduced the risk of prostate cancer. However, a trial of 35 533 healthy men found that dietary supplementation with vitamin E significantly increased the risk of prostate cancer.<sup>14</sup> Similarly, laboratory experiments suggested that antioxidants may slow or prevent atherosclerotic plaque formation, but a trial of 14 641 male physicians found that neither vitamin E nor vitamin C supplementation reduced the risk of major cardiovascular events.<sup>15</sup>

There is no one correct way to assess the risk of bias.<sup>16</sup> Some reviewers use long checklists to evaluate risk of bias, whereas others focus on 3 or 4 key aspects of the study. When considering whether to trust the results of a review, check to see whether the authors examined criteria similar to those we have presented in other chapters of this book (see Chapter 6, Therapy [Randomized Trials]; Chapter 10, Harm [Observational Studies]; Chapter 12, Diagnostic Tests; and Chapter 13, Prognosis). Reviewers should apply these criteria with a relatively low threshold (such as restricting eligibility to RCTs) in selecting studies (Box 14-3) and more comprehensively (such as considering *concealment*, *blinding*, and *stopping early* for benefit) in assessing the risk of bias of the included studies. The authors of systematic reviews should explicitly report the extent of the risk of bias of each included study in their review.

### **Did the Review Address Possible Explanations of Between-Study Differences in Results?**

Studies included in a systematic review are unlikely to show identical results. Whether or not their review includes a meta-analysis, systematic review authors should attempt to explain the reasons for variability in results. When the studies are combined in a meta-analysis, the difference in results becomes easily quantifiable. Chance always represents a possible explanation.

Alternatively, differences in the characteristics of the patients enrolled, in the way the intervention was administered, in the way the outcome was assessed, or in the risk of bias may be responsible. For example, the intervention may be more effective in older patients than in younger patients or in those with diabetes than in those without diabetes. We often refer to inconsistency in results among studies as heterogeneity.

Systematic review authors should hypothesize possible explanations for heterogeneity (a priori, when they plan the review) and test their hypotheses in a subgroup analysis. Subgroup analyses may provide important insights, but they also may be misleading. In Chapter 15, *Understanding and Applying the Results of a Systematic Review and Meta-analysis*, we discuss how to evaluate heterogeneity and how it affects the confidence in estimates.

### Did the Review Present Results That Are Ready for Clinical Application?

If you and your patients are told that treatment lowers the risk of myocardial infarction by 50%, it sounds impressive, but that could mean a reduction from 1% to 0.5% or from 40% to 20%. In the former situation, when the *risk difference* (also referred to as *absolute risk reduction*) is 0.5%, your patient may decide to decline a treatment with appreciable adverse effect, *burden*, or cost. In the latter situation, that is much less likely to be the case. Therefore, you and your patients need to know the absolute effect of the intervention. The absolute benefit (or harm) that patients will achieve with therapy depends on their *baseline risk* (the likelihood of the outcome when receiving no or standard therapy).

For example, statins reduce fatal and nonfatal cardiovascular events<sup>17</sup> by approximately 25% (*relative risk* [RR], 0.75); the absolute benefit, however, may be greater for a patient with an elevated Framingham risk score (or other risk stratification method) than for a patient with a low score (Box 14-4).

## BOX 14-4

### The Impact of Baseline Risk on the Magnitude of Absolute Risk Reduction

Patient 1	Patient 2
65-year-old male smoker with cholesterol level of 250 mg/dL, high-density lipoprotein (HDL) of 30 mg/dL, and systolic blood pressure of 140 mm Hg	50-year-old female smoker with cholesterol of 170 mg/dL, HDL of 55 mg/dL, and systolic blood pressure of 130 mm Hg
Absolute risk of having a cardiac event during the next 10 years: 28%	Absolute risk of having a cardiac event during the next 10 years: 2%
Risk after treatment with statin: $28\% \times 0.75 = 21\%$	Risk after treatment with statin: $2\% \times 0.75 = 1.5\%$
Absolute risk reduction: $28\% - 21\% = 7\%$	Absolute risk reduction: $2\% - 1.5\% = 0.5\%$

Although we are primarily interested in absolute effects, relative effects tend to be much more consistent across studies (see Chapter 8, Does Treatment Lower Risk? Understanding the Results). That is the reason that meta-analyses of binary outcomes usually should and do combine and present relative effects, such as the relative risk, *odds ratio*, or occasionally *hazard ratio*. So how, then, do we determine the absolute effects in which we are really interested? The best way is to obtain an estimate of the patients' baseline risk (ideally from an observational study of a representative population, from a risk-stratification instrument, or, if neither is available, from the randomized trials in the meta-analysis)<sup>18</sup> and then use the relative risk<sup>19</sup> to estimate that patient's risk difference.

Review authors also can present outcomes that are *continuous variables* in ways that are more or less useful and applicable. For

instance, the weighted mean difference and standardized mean difference represent common statistical approaches for pooling across studies. Clinicians, however, may have difficulty grasping the significance of the effect of a respiratory rehabilitation program presented as a weighted mean difference of 0.71 units on the Chronic Respiratory Questionnaire (CRQ) scale. They may have less difficulty if told that the *minimal important difference* on the CRQ is 0.5 units. Clinicians are likely to have at least equal difficulty if told that the treatment effect on disease-specific *health-related quality of life* is a standardized mean difference of 0.71. Again, they may have less difficulty if told that 0.2, 0.5, and 0.8 may represent small, moderate, and large effects. Clinicians are likely to have the least amount of difficulty if told that 30% of patients have an important improvement in function as a result of the program (a *number needed to treat* of approximately 3).<sup>20</sup>

### **Were Selection and Assessments of Studies Reproducible?**

As we have seen, authors of systematic reviews must decide which studies to include, the extent of risk of bias, and what data to abstract. These decisions always require judgment by the reviewers and are subject to both mistakes (ie, *random errors*) and bias (ie, *systematic errors*). Having 2 or more people participate in each decision guards against errors, and if there is good agreement beyond chance among the reviewers, the clinician can have more confidence in the results of the systematic review. Systematic reviewers often report a measure of agreement (eg, a measure of *chance-corrected agreement* such as the  $\kappa$  statistic) to quantify their level of agreement on study selection and appraisal of the risk of bias.

### **Did the Review Address Confidence in Effect Estimates?**

As we have pointed out, a review can follow optimal systematic review and meta-analytic methods, and the evidence may still

warrant low confidence in estimates of effect. Ideally, systematic review authors will explicitly address the risk of bias that can diminish confidence in estimates as well as imprecision (ie, wide CIs) and inconsistency (ie, large variability in results from study to study). If systematic review authors do not make explicit assessments themselves, they should at least provide the information you need to make your own assessment. The next chapter (Chapter 15, Understanding and Applying the Results of a Systematic Review and Meta-analysis) describes in detail how the systematic review authors—or you, in the absence of the authors doing so explicitly—can address these issues to make an appropriate rating of confidence in estimates of effect.

## CLINICAL SCENARIO RESOLUTION

**R**eturning to our opening scenario, the systematic review and meta-analysis you located included 11 trials that enrolled more than 10 000 patients who were having noncardiac surgery and were randomized to either  $\beta$ -blockers or a *control group*.<sup>2</sup> The trials addressed the main outcomes of interest (death, nonfatal myocardial infarction, and nonfatal stroke). The  $\beta$ -blocker, dose, timing, and duration of administration all varied across the trials.

The systematic review authors had searched MEDLINE, EMBASE, CINAHL, the Cochrane Library Central Register of Randomised Controlled Trials, and other trial databases and registries. They also checked the reference lists of identified articles and previous systematic reviews for additional references. They did not restrict the search to a particular language or location. They had 2 independent reviewers assess trial eligibility and select studies, and disagreements were resolved by a third review author. They did not quantitatively report the agreement level among reviewers, a feature you would have preferred to know.

The systematic review authors used the *Cochrane Collaboration* risk of bias assessment methods. They explicitly described

the risk of bias of each trial by reporting on the adequacy of generation of the allocation sequence, allocation concealment, and blinding of participants, personnel, and outcome assessors. As part of the meta-analysis, the authors conducted a separate *sensitivity analysis* that excluded the studies with a higher risk of bias. They tested for publication bias. They did not report that they had contacted authors of primary studies, which you would have preferred they did.

Overall, you conclude that the credibility of the methods of this systematic review and meta-analysis is moderate to high, and you decide to examine the estimates of effect and the associated confidence in these estimates.

## References

1. Bangalore S, Wetterslev J, Pranesh S, Sawhney S, Gluud C, Messerli FH. Perioperative beta blockers in patients having non-cardiac surgery: a meta-analysis. *Lancet*. 2008;372(9654):1962-1976.
2. Bouri S, Shun-Shin MJ, Cole GD, Mayet J, Francis DP. Meta-analysis of secure randomised controlled trials of  $\beta$ -blockade to prevent perioperative death in non-cardiac surgery. *Heart*. 2014;100(6):456-464.
3. Alkin M. *Evaluation Roots: Tracing Theorists' Views and Influences*. Thousand Oaks, CA: Sage Publications Inc; 2004.
4. Oxman AD. Checklists for review articles. *BMJ*. 1994;309(6955):648-651.
5. Irwig L, Tosteson AN, Gatsonis C, et al. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med*. 1994;120(8):667-676.
6. Oxman AD, Guyatt GH. The science of reviewing research. *Ann N Y Acad Sci*. 1993;703:125-134.
7. The Cochrane Collaboration. *Cochrane Handbook for Systematic Reviews of Interventions*. Version 5.1.0. <http://handbook.cochrane.org/>. Accessed July 26, 2014.
8. MacLean CH, Morton SC, Ofman JJ, Roth EA, Shekelle PG. How useful are unpublished data from the Food and Drug Administration in meta-analysis? *J Clin Epidemiol*. 2003;56(1):44-51.
9. McDonagh MS, Peterson K, Balshem H, Helfand M. US Food and Drug Administration documents can provide unpublished evidence relevant to systematic reviews. *J Clin Epidemiol*. 2013;66(10):1071-1081.

10. Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ*. 1997;315(7109):640-645.
11. Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *JAMA*. 1998;279(4):281-286.
12. Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet*. 1998;352(9128):609-613.
13. Odgaard-Jensen J, Vist GE, Timmer A, et al. Randomisation to protect against selection bias in healthcare trials. *Cochrane Database Syst Rev*. 2011;(4):MR000012.
14. Klein EA, Thompson IM Jr, Tangen CM, et al. Vitamin E and the risk of prostate cancer: the Selenium and Vitamin E Cancer Prevention Trial (SELECT). *JAMA*. 2011;306(14):1549-1556.
15. Sesso HD, Buring JE, Christen WG, et al. Vitamins E and C in the prevention of cardiovascular disease in men: the Physicians' Health Study II randomized controlled trial. *JAMA*. 2008;300(18):2123-2133.
16. Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA*. 1999;282(11):1054-1060.
17. Taylor F, Huffman MD, Macedo AF, et al. Statins for the primary prevention of cardiovascular disease. *Cochrane Database Syst Rev*. 2013;1:CD004816.
18. Guyatt GH, Eikelboom JW, Gould MK, et al; American College of Chest Physicians. Approach to outcome measurement in the prevention of thrombosis in surgical and medical patients: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest*. 2012;141(2)(suppl):e185S-e194S.
19. Murad MH, Montori VM, Walter SD, Guyatt GH. Estimating risk difference from relative association measures in meta-analysis can infrequently pose interpretational challenges. *J Clin Epidemiol*. 2009;62(8):865-867.
20. Thorlund K, Walter SD, Johnston BC, Furukawa TA, Guyatt GH. Pooling health-related quality of life outcomes in meta-analysis—a tutorial and review of methods for enhancing interpretability. *Res Synth Methods*. 2012;2(3):188-203.

# Understanding and Applying the Results of a Systematic Review and Meta-analysis

M. Hassan Murad, Victor M. Montori, John P. A. Ioannidis, Ignacio Neumann, Rose Hatala, Maureen O. Meade, PJ Devereaux, Peter Wyer, and Gordon Guyatt

## IN THIS CHAPTER

---

### Clinical Scenario

#### Understanding the Summary Estimate of a Meta-analysis

Understanding the Estimate of Absolute Effect

#### Rating Confidence in the Estimates (The Quality of Evidence)

The GRADE Approach

How Serious Is the Risk of Bias in the Body of Evidence?

Are the Results Consistent Across Studies?

(continued on following page)



How Precise Are the Results?

Do the Results Directly Apply to My Patient?

Is There Concern About Reporting Bias?

Are There Reasons to Increase the Confidence Rating?

**An Evidence-Based Summary of the Findings:**

**The Evidence Profile**

**Clinical Scenario Resolution**

---

In the previous chapter (Chapter 14, The Process of a Systematic Review and Meta-analysis), we provided guidance on how to evaluate the *credibility* of the process of a *systematic review* with or without a *meta-analysis*. In this chapter, we address how—if the systematic review is sufficiently credible—to decide on the degree of confidence in the estimates that the *evidence* warrants. As you will see, systematic review authors may have conducted a credible review and analysis and one may still have little confidence in the estimates of effect. We will return to the clinical scenario discussed in the previous chapter and obtain the relative and absolute effects of the intervention from a credible systematic review and meta-analysis<sup>1</sup> and determine the confidence in these estimates (quality of evidence). The general framework for judging confidence in estimates is based on the approach offered by the *GRADE (Grading of Recommendations Assessment, Development and Evaluation)* Working Group.<sup>2</sup> This chapter focuses on questions of therapy or harm. This framework can, however, be adapted for other types of questions, such as issues of prognosis<sup>3</sup> or diagnosis.<sup>4</sup>

## CLINICAL SCENARIO

**W**e continue with the scenario of a 66-year-old male smoker with type 2 diabetes and hypertension undergoing noncardiac surgery for whom we are considering prescribing perioperative  $\beta$ -blockers to prevent the cardiovascular complications of nonfatal infarction, death, and nonfatal stroke.

## UNDERSTANDING THE SUMMARY ESTIMATE OF A META-ANALYSIS

If the systematic review authors decide that combining results to generate a single estimate of effect is inappropriate, a systematic review will likely end with a table or tables describing results of

individual *primary studies*. Often, however, systematic reviews include a meta-analysis with a best estimate of effect (often called a summary or *pooled estimate*) from the weighted averages of the results of the individual studies. The weighting process depends on sample size or number of events or, more specifically, study precision. Studies that are more precise have narrower *confidence intervals* (CIs) and larger weight in meta-analysis.

In a meta-analysis of a therapeutic question looking at *dichotomous outcomes* (yes/no) for estimates of the magnitude of the benefits or *risks*, you should look for the *relative risk* (RR) and *relative risk reduction* (RRR) or the *odds ratio* (OR) and *relative odds* reduction (see Chapter 8, Does Treatment Lower Risk? Understanding the Results). When the outcome is analyzed using time-to-event methods (eg, *survival analysis*), the results could be presented as a *hazard ratio*. In a meta-analysis addressing diagnosis, you should look for summary estimates of *likelihood ratios* or diagnostic ORs (see Chapter 12, Diagnostic Tests).

In the setting of *continuous variables* rather than dichotomous outcomes, meta-analysts typically use 1 of 2 options to aggregate data across studies. If the outcome is measured the same way in each study (eg, duration of hospitalization), the results from each study are combined, taking into account each study's precision to calculate what is called a *weighted mean difference*. This measure has the same units as the outcomes reported in the individual studies (eg, pooled estimate of reduction in hospital stay with treatment, 1.1 days).

Sometimes the outcome measures used in the primary studies are similar but not identical. For example, one trial might measure *health-related quality of life* using a validated questionnaire (the Chronic Respiratory Questionnaire), and another trial might use a different validated questionnaire (the St. George's Questionnaire). Another example of this situation is a meta-analysis of studies using different measures of severity of depression.

If the patients and the interventions are similar, generating a pooled estimate of the effect of the intervention on quality of

life or depression, even when investigators have used different measurement instruments, is likely to be worthwhile. One way of generating the pooled estimate in this instance is to standardize the measures by looking at the mean difference between treatment and control and dividing this by the SD.<sup>5</sup> The *effect size* that results from this calculation provides a summary estimate of the treatment effect expressed in SD units (eg, an effect size of 0.5 means that the mean effect of treatment across studies is half of an SD unit). A rule of thumb for understanding effect sizes suggests that 0.2 SD represents small effects; 0.5 SD, moderate effects; and 0.8 SD, large effects.<sup>6</sup>

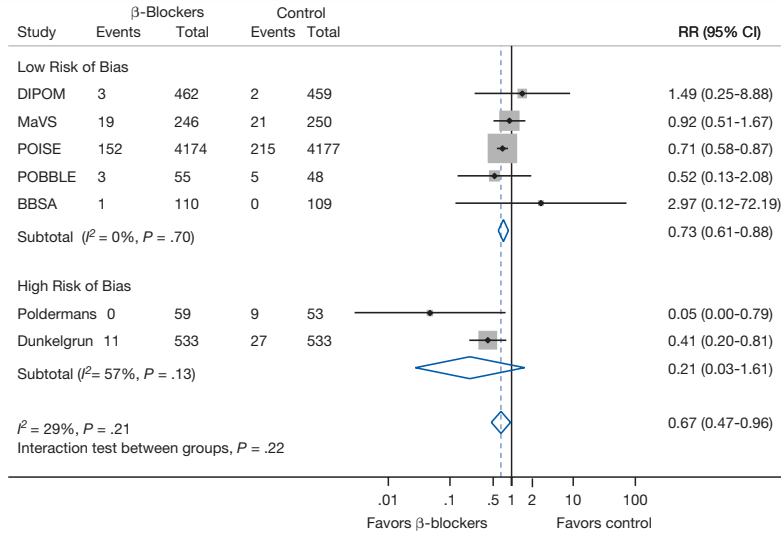
Clinicians may be unfamiliar with how to interpret effect size, and systematic review authors may help you interpret the results by using one of a number of alternative presentations. One is to translate the summary effect size back into natural units.<sup>7</sup> For instance, clinicians may have become familiar with the significance of differences in walk test scores in patients with chronic lung disease. Investigators can then convert the effect size of a treatment on a number of measures of functional status (eg, the walk test and stair climbing) back into differences in walk test scores.<sup>8</sup>

Even better may be the translation of continuous outcomes into dichotomies: the proportion of patients who, for instance, have experienced an important reduction in pain, fatigue, or dyspnea. Methods of making such translations are increasingly well developed.<sup>9,10</sup> For examples on how systematic review authors can present results that are ready for clinical applications, see Chapter 14, The Process of a Systematic Review and Meta-analysis.

The results of a traditional meta-analysis are usually depicted in what is called a *forest plot* (Figures 15-1, 15-2, and 15-3). This forest plot shows the effect (ie, result) from every study; the *point estimate* is presented as a square with a size that is proportional to the weight of the study, and the CI is presented as a horizontal line. The solid line at 1.0 indicates no effect, and the dashed line

**FIGURE 15-1**

**Results of a Meta-analysis of the Outcomes of Nonfatal Infarction in Patients Receiving Perioperative  $\beta$ -Blockers**



Abbreviations: BBSA, Beta Blocker in Spinal Anesthesia study; CI, confidence interval; DIPOM, Diabetic Postoperative Mortality and Morbidity trial; MaVS, Metoprolol after Vascular Surgery study; POBBLE, Perioperative  $\beta$ -blockade trial; POISE, PeriOperative ISchemic Evaluation trial.

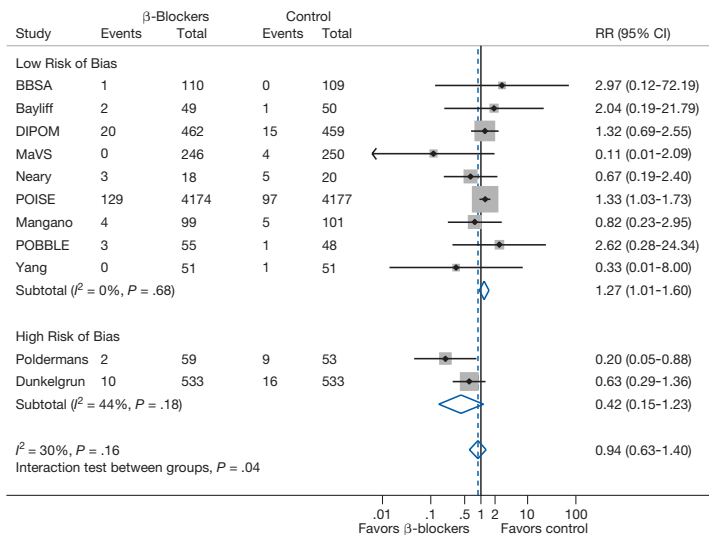
Solid line indicates no effect. Dashed line is centered on meta-analysis pooled estimate.

Data are from Bouri et al.<sup>1</sup>



FIGURE 15-2

### Results of a Meta-analysis of the Outcomes of Death in Patients Receiving Perioperative $\beta$ -Blockers



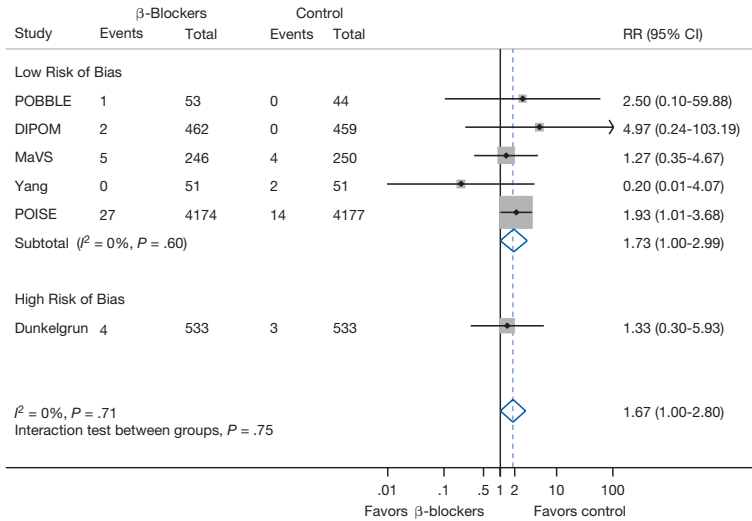
Abbreviations: BBSA, Beta Blocker in Spinal Anesthesia study; CI, confidence interval; DIPOM, Diabetic Postoperative Mortality and Morbidity trial; MaVS, Metoprolol after Vascular Surgery study; POBBLE, Perioperative  $\beta$ -blockade trial; POISE, PeriOperative ISchemic Evaluation trial. Solid line indicates no effect. Dashed line is centered on meta-analysis pooled estimate.

Data from Bouri et al.<sup>1</sup>



**FIGURE 15-3**

**Results of a Meta-analysis of the Outcomes of Nonfatal Stroke in Patients Receiving Perioperative  $\beta$ -Blockers**



Abbreviations: CI, confidence interval; DIPOM, Diabetic Postoperative Mortality and Morbidity trial; MaVS, Metoprolol after Vascular Surgery study; POBBLE, Perioperative  $\beta$ -blockade trial; POISE, PeriOperative ISchemic Evaluation trial.

Solid line indicates no effect. Dashed line is centered on meta-analysis pooled estimate.

Data from Bouri et al.1

is centered on the meta-analysis combined summary effect. The combined summary effect is usually presented as a diamond, with its width representing the CI for the combined effect. As the CI widens, uncertainty about the magnitude of effect increases; when the CI crosses no effect (RR or OR of 1.0), there is uncertainty about whether the intervention has any effect at all (see Chapter 9, Confidence Intervals: Was the Single Study or Meta-analysis Large Enough?).

## USING THE GUIDE

**R**eturning to the perioperative  $\beta$ -blockers scenario, you found a systematic review that you considered as having a credible process that included a meta-analysis for the outcomes of nonfatal infarction, mortality, and nonfatal stroke.<sup>1</sup> The forest plots reveal the estimates of effect for these outcomes from the relevant *randomized trials* (Figure 15-1, 15-2, and 15-3).

Perioperative administration of  $\beta$ -blockers decreases the risk of 1 adverse outcome—nonfatal myocardial infarction (RR, 0.67; 95% CI, 0.47-0.96). The summary effect reached the threshold for *statistical significance* because the CI does not cross 1.0 (no effect) (Figure 15-1). However,  $\beta$ -blockers likely increased the risk of nonfatal stroke, the lower boundary of the CI just touching no effect (RR, 1.67; 95% CI, 1.00-2.80) (Figure 15-2). You are not sure about the effect of  $\beta$ -blockers on the outcome of death because the CI crosses 1.0 and is wide, including a large reduction (37%) and a large increase (40%) in death (RR, 0.94; 95% CI, 0.63-1.40) (Figure 15-3).

You note, however, that there is appreciable *inconsistency* in the results for the *end points* of death and myocardial infarction and that, in particular, the studies with low or high *risk of bias* studies yield different results. This raises the question of which results are more credible, an issue to which we return later in this chapter.



## Understanding the Estimate of Absolute Effect

The goal of a systematic review and meta-analysis is often to present evidence users (clinicians, patients, and policy-makers) with best estimates of the effect of an intervention on each *patient-important outcome*. When interpreting and applying the results, you and your patients must balance the desirable and undesirable consequences to decide on the best course of action.

As we pointed out in the previous chapter (Chapter 14, The Process of a Systematic Review and Meta-analysis), knowledge of the RRs associated with the intervention is insufficient for making a decision about the trade-off between desirable and undesirable consequences; rather, it requires knowledge of the *absolute risk* associated with the intervention. For instance, the relative estimates we have presented so far suggest an RRR of myocardial infarction of 33% with use of  $\beta$ -blockers in non-cardiac surgery but an increase in nonfatal strokes of 67%. The decision about whether to use  $\beta$ -blockers will be different, depending on whether the reduction in myocardial infarction is from 10% to 7% or from 1% to 0.7% and whether the increase in nonfatal strokes is from 0.5% to 0.8% or from 5% to 8%.

However, before we arrive at the best estimates of absolute effect we need to resolve a pending question: does the most trustworthy estimate of relative effect come from all of the studies or does it come from the studies with low risk of bias? We resolve this issue and present the best estimates of absolute effect later in this chapter.

## RATING CONFIDENCE IN THE ESTIMATES (THE QUALITY OF EVIDENCE)

---

Consistent with the second principle of *evidence-based practice*—some evidence is more trustworthy and some less so—application of evidence requires a rating of how confident we

are in our estimates of the magnitude of intervention effects on the outcomes of interest. This confidence rating is important for *clinical practice guideline* developers when they make their recommendations and for clinicians and patients when they decide on their course of action (see Chapter 17, How to Use a Patient Management Recommendation: Clinical Practice Guidelines and Decision Analyses).

The judgment about our confidence in the effect estimates applies not to a single study but rather a body of evidence. For any management decision, confidence in estimates can differ across outcomes. Historically, the word “quality” has been used synonymously with both risk of bias and confidence in estimates. Because of the ambiguity, we avoid the use of the word “quality” (although when we do use it, it is synonymous with confidence). Instead, we use the other 2 terms (risk of bias and confidence in estimates). In this chapter, the focus is on confidence in effect estimates.

## The GRADE Approach

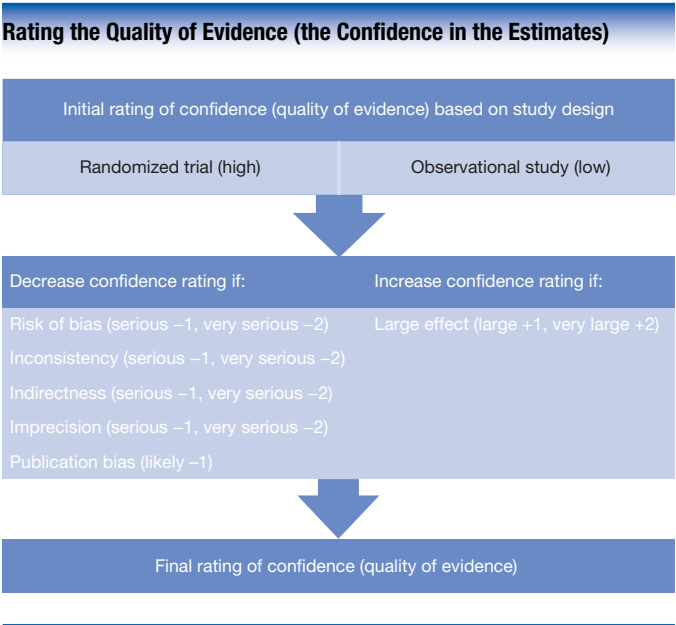
The GRADE approach is one of several systems to rate the quality of evidence. The GRADE Working Group is a group of health care professionals, researchers, and guideline developers who, in 2000, began to work together to develop an optimal system of rating confidence in estimates for systematic reviews and health technology assessments of questions of the impact of interventions and to determine the strength of recommendations for clinical practice guidelines.<sup>2</sup> The GRADE approach has been disseminated widely and endorsed by more than 70 organizations worldwide,<sup>11,12</sup> including the Cochrane Collaboration, the UK National Institutes of Clinical Excellence, the World Health Organization, and the American College of Physicians. Several hundred publications have since described, demonstrated the feasibility and usefulness, evaluated the use of, and provided guidance on the GRADE approach.

GRADE suggests rating confidence in estimates of effect in 4 categories: high, moderate, low, or very low. Some organizations, including UpToDate, combine the low and very low. The

lower the confidence, the more likely the underlying true effect is substantially different from the observed estimate of effect, and thus, it is more likely that further research would reveal a change in the estimates.<sup>13</sup>

Confidence ratings begin by considering study design. Randomized trials are initially assigned high confidence and *observational studies* are given low confidence, but a number of factors may modify these initial ratings (Figure 15-4). Confidence ratings may decrease when there is increased risk of bias, inconsistency, *imprecision*, *indirectness*, or concern about *publication bias*. An increase in confidence rating is uncommon and mainly occurs when the effect size is large (Figure 15-4).

FIGURE 15-4



These factors defined by GRADE should affect our confidence in estimates whether or not systematic review authors formally use GRADE. In one way or another, therefore, your consideration of evidence from a systematic review of alternative management strategies must include consideration of these issues. We now provide a description of how authors of systematic reviews and meta-analyses apply these criteria.

### How Serious Is the Risk of Bias in the Body of Evidence?

Authors of systematic reviews evaluate the risk of bias for each of the outcomes measured in each individual study. *Bias* represents systematic rather than random error (see Chapter 5, Why Study Results Mislead: Bias and Random Error).

For randomized trials, risk of bias increases if there are problems with the *randomization* (defects in generation of the randomization sequence or lack of appropriate *allocation concealment*); if patients, caregivers, and study personnel are not *blinded*; or if a large number of patients are *lost to follow-up* (see Chapter 6, Therapy [Randomized Trials]). The effect of these problems can differ across outcomes. For example, lack of blinding and inadequate allocation concealment lead to greater bias for subjective outcomes than for objective hard clinical outcomes, such as death.<sup>14</sup> Stopping trials early because of a large apparent effect also can exaggerate the treatment effects.<sup>15</sup> In observational studies, the main concerns associated with increased risk of bias include inappropriate measurement of *exposure* and outcome, inadequate statistical adjustment for prognostic imbalance, and loss to follow-up (see Chapter 10, Harm [Observational Studies]).

Ideally, the authors of systematic reviews will present a risk of bias evaluation for every individual study and provide a statement about the overall risk of bias for all of the included studies. The reproducibility of this judgment affects the credibility of the process of the systematic review (see Chapter 14, The Process of a Systematic Review and Meta-analysis). Following the GRADE

approach, the risk of bias can be expressed as “not serious,” “serious,” or “very serious.” The assessment of the level of risk of bias can then result in no decrease in the confidence rating in estimates of effect or a decrease by 1 or 2 levels (eg, from high to moderate or low confidence) (Figure 15-4).<sup>13</sup>

## USING THE GUIDE

**T**he authors of the systematic review and meta-analysis addressing perioperative  $\beta$ -blockers<sup>1</sup> used the Cochrane Collaboration risk of bias assessment methods (see Chapter 14, The Process of a Systematic Review and Meta-analysis). They explicitly described the risk of bias of each trial and reported on the adequacy of generation of the allocation sequence; allocation concealment; blinding of participants, personnel, and outcome assessors; the extent of loss to follow-up; and the use of the *intention-to-treat principle*.

Of the 11 trials included in the analysis, 2 were considered to have high risk of bias<sup>16,17</sup>; limitations included lack of blinding and, in one trial, *stopping early* because of large apparent benefit.<sup>17</sup> The results of these 2 trials became even more questionable when, subsequently, concerns were raised about the integrity of the data.<sup>1</sup> The remaining 9 trials were deemed by the systematic review authors to have adequate bias protection measures and represented a body of evidence that was overall at low risk of bias for the 3 key outcomes—nonfatal myocardial infarction, death, and nonfatal stroke.

### Are the Results Consistent Across Studies?

The starting assumption of a meta-analysis that provides a summary estimate of treatment effect is that across the range of study patients, interventions, and outcomes included in the analysis, the effect of interest is more or less the same (see Chapter 14, The

Process of a Systematic Review and Meta-analysis). On the one hand, a meta-analysis question framed to include a broad range of patients, interventions, and ways of measuring outcome helps avoid spurious effects from *subgroup analyses*, leads to narrower CIs, and increases applicability across a broad range of patients. On the other hand, combining the results of diverse studies may violate the starting assumption of the analysis and lead to spurious conclusions (for instance, that the same estimate of effect applies to different patient groups or different ways of administering an intervention, when it in fact does not).

The solution to this dilemma is to evaluate the extent to which results differ from study to study, that is, the variability or *heterogeneity* of study results. [Box 15-1](#) summarizes 4 approaches to evaluating variability in study results, and the subsequent discussion expands on these principles.<sup>18</sup>

### Visual Assessment of Variability

Studies combined in a meta-analysis and depicted in a forest plot will inevitably have some inconsistency (heterogeneity) of their point estimates. The question is whether that heterogeneity

## BOX 15-1

### Evaluating Variability in Study Results

#### Visual evaluation of variability

How similar are the point estimates?

To what extent do the confidence intervals overlap?

#### Statistical tests evaluating variability

Yes-or-no tests for heterogeneity that generate a *P* value

*I*<sup>2</sup> test that quantifies the variability explained by between-study differences in results

is sufficiently great to make us uncomfortable with combining results from a group of related studies to generate a single summary effect.<sup>19</sup>

Consider the results of the 2 meta-analyses shown in [Figure 15-5A](#) and [B](#) (meta-analysis A and meta-analysis B, respectively). When reviewing the results of these studies, would clinicians be comfortable with a single summary result in either or both meta-analyses? Although the results of meta-analysis A seem extremely unlikely to meet the assumption of a single underlying treatment effect across studies, the results of meta-analysis B are completely consistent with the assumption. Therefore, we would be uncomfortable applying the pooled estimate to all studies in A but comfortable doing so in B.

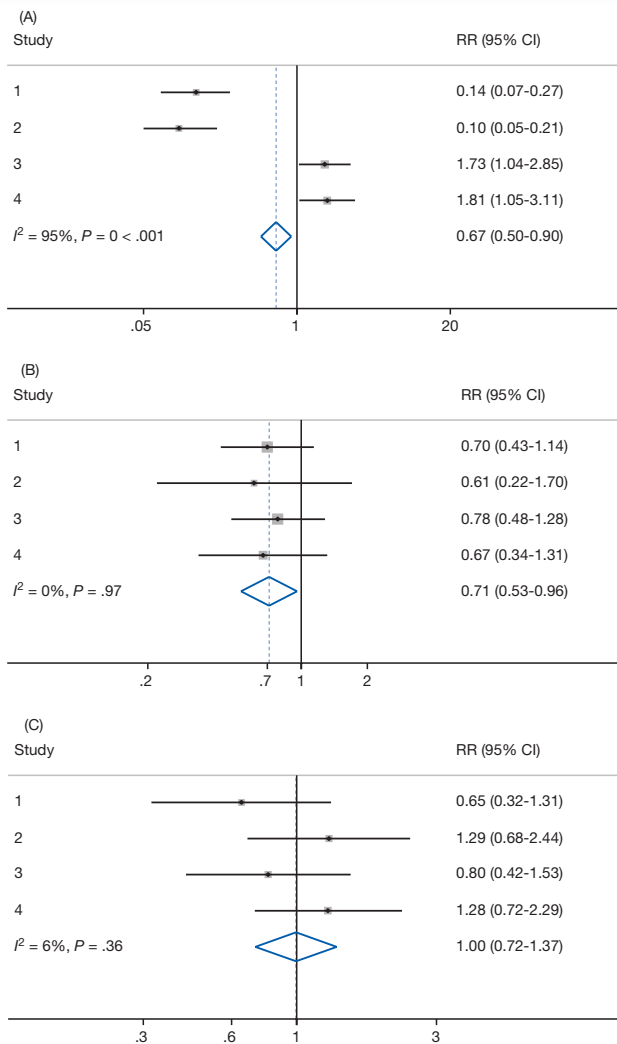
Constructing a rule to capture these inferences, one might suggest that “we are comfortable with a single summary effect when all studies suggest benefit or all studies suggest harm” (the case for B but not A). [Figure 15-5C](#), however, highlights the limitation of such a rule: this hypothetical meta-analysis C also shows point estimates on both sides of the line of no effect, but here we would be comfortable combining the results.

A better approach to assessing heterogeneity focuses on the magnitude of the differences in the point estimates of the studies. Large differences in point estimates make clinicians less confident in the pooled estimate (as in meta-analysis A). Small differences in the magnitude of point estimates (as in meta-analyses B and C) support the underlying assumption that, across the range of study patients, interventions, and outcomes included in the meta-analysis, the effect of interest is more or less the same.

There is a second, equally important criterion that clinicians should apply when judging whether combining the studies is appropriate. If CIs overlap widely (as in meta-analyses B and C), *random error*, or chance, remains a plausible explanation for the differences in the point estimates. When CIs do not overlap (as in meta-analysis A), random error becomes an unlikely explanation for differences in apparent treatment effect

FIGURE 15-5

## Results of Hypothetical Meta-analyses



Abbreviations: CI, confidence interval; RR, relative risk.



across studies. Visual assessment of heterogeneity is useful; formal statistical testing can provide complementary information.

### Yes-or-No Statistical Tests of Heterogeneity

The *null hypothesis* of the *test for heterogeneity* is that the underlying effect is the same in each of the studies<sup>20</sup> (eg, the RR derived from study 1 is the same as that from studies 2, 3, and 4). Therefore, the null hypothesis assumes that all of the apparent variability among individual study results is due to chance. *Cochran Q*, the most commonly used test for heterogeneity, generates a *probability* based on a  $\chi^2$  distribution that between-study differences in results equal to or greater than those observed are likely to occur simply by chance.

Meta-analysts may consider different thresholds for the significance of the test of heterogeneity (eg, a conventional threshold of  $P < .05$  or a more conservative threshold of  $P < .10$ ). As a general principle, however, a low  $P$  value of the test for heterogeneity means that random error is an unlikely explanation for the differences in results from study to study. Thus, a low  $P$  value decreases confidence in a single summary estimate that represents the treatment effect for all patients and all variations in the administration of a treatment. A high  $P$  value of the test of heterogeneity, on the other hand, increases our confidence that the assumption underlying combining studies holds true.

In Figure 15-5A, the  $P$  value associated with the test for heterogeneity is small ( $P < .001$ ), indicating that it is unlikely that we would observe results this disparate if all studies had the same underlying effect. On the other hand, the corresponding  $P$  values in Figure 15-3B and C are fairly large (.97 and .36, respectively). Therefore, in these 2 meta-analyses, chance is a likely explanation for the observed differences in effect.

When a meta-analysis includes studies with small sample sizes and a correspondingly small number of events, the test of heterogeneity may not have sufficient power to detect existing heterogeneity. Conversely, in a meta-analysis that includes studies with large sample sizes and a large number of events, the test for heterogeneity may provide potentially misleading results

that reveal statistically significant but unimportant differences in point estimates. This is another reason why clinicians need to use their own visual assessments of heterogeneity (similarity of point estimates, overlap of CIs) and consider the results of formal statistical tests in that context.

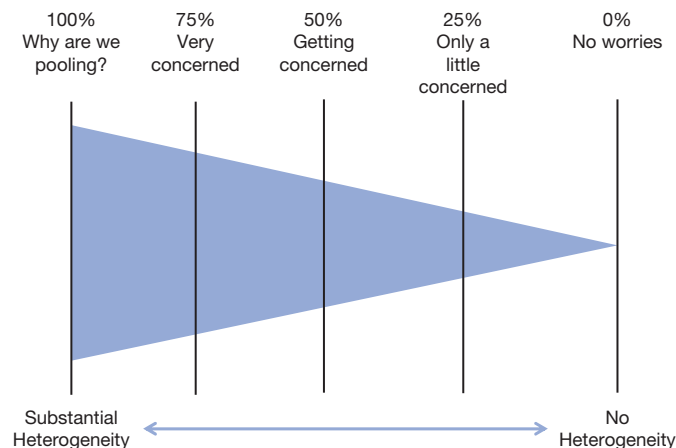
### Magnitude of Heterogeneity Statistical Tests

The  $I^2$  statistic is a preferred alternative approach for evaluating heterogeneity that focuses on the magnitude of variability rather than the statistical significance of variability.<sup>21</sup>

When the  $I^2$  is 0%, chance provides a satisfactory explanation for the variability in the individual study point estimates, and clinicians can be comfortable with a single summary estimate of treatment effect. As the  $I^2$  increases, we become progressively less comfortable with a single summary estimate, and the need to look for explanations of variability other than chance becomes more compelling. Figure 15-6 provides a guide for interpreting the  $I^2$ .

**FIGURE 15-6**

#### Interpretation of the $I^2$ Statistic



If provided by the meta-analysis authors, a 95% CI associated with the  $I^2$  can provide further insight regarding assessment of inconsistency. In most meta-analyses with a limited number of relatively small studies, this CI is quite large, suggesting the need for caution in making strong inferences regarding inconsistency.<sup>22</sup>

The results in Figure 15-5A generate an  $I^2$  of more than 75% (suggesting high heterogeneity), whereas the results in Figure 15-5B and C yield low  $I^2$  percentages of 0% and 6%, respectively (suggesting low heterogeneity).

### What to Do When Between-Study Variability in Results Is Large?

One of the credibility criteria introduced in Chapter 14 is whether the authors have addressed possible explanations of heterogeneity. When between-study variability is large, such an exploration becomes crucial.

Differences between study results can arise from differences in the population enrolled (eg, large effects in the more ill, smaller in the less ill), differences in the interventions (eg, if large doses are more effective than small doses), differences in the comparators (eg, smaller effects when standard care is optimal than when it is not), and study methods (eg, larger effect in studies with high risk of bias vs those with low risk of bias). Meta-analysis authors should conduct a test of interaction to determine whether the difference in effect estimates among subgroups is attributable to chance. Apparent subgroup effects are more likely to be true when they are based on within-trial rather than between-trial comparisons, are very unlikely to be due to chance, and are based on a small number of hypotheses specified a priori, including a specified direction. If these criteria are not met, any subgroup hypothesis warrants a high level of skepticism.

What if, in the end, we are left with a large degree of unexplained between-study heterogeneity for which chance does not provide an adequate explanation? This is not an uncommon

situation. Some argue that, in this situation, meta-analysis authors should not combine the results. Clinicians and patients, however, still need a best estimate of the treatment effect to inform their decisions. Pending further research that may explain the differences between results of different studies that address the same question, the summary estimate remains the best estimate of the treatment effect. Although clinicians and patients must use the best estimate to make their decisions, substantial unexplained inconsistency between studies appreciably reduces confidence in the summary estimate.<sup>23</sup>

## USING THE GUIDE

In Figure 15-1 and 15-2, for both nonfatal myocardial infarction and death, we note substantial differences in point estimates across studies. In the case of death, there is minimal CI overlap. Although the heterogeneity  $P$  values of .21 and .16 are not statistically significant, the  $I^2$  of 29% for nonfatal myocardial infarction and 30% for death suggest the presence of variability for which seeking a possible explanation is worthwhile.

Examining the data, we find that trials with a high risk of bias reveal a substantially larger reduction in the risk of nonfatal myocardial infarction. A test of interaction between the 2 groups of studies (those with high risk of bias and those with low risk of bias) yields a nonsignificant  $P$  value of .22, which indicates that the difference in the reduction in nonfatal myocardial infarction risk between these 2 subgroups of studies could be due to chance.

However, for the outcome of death, a test of interaction between the 2 groups of studies yields a significant  $P$  value of .04, which suggests that the risk of bias explains the observed heterogeneity (Figure 15-2). As we have mentioned previously, our inclination to use only the studies with low risk of bias is

reinforced by our awareness of the doubts that have been raised regarding the integrity of the data from the 2 studies with high risk of bias. Results of the studies with low risk of bias are consistent ( $I^2$  of 0% and  $P$  value for heterogeneity test of .68).

Meta-analysis of the outcome of nonfatal stroke reveals consistent results across trials with an  $I^2$  value of 0% and  $P$  value for the heterogeneity test of .71 (Figure 15-3).

### How Precise Are the Results?

Meta-analysis generates an estimate of the mean effect across studies and a CI around that estimate, that is, a range of values with a specified probability (typically 95%) of including the true effect (see Chapter 9, Confidence Intervals: Was the Single Study or Meta-analysis Large Enough?). When applying research evidence to a clinical question, one should determine whether clinical action would differ if the upper or the lower boundaries of the CI represented the truth. If the clinical decision is the same whether the upper or lower boundary of the CI represents the true effect, then the evidence is sufficiently precise. If across the range of the CI values our decision making would change, then we should have less confidence in the evidence and lower the confidence rating (eg, from high to moderate confidence).<sup>24</sup>

## USING THE GUIDE

**T**o determine the precision of the estimate of the effect of perioperative  $\beta$ -blockers on the risk of nonfatal myocardial infarction, you need to calculate the absolute effect, which requires knowledge of the RR and the *control event rate* (ie, the event rate in patients who did not receive  $\beta$ -blockers). Having decided that the best estimate of RR comes from focusing on the trials with low risk of bias rather than all trials included in

the meta-analysis, we note that the RR is 0.73 (95% CI, 0.61-0.88) (Figure 15-1). We obtain the control event rate from the trial that is by far the largest—and the one that likely enrolled the most representative population<sup>25</sup>—which was 215/4177 or approximately 52 per 1000. You can then calculate the decreased risk of nonfatal myocardial infarction in those using  $\beta$ -blockers as follows:

Risk with intervention = risk with control  $\times$  relative risk =  
 $52/1000 \times 0.73 =$  approximately 38 per 1000

Risk difference = risk with control – risk with intervention =  
 $52/1000 - 38/1000 = -14$  (approximately 14 fewer myocardial infarctions per 1000)

You can use the same process to calculate the CIs around the *risk difference*, substituting the boundaries of the CI (in this case, 0.61 and 0.88) for the point estimate (in this case, 0.73). For instance, for the upper boundary of the CI:

Risk with intervention =  $52/1000 \times 0.88 =$  approximately 46 per 1000

Risk with intervention – risk with control =  $46 - 52 = -6$   
 (approximately 6 fewer per 1000)

The estimate of absolute difference in nonfatal myocardial infarction when using  $\beta$ -blockers is therefore approximately 14 fewer per 1000, with a CI of approximately 6 to 20 fewer per 1000.

The corresponding absolute difference for nonfatal stroke is 2 more nonfatal strokes per 1000, with a CI of approximately 0 to 6 more per 1000; for death, the absolute difference is 6 deaths more per 1000 with a CI of approximately 0 to 13 more per 1000 (Table 15-1).

Lowering a confidence rating because of imprecision is always a judgment call. There seems to be no doubt about the need to lower confidence for nonfatal stroke (the effect ranges from no difference to an appreciable increase in nonfatal stroke)

TABLE 15-1

**Evidence Profile: Explicit Presentation of the Best Estimates of the Effect of Perioperative  $\beta$ -Blockers and the Confidence in Estimates**

Outcome	Confidence Assessment			
	No. of Participants (No. of Studies)	Risk of Bias	Consistency	Directness
Myocardial infarction	10 189 (5)	No serious limitations	No serious limitations	No serious limitations
Stroke	10 186 (5)	No serious limitations	No serious limitations	No serious limitations
Death	10 529 (9)	No serious limitations	No serious limitations	No serious limitations

Abbreviation: CI, confidence interval.

and likely for death (some may consider 1 additional death in 1000 acceptable given the reduction in myocardial infarction; most would not consider 6 in 1000 trivial). Regarding nonfatal myocardial infarction, our judgment was not to lower confidence for imprecision (Table 15-1).

### Do the Results Directly Apply to My Patient?

The optimal evidence for decision making comes from research that directly compared the interventions in which we are interested, evaluated in the populations in which we are interested, and measured outcomes important to patients. If populations, interventions, and outcomes in studies differ from those of interest (ie, the patient before us), we lose confidence in estimates of effect. In GRADE, the term “indirectness” is used as a label for these issues.<sup>26</sup>

So, for instance, the patient at hand may be very elderly and the trials may have included few, if any, such patients. The dose of a drug tested in the trials may be greater than the dose your patient can tolerate.

		Summary of Findings		
Precision	Reporting Bias	Confidence	Relative Effect (95% CI)	Risk Difference per 1000 Patients
No serious limitations	Not detected	High	0.73 (0.61-0.88)	14 fewer (6 fewer to 20 fewer)
Imprecise	Not detected	Moderate	1.73 (1.00-2.99)	2 more (0 more to 6 more)
Imprecise	Not detected	Moderate	1.27 (1.01-1.60)	6 more (0 more to 13 more)

Decisions regarding indirectness of patients and interventions depend on an understanding of whether biologic or social factors are sufficiently different that one might expect substantial differences in the magnitude of effect. Do elderly patients metabolize a drug differently from younger patients? Are there competing risks that will be responsible for the demise of elderly patients long before they experience the benefits of the intervention? Is there evidence that the tissue effect of a medication is highly dose dependent?

Another issue of indirectness arises when outcomes assessed in the studies differ from those of interest to patients. Trials often measure laboratory or *surrogate outcomes* that are not themselves important but are measured in the presumption that changes in the surrogate reflect changes in an outcome important to patients. For instance, we have excellent information about the effect of medications used in type 2 diabetes on hemoglobin A<sub>1C</sub> but limited information on their effect on macrovascular and microvascular disease. In almost every instance, we should reduce our confidence in estimates of effect on patient-important outcomes when all we have available is the effect on surrogates.



Lastly, a different type of indirectness occurs when clinicians must choose among interventions that have not been tested in head-to-head comparisons. For instance, we may want to choose among alternative bisphosphonates for managing osteoporosis. We will find many trials that compare each agent to placebo, but few, if any, that have compared them directly against one another.<sup>27</sup> Making comparisons among treatments under these circumstances requires extrapolating results for the existing comparisons and requires multiple assumptions (see Chapter 16, Network Meta-analysis).<sup>26</sup>

## USING THE GUIDE

**A**ssessing directness regarding the evidence bearing on the use of  $\beta$ -blockers in noncardiac surgery,<sup>1</sup> we note that the age of most patients enrolled across the trials ranged from 50 to 70 years, similar to your patient, who is 66 years old. Almost all of the trials enrolled patients undergoing surgical procedures classified as intermediate surgical risk, similar to the hip surgery of your patient. Most of the trials enrolled many patients who, like yours, had risk factors for heart disease. Although the drug used and the dose varied across trials, the consistent results suggest you can use a modest dose of the  $\beta$ -blocker with which you are most familiar. The outcomes of death, nonfatal stroke, and nonfatal infarction are the key outcomes of importance to your patients. Overall, the available evidence presented in the systematic review is direct and applicable to your patient and addresses the key outcomes (benefits and harms) needed for decision making.

### Is There Concern About Reporting Bias?

The most difficult types of bias for systematic review authors to address stem from the inclination of authors of original studies to publish material, either entire studies or specific outcomes,

based on the magnitude, direction, or statistical significance of the results. We call the systematic error in the body of evidence that results from this inclination *reporting bias*. When an entire study remains unreported, the standard term is publication bias. The reason for publication bias is that studies without statistically significant results (*negative studies*) are less likely to be published than studies that reveal apparent differences (*positive studies*). The magnitude and direction of a study's results may be more important determinants of publication than study design, relevance, or quality,<sup>28</sup> and positive studies may be as much as 3 times more likely to be published than negative studies.<sup>29</sup> When authors or study sponsors selectively manipulate and report specific outcomes and analyses, we use the term *selective outcome reporting bias*.<sup>30</sup> Selective reporting bias can be a serious problem. Empirical evidence suggests that half of the analysis plans of randomized trials are different in protocols than in published reports.<sup>31</sup> When the publication is delayed because of the lack of significance of results, authors have used the term *time lag bias*.<sup>32</sup>

Selective outcome reporting can also create misleading estimates of effect. A study of US Food and Drug Administration (FDA) reports found that they often included numerous unpublished studies and the findings of these studies can appreciably alter the estimates of effect.<sup>33</sup>

Reporting bias can intrude at virtually all stages of the planning, implementation, and dissemination of research. Even if studies with negative results succeed and get published, they may still suffer from *dissemination bias*: they may be published in less prominent journals, may not receive adequate attention from policymakers, may be omitted (whether identified or not) in narrative reviews, may be omitted (if unidentified) from systematic reviews, and may have minimal or no effect on formulation of policy guidelines. On the other hand, studies with positive results may receive disproportionate attention. For instance, they are more likely to appear in subsequent evidence summaries and in an evidence *synopsis*.<sup>34</sup>

The consequences of publication and reporting bias can corrupt the body of evidence, usually exaggerating estimates of magnitude of *treatment effect*. Systematic reviews that fail to identify and include unpublished studies face a risk of presenting overly sanguine estimates of treatment effectiveness.

The risk of publication bias is probably higher for systematic reviews and meta-analyses that are based on small studies. Small studies are more likely to produce nonsignificant results due to lack of statistical power and are easier to hide. Larger studies are not, however, immune. Sponsors and authors who are not pleased with the results of a study may delay publication or choose to publish their study in a journal with limited readership or a lower impact factor.<sup>32</sup>

An example of reporting bias is the Salmeterol Multicenter Asthma Research Trial, which was a randomized trial designed to examine the effect of salmeterol or placebo on a *composite endpoint* of respiratory-related deaths and life-threatening experiences. In September 2002, after a data safety and monitoring board review of 25 858 randomized patients that found a nearly significant increase in the primary outcome in salmeterol-treated patients, the sponsor terminated the study. In a significant deviation from the original protocol, the sponsor submitted to the FDA an analysis, including events in the 6 months after the termination of the trial, which produced an apparent diminution of the dangers associated with salmeterol. The FDA, through specific inquiry, eventually obtained the data and the results were finally published in January 2006, revealing the increased likelihood of respiratory-related deaths with salmeterol.<sup>35,36</sup>

### Strategies to Address Reporting Bias

Several tests have been developed to detect publication bias (Box 15-2); unfortunately, all have serious limitations. The tests require a large number of studies (ideally 30 or more), although many meta-analysis authors use them in analyses including few studies. Moreover, none of these tests has been validated against a *criterion standard* (or *gold standard*) of real data in which we know whether publication bias or other biases existed or not.<sup>37</sup>

## BOX 15-2

**Four Strategies to Address Reporting Bias**

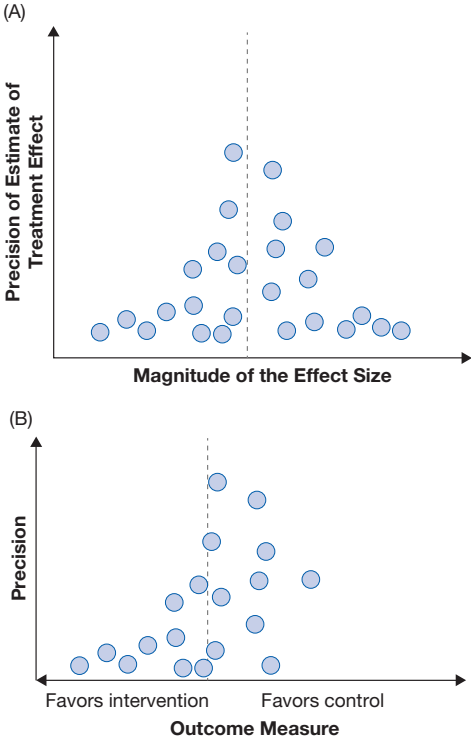
1. Examine whether the smaller studies show bigger effects
  - a. Funnel plots, visually assessed
  - b. Funnel plots, statistical analysis
2. Reconstruct evidence by restoring the picture after accounting for postulated publication bias
  - a. Trim and fill
3. Estimate the chances of publication according to the statistical significance level
4. Examine the evolution of effect size over time as more data appear

The first category of tests examines whether small studies differ from larger ones in their results. In a figure that relates the precision (as measured by sample size, inverse of *standard error* or *variance*) of studies included in a meta-analysis to the magnitude of treatment effect, the resulting display should resemble an inverted funnel (Figure 15-7A). Such *funnel plots* should be symmetric, around the point estimate (dominated by the largest trials) or the results of the largest trials themselves. A gap or empty area in the funnel suggests that studies have been conducted and not published (Figure 15-7B). Because visual determination of symmetry can be subjective, meta-analysts sometimes apply statistical tests for the symmetry of the funnel.<sup>37</sup>

Even when the funnel shape or the tests suggest publication bias, other explanations for asymmetry are possible. The small studies may have a higher risk of bias, which may explain their larger effects. On the other hand, the small studies may have chosen a more responsive patient group or administered the intervention more meticulously. Finally, there is always the possibility of a chance finding.

FIGURE 15-7

**Funnel Plot Showing No Publication Bias (A) and Showing Possible Publication Bias (B)**



A, The circles represent the point estimates of the trials. The pattern of distribution resembles an inverted funnel. Larger studies tend to be closer to the summary estimate (the dashed line). In this case, the effect sizes of the smaller studies are more or less symmetrically distributed around the summary estimate. B, This funnel plot shows that the smaller studies are not symmetrically distributed around either the point estimate (dominated by the larger trials) or the results of the larger trials themselves. The trials expected in the bottom right quadrant are missing. This suggests publication bias and an overestimate of the treatment effect relative to the underlying truth.

A second set of tests imputes and corrects for missing information and address its effect (*trim-and-fill method*). Again, the availability of few studies and the presence of heterogeneity make this second strategy inappropriate for most meta-analyses.

A third set of tests estimates whether there are differential chances of publication according to the level of statistical significance.<sup>38-40</sup> The excess significance test can be used in single meta-analyses and collections of multiple meta-analyses in the same field where similar biases may be operating.

Finally, a set of tests aims to examine whether evidence changes over time as more data accumulate. Continuously diminishing effects are characteristic of time lag bias.<sup>41</sup>

More compelling than any of these theoretical exercises is the success of systematic review authors in obtaining the results of unpublished studies that appear to be a complete collection of all of the studies that have been undertaken.

Prospective study registration with accessible results represents the best solution to reporting bias.<sup>42,43</sup> Prospective registration makes publication bias potentially identifiable; however, more detailed information is necessary to identify potential selective outcome and analysis reporting bias. Until complete reporting becomes a reality,<sup>44</sup> clinicians using research reports to guide their practice must remain cognizant of the dangers of reporting biases.

## USING THE GUIDE

**T**he authors of the systematic review and meta-analysis<sup>1</sup> addressing perioperative  $\beta$ -blockers constructed funnel plots that appear to be symmetrical, and the statistical tests for the symmetry of the plot were nonsignificant. The total number of patients included (>10 000) further reduces concern about publication bias, leaving no reason for lowering our confidence rating due to publication or reporting bias.

## Are There Reasons to Increase the Confidence Rating?

Some uncommon situations warrant an increase in the confidence rating of effect estimates from observational studies. Consider our confidence in the effect of hip replacement on reducing pain and functional limitations in severe osteoarthritis, epinephrine to prevent mortality in anaphylaxis, insulin to prevent mortality in diabetic ketoacidosis, or dialysis to prolong life in patients with end-stage renal failure.<sup>45</sup> In each of these situations, we are confident of a substantial treatment effect despite the absence of randomized trials. Why is that? The reason is a very large treatment effect that was achieved during a short period among patients with a condition that would have inevitably worsened in the absence of an intervention.

The GRADE approach provides specific guidance regarding large effect sizes: consider increasing the confidence rating by 1 level when there is a 2-fold reduction or increase in risk and consider increasing the confidence rating by 2 levels in the presence of a 5-fold reduction or increase in risk. For example, a systematic review and meta-analysis of observational studies examining the relationship between infant sleeping position and sudden infant death syndrome (SIDS) found an OR of 4.9 (95% CI, 3.6-6.6) of SIDS occurring with front vs back sleeping positions.<sup>46</sup> The “back to sleep” campaigns that were started in the 1980s were associated with a relative decrease in the incidence of SIDS by 50% to 70% in numerous countries.<sup>46</sup> This large effect increases our confidence in a true association.<sup>45</sup>

## AN EVIDENCE-BASED SUMMARY OF THE FINDINGS: THE EVIDENCE PROFILE

---

To optimally apply evidence summarized in a systematic review, practitioners need succinct, easily digestible presentations of confidence in effect estimates (quality of evidence) and magnitude of effects. They need this information to trade benefits and harms and communicate risks to their patients. They need

to know the confidence we have in a body of evidence to convey the uncertainty to their patients.

Systematic reviews may provide this summary in different ways. The GRADE Working Group recommends what are called *evidence profiles* (or a shortened version called *summary of findings tables*). Such tables present the relative and absolute effects of an intervention on each of the critical outcomes most important to patients, including a confidence rating. If stratifying patients' baseline risk for the outcome is possible, the absolute effect is presented for each risk strata separately.

## CLINICAL SCENARIO RESOLUTION

**T**able 15-1 presents the evidence profile summarizing the results of the systematic review addressing perioperative  $\beta$ -blockers. We see that evidence warranting high confidence suggests that individuals with underlying cardiovascular disease or risk factors for disease can expect a reduction in their risk of a perioperative nonfatal infarction of 14 in 1000 (from approximately 20 per 1000 to 6 per 1000). Unfortunately, they can also expect an increase in their risk of dying or experiencing a nonfatal stroke. Because most people are highly averse to the disability associated with stroke and at least equally averse to death, it is likely that most patients faced with this evidence would decline  $\beta$ -blockers as part of their perioperative regimen. Indeed, that is what our 66-year-old man with diabetes decides when you discuss the evidence with him.

## References

1. Bouri S, Shun-Shin MJ, Cole GD, Mayet J, Francis DP. Meta-analysis of secure randomised controlled trials of  $\beta$ -blockade to prevent perioperative death in non-cardiac. *Heart*. 2014;100(6):456-464.
2. Guyatt GH, Oxman AD, Vist GE, et al; GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336(7650):924-926.



3. Spencer FA, Iorio A, You J, et al. Uncertainties in baseline risk estimates and confidence in treatment effects. *BMJ*. 2012;345:e7401.
4. Schünemann HJ, Oxman AD, Brozek J, et al; GRADE Working Group. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ*. 2008;336(7653):1106-1110.
5. Rosenthal R. *Meta-analytic Procedures for Social Research*. 2nd ed. Newbury Park, CA: Sage Publications; 1991.
6. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Earlbaum Associates; 1988.
7. Smith K, Cook D, Guyatt GH, Madhavan J, Oxman AD. Respiratory muscle training in chronic airflow limitation: a meta-analysis. *Am Rev Respir Dis*. 1992;145(3):533-539.
8. Lacasse Y, Martin S, Lasserson TJ, Goldstein RS. Meta-analysis of respiratory rehabilitation in chronic obstructive pulmonary disease. A Cochrane systematic review. *Eura Medicophys*. 2007;43(4):475-485.
9. Thorlund K, Walter S, Johnston B, Furukawa T, Guyatt G. Pooling health-related quality of life outcomes in meta-analysis—a tutorial and review of methods for enhancing interpretability. *Res Synth Methods*. 2011;2(3):188-203.
10. Guyatt GH, Thorlund K, Oxman AD, et al. GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles—continuous outcomes. *J Clin Epidemiol*. 2013;66(2):173-183.
11. Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. *J Clin Epidemiol*. 2011;64(4):380-382.
12. Organizations. The GRADE Working Group. <http://www.gradeworkinggroup.org/society/index.htm>. Accessed April 9, 2014.
13. Balshem H, Helfand M, Schünemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol*. 2011;64(4):401-406.
14. Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ*. 2008;336(7644):601-605.
15. Bassler D, Briel M, Montori VM, et al; STOPIT-2 Study Group. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. *JAMA*. 2010;303(12):1180-1187.
16. Dunkelgrun M, Boersma E, Schouten O, et al; Dutch Echocardiographic Cardiac Risk Evaluation Applying Stress Echocardiography Study Group. Bisoprolol and fluvastatin for the reduction of perioperative cardiac mortality and myocardial infarction in intermediate-risk patients undergoing noncardiovascular surgery: a randomized controlled trial (DECREASE-IV). *Ann Surg*. 2009;249(6):921-926.
17. Poldermans D, Boersma E, Bax JJ, et al; Dutch Echocardiographic Cardiac Risk Evaluation Applying Stress Echocardiography Study Group. The effect of bisoprolol on perioperative mortality and myocardial infarction in high-risk patients undergoing vascular surgery. *N Engl J Med*. 1999;341(24):1789-1794.

18. Hatala R, Keitz S, Wyer P, Guyatt G; Evidence-Based Medicine Teaching Tips Working Group. Tips for learners of evidence-based medicine: 4. Assessing heterogeneity of primary studies in systematic reviews and whether to combine their results. *CMAJ*. 2005;172(5):661-665.
19. Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. *Lancet*. 1998;351(9096):123-127.
20. Lau J, Ioannidis JP, Schmid CH. Quantitative synthesis in systematic reviews. *Ann Intern Med*. 1997;127(9):820-826.
21. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327(7414):557-560.
22. Ioannidis JP, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. *BMJ*. 2007;335(7626):914-916.
23. Guyatt GH, Oxman AD, Kunz R, et al; GRADE Working Group. GRADE guidelines: 7. Rating the quality of evidence—inconsistency. *J Clin Epidemiol*. 2011;64(12):1294-1302.
24. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines 6. Rating the quality of evidence—imprecision. *J Clin Epidemiol*. 2011;64(12):1283-1293.
25. Devereaux PJ, Yang H, Yusuf S, et al; POISE Study Group. Effects of extended-release metoprolol succinate in patients undergoing non-cardiac surgery (POISE trial): a randomised controlled trial. *Lancet*. 2008;371(9627):1839-1847.
26. Guyatt GH, Oxman AD, Kunz R, et al; GRADE Working Group. GRADE guidelines: 8. Rating the quality of evidence—indirectness. *J Clin Epidemiol*. 2011;64(12):1303-1310.
27. Murad MH, Drake MT, Mullan RJ, et al. Clinical review. Comparative effectiveness of drug treatments to prevent fragility fractures: a systematic review and network meta-analysis. *J Clin Endocrinol Metab*. 2012;97(6):1871-1880.
28. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet*. 1991;337(8746):867-872.
29. Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ*. 1997;315(7109):640-645.
30. Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA*. 2004;291(20):2457-2465.
31. Saquib N, Saquib J, Ioannidis JP. Practices and impact of primary outcome adjustment in randomized controlled trials: meta-epidemiologic study. *BMJ*. 2013;347:f4313.
32. Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *JAMA*. 1998;279(4):281-286.
33. McDonagh MS, Peterson K, Balshem H, Helfand M. US Food and Drug Administration documents can provide unpublished evidence relevant to systematic reviews. *J Clin Epidemiol*. 2013;66(10):1071-1081.
34. Carter AO, Griffin GH, Carter TP. A survey identified publication bias in the secondary literature. *J Clin Epidemiol*. 2006;59(3):241-245.

35. Lurie P, Wolfe SM. Misleading data analyses in salmeterol (SMART) study. *Lancet*. 2005;366(9493):1261-1262, discussion 1262.
36. Nelson HS, Weiss ST, Bleecker ER, Yancey SW, Dorinsky PM; SMART Study Group. The Salmeterol Multicenter Asthma Research Trial: a comparison of usual pharmacotherapy for asthma or usual pharmacotherapy plus salmeterol. *Chest*. 2006;129(1):15-26.
37. Lau J, Ioannidis JP, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. *BMJ*. 2006;333(7568):597-600.
38. Hedges L, Vevea J. Estimating effect size under publication bias: small sample properties and robustness of a random effects selection model. *J Educ Behav Stat*. 1996;21(4):299-333.
39. Vevea J, Hedges L. A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*. 1995;60(3):419-435.
40. Ioannidis JP, Trikalinos TA. An exploratory test for an excess of significant findings. *Clin Trials*. 2007;4(3):245-253.
41. Ioannidis JP, Contopoulos-Ioannidis DG, Lau J. Recursive cumulative meta-analysis: a diagnostic for the evolution of total randomized evidence from group and individual patient data. *J Clin Epidemiol*. 1999;52(4):281-291.
42. Boissel JP, Haugh MC. Clinical trial registries and ethics review boards: the results of a survey by the FICHTRE project. *Fundam Clin Pharmacol*. 1997;11(3):281-284.
43. Horton R, Smith R. Time to register randomised trials. The case is now unanswerable. *BMJ*. 1999;319(7214):865-866.
44. Dickersin K, Rennie D. The evolution of trial registries and their use to assess the clinical trial enterprise. *JAMA*. 2012;307(17):1861-1864.
45. Guyatt GH, Oxman AD, Sultan S, et al; GRADE Working Group. GRADE guidelines: 9. Rating up the quality of evidence. *J Clin Epidemiol*. 2011;64(12):1311-1316.
46. Gilbert R, Salanti G, Harden M, See S. Infant sleeping position and the sudden infant death syndrome: systematic review of observational studies and historical review of recommendations from 1940 to 2002. *Int J Epidemiol*. 2005;34(4):874-887.

# Network Meta-analysis

Edward J. Mills, John P. A. Ioannidis,  
Kristian Thorlund, Holger J. Schünemann,  
Milo A. Puhan, and Gordon Guyatt

## IN THIS CHAPTER

---

### Clinical Scenario

### Finding the Evidence

### Introduction

### How Serious Is the Risk of Bias?

Did the Meta-analysis Include Explicit and Appropriate Eligibility Criteria?

Was Biased Selection and Reporting of Studies Unlikely?

Did the Meta-analysis Address Possible Explanations of Between-Study Differences in Results?

Did the Authors Rate the Confidence in Effect Estimates for Each Paired Comparison?

(continued on following page)

### **What Are the Results?**

What Was the Amount of Evidence in the Treatment Network?

Were the Results Similar From Study to Study?

Were the Results Consistent in Direct and Indirect Comparisons?

How Did Treatments Rank and How Confident Are We in the Ranking?

Were the Results Robust to Sensitivity Assumptions and Potential Biases?

### **How Can I Apply the Results to Patient Care?**

Were All Patient-Important Outcomes Considered?

Were All Potential Treatment Options Considered?

Are Any Postulated Subgroup Effects Credible?

### **Clinical Scenario Resolution**

### **Conclusion**

---

## CLINICAL SCENARIO

**Y**our patient is a 45-year-old woman who experiences frequent migraine headaches that last from 4 to 24 hours and prevent her from attending work or looking after her children. She has exhausted efforts to manage the symptoms with nonsteroidal anti-inflammatory drugs and seeks additional treatment. You decide to recommend a triptan for the patient's migraine headaches but are wondering how to choose from the 7 available triptans. You retrieve a *network meta-analysis* (NMA) that evaluates the different triptans among this patient population.<sup>1</sup> You are not familiar with this type of study, and you wonder if there are special issues to which you should attend in evaluating its methods and results.

## FINDING THE EVIDENCE

---

You start by typing “migraine triptans” in the search box of an evidence-based summary website with which you are familiar. You find several chapters related to the management of migraine and drug information on the different drugs that are available. However, despite the profusion of *evidence* comparing single regimens, you wonder if all triptans have been compared, ideally in a single *systematic review*. To search for such a review, you type “migraine triptans comparison” in PubMed’s Clinical Queries (<http://www.ncbi.nlm.nih.gov/pubmed/clinical>; see Chapter 4, Finding Current Best Evidence). In the results page, the middle column, which applies a broad filter for potential systematic reviews, retrieves 21 citations. The first strikes you as the most relevant to your question. It is a network meta-analysis that evaluates the different triptans among your patient population.<sup>1</sup> You are not familiar with this type of study, and you wonder if there are special issues to which you should attend in evaluating its methods and results.

## INTRODUCTION

---

Traditionally, a meta-analysis addresses the merits of one intervention vs another (eg, *placebo* or another active intervention). Data are combined from all studies—often *randomized clinical trials* (RCTs)—that meet eligibility criteria in what we will term a pairwise meta-analysis. Compared with a single RCT, a *meta-analysis* improves the power to detect differences and also facilitates examination of the extent to which there are important differences in *treatment effects* across eligible RCTs—variability that is frequently called *heterogeneity*.<sup>2,3</sup> Large unexplained heterogeneity may reduce a reader's confidence in estimates of treatment effect (see Chapter 15, Understanding and Applying the Results of a Systematic Review and Meta-analysis).

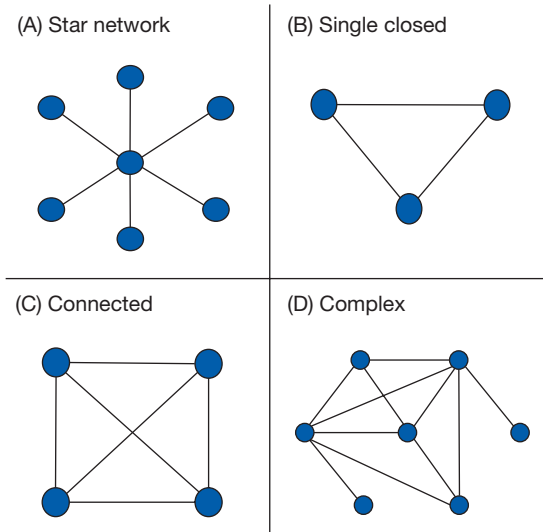
A drawback of traditional pairwise meta-analysis is that it evaluates the effects of only 1 intervention vs 1 comparator and does not permit inferences about the relative effectiveness of several interventions. For many medical conditions, however, there are a selection of interventions that have most frequently been compared with placebo and occasionally with one another.<sup>4,5</sup> For example, despite 91 completed and ongoing RCTs that address the effectiveness of the 9 biologic drugs for the treatment of rheumatoid arthritis, only 5 compare biologics directly against each other.<sup>4</sup>

Recently, another form of meta-analysis, called an NMA (also known as a multiple or mixed treatment comparison meta-analysis) has emerged.<sup>6,7</sup> The NMA approach provides estimates of effect sizes for all possible pairwise comparisons whether or not they have actually been compared head to head in RCTs. [Figure 16-1](#) displays examples of common networks of treatments.

Our ability to provide estimates of relative effect when 2 interventions, A and B, have not been tested head to head against one another comes from what are called indirect comparisons. We can make an indirect comparison if the 2 interventions

FIGURE 16-1

## Examples of Possible Network Geometry



The figure shows 4 network graphs. In each graph, the lines show where direct comparisons exist from 1 or more trials. Figure 16-1A shows a star network, where all interventions have just 1 mutual comparator. Figure 16-1B shows a single closed loop that involves 3 interventions and can provide data to calculate both direct comparisons and indirect comparisons. Figure 16-1C shows a well-connected network, where all interventions have been compared against each other in multiple randomized clinical trials. Figure 16-1D is an example of a complex network with multiple loops and also arms that have sparse connections.

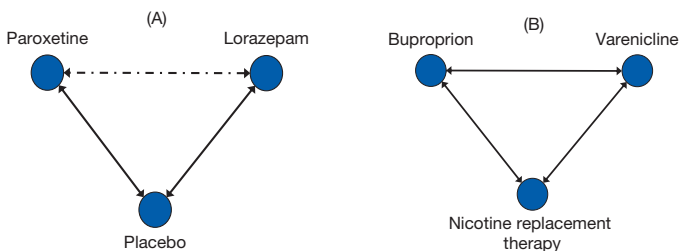
(eg, paroxetine and lorazepam in [Figure 16-2A](#)) have each been compared directly against another intervention, C (eg, placebo).

For instance, assume that A (eg, paroxetine) substantially reduces the odds of an adverse outcome relative to C (placebo) (*odds ratio* [OR], 0.5). Intervention B (eg, lorazepam), on the other hand, has no impact relative to C on that outcome (OR, 1.0).



FIGURE 16-2

## A Simple Indirect Comparison and Simple Closed Loop



In the first example (A), there is direct evidence from paroxetine compared with placebo and direct evidence of lorazepam compared with placebo. Therefore, the indirect comparison can be applied to determine the effect of paroxetine compared with lorazepam, even if no direct head-to-head comparison exists on these 2 agents. In the second example (B), there is direct evidence that compares nicotine replacement therapy with both varenicline and bupropion. There is also direct evidence that compares bupropion with varenicline. Therefore, one has enough information to evaluate whether the results are coherent between direct and indirect evidence.

One might then reasonably deduce that A is substantially superior to C—indeed, our best estimate of the OR of A vs B would be 0.5/1.0 or 0.5. The ratio of the OR in such a situation is our way of estimating the effect of A vs B on the outcome of interest.<sup>8</sup>

Network meta-analyses, which simultaneously include both direct and *indirect evidence* (see Figure 16-2B for an example in which both direct and indirect evidence is available, sometimes called a closed loop), are subject to 3 chief considerations. The first is an assumption that is also necessary for a conventional meta-analysis (see Chapter 15, Understanding and Applying the Results of a Systematic Review and Meta-analysis). Among trials available for pairwise comparisons, are the studies sufficiently *homogenous* to combine for each intervention?

Second, are the trials in the network sufficiently similar, with the exception of the intervention (eg, in important features, such as populations, design, or outcomes)?<sup>9</sup> For instance, if trials of drug

A vs placebo differ substantially in the characteristics of the population studied from the population in drug B vs placebo, inferences about the relative effect of A and B on the basis of how each did against placebo become questionable. Third, where direct and indirect evidence exist, are the findings sufficiently consistent to allow confident pooling of direct and indirect evidence together?

By including evidence from both direct and indirect comparisons, an NMA may increase precision of estimates of the relative effects of treatments and facilitate simultaneous comparisons, or even ranking, of these treatments.<sup>7</sup> However, because NMAs are methodologically sophisticated, they are often challenging to interpret.<sup>10</sup>

One challenge clinicians will face with NMAs is that they usually use *Bayesian analysis* approaches rather than the *frequentist analysis* approaches with which most of us are more familiar. Clinicians need not worry further about this, and the main reason for pointing it out is as an alert to a difference in terms. Clinicians are used to considering *confidence intervals* (CIs) around estimates of treatment effect. The Bayesian equivalent are called *credible intervals* and can be interpreted in conceptually the same way as CIs.

Here, we demystify NMAs by using the 3 questions of *risk of bias*, results, and applicability of results. [Box 16-1](#) includes all issues relevant to evaluating systematic reviews. Our discussion in this chapter does not include all of the issues but rather highlights those that are most important, or differ, in NMAs.

## BOX 16-1

### Users' Guides Critical Appraisal Tool

#### How serious is the risk of bias?

Did the review include explicit and appropriate eligibility criteria?

Was biased selection and reporting of studies unlikely?

Did the review address possible explanations of between-study differences in results?

Were selection and assessments of studies reproducible?

Did the authors rate the confidence in effect estimates for each paired comparison?

#### **What are the results?**

What was the amount of evidence in the network?

Were the results similar from study to study?

Were the results consistent in direct and indirect comparisons?

How did treatments rank and how secure are we in the rankings?

Were the results robust to sensitivity assumptions and potential biases?

#### **How can I apply the results to patient care?**

Were all patient-important outcomes considered?

Were all potential treatment options considered?

Are any postulated subgroup effects credible?

What is the overall quality and what are limitations of the evidence?

## **HOW SERIOUS IS THE RISK OF BIAS?**

---

### **Did the Meta-analysis Include Explicit and Appropriate Eligibility Criteria?**

One can formulate questions of optimal patient management in terms of the *PICO* framework of patients (P), interventions (I), comparisons (C), and outcomes (O).

Broader eligibility criteria may enhance *generalizability* of the results but may be misleading if participants are too dissimilar and as a consequence heterogeneity is large. Diversity of interventions may also be excessive if authors pool results from

different doses or even different agents in the same class (eg, all statins), based on the assumption that effects are similar. You should ask whether investigators have been too broad in their inclusion of different populations, different doses or different agents in the same class, or different outcomes to make comparisons across studies credible.

### Was Biased Selection and Reporting of Studies Unlikely?

Some NMAs apply the search strategies from other systematic reviews as the basis for identifying potentially eligible trials. Readers can be confident in such approaches only if authors have updated the search to include recently published trials.<sup>11</sup>

The eligible interventions can be unrestricted. Sometimes, however, the authors may choose to include only a specific set of interventions, eg, those available in their country. Some industry-initiated NMAs may choose to consider only a sponsored agent and its direct competitors.<sup>12</sup> This may omit the optimal agent for some situations and tends to give a fragmented picture of the evidence. It is typically best to include all interventions<sup>13</sup> because data on clearly suboptimal or abandoned interventions may still offer indirect evidence for other comparisons.<sup>14</sup>

In an NMA of 12 treatments for major depression, the authors chose to exclude placebo-controlled RCTs and included only head-to-head active treatment RCTs.<sup>15</sup> However, *publication bias* in the antidepressant literature is well acknowledged,<sup>16,17</sup> and by excluding placebo-controlled trials, the analysis loses the opportunity to benefit from additional available evidence.<sup>18</sup> Exclusion of eligible interventions, in this case placebo, may not just decrease statistical power but may also change the overall results.<sup>14</sup> Placebo-controlled trials may be different

than head-to-head comparison trials in their conduct or in the degree of bias (eg, they may have more or less publication bias or *selective outcome reporting* and selective analysis reporting). Thus, their exclusion may also have an impact on the *point estimates* of the effects of pairwise comparisons and may affect the relative ranking of regimens.<sup>14</sup> When an NMA of second-generation antidepressants was later conducted and included placebo-controlled trials, relying only on the relative differences among treatments using the same depression scale, the authors reached a different interpretation than the earlier NMA.<sup>15,19,20</sup>

Finally, original trials often address multiple outcomes. Selection of NMA outcomes should not be data driven but based on importance for patients and consider both outcomes of benefit and *harm*.

### **Did the Meta-analysis Address Possible Explanations of Between-Study Differences in Results?**

When substantial clinical variability is present (this is usually, and appropriately, the case), authors may conduct *subgroup analyses* or *meta-regression* to explain heterogeneity. If such analyses are successful in explaining heterogeneity, the NMA may provide results that more optimally fit the clinical setting and characteristics of the patient you are treating.<sup>21</sup> For example, in an NMA evaluating different statins for cardiovascular disease protection, the authors used meta-regression to address whether it was appropriate to combine results across primary and secondary prevention populations, different statins, and different doses of statins.<sup>22</sup> Meta-regression suggested heightened efficacy in those with prior cardiac events and those with a

history of hypertension, possibly suggesting a more compelling case for statin use in such populations.

Inclusion of multiple control interventions (eg, placebo, no intervention, older standard of care) may enhance the robustness and connectedness of the treatment network. It is, however, important to gauge and account for potential differences between control groups. For example, because of potential placebo effects, patients receiving placebo in a *blinded* RCT may have differing responses than patients receiving no intervention in a nonblinded RCT. Thus, if an active treatment, A, has been compared with placebo and another active treatment, B, has been compared with no intervention, the different choice of control groups may produce misleading results (B may appear superior, but the use of placebo as the comparator in the A trials may be responsible for the difference). As with active interventions, meta-regression may address this problem.

For example, in an NMA evaluating the effectiveness of smoking cessation therapies, the authors combined placebo-controlled arms with standard-of-care control arms and then used meta-regression to examine whether the choice of control changed the *effect size*.<sup>23</sup> The authors found that trials that used placebo controls had smaller effect sizes than those that used standard of care, which explained the heterogeneity.

### Did the Authors Rate the Confidence in Effect Estimates for Each Paired Comparison?

The treatment effects in an NMA are typically reported with common effect sizes along with 95% credible intervals. Credible intervals are the Bayesian equivalent to the more commonly understood CIs. When there are  $K$  interventions included in the treatment network, there are  $K^*(K-1)/2$  possible pairwise

comparisons. For example, if there are 7 interventions, then there are  $7 \times (7-1)/2$ , or 21, possible pairwise comparisons. Like authors of conventional meta-analyses, authors of NMAs need to address confidence in estimates of effect for each paired comparison (A vs B, A vs C, B vs C, etc—15 comparisons in the NMA example with 7 interventions). The necessity for these confidence ratings is that evidence may warrant strong inferences (ie, high confidence in estimates) for the superiority of one treatment over another (A vs B, for instance) and only weak inferences (ie, very low confidence in estimates) for the judgment of superiority of another pairing (A vs C).

The GRADE Working Group has provided a framework that is well suited to addressing confidence in estimates (see Chapter 15, Understanding and Applying the Results of a Systematic Review and Meta-analysis). We lose confidence in direct comparisons of alternative treatments if the relevant randomized trials have failed to protect against risk of bias by *allocation concealment*, blinding, and preventing loss to *follow-up* (see Chapter 6, Therapy [Randomized Trials]). We also lose confidence when CIs (or in the case of a Bayesian NMA, credible intervals) on pooled estimates are wide (imprecision); results vary from study to study and we cannot explain the differences (*inconsistency*); the population, intervention, or outcome differ from that of primary interest (*indirectness*); or we are concerned about publication bias.

Ideally, for each paired comparison, authors will present the pooled estimate for the direct comparison (if there is one) and its associated rating of confidence, the indirect comparison(s) that contributed to the pooled estimate from the NMA and its associated rating of confidence, and the NMA estimate and the associated rating of confidence. Criteria for judging confidence in estimates for direct comparisons are well established. Although these criteria provide considerable guidance in assessing confidence in indirect estimates, judgments regarding confidence in estimates from indirect comparisons present additional challenges. Criteria for addressing these challenges are still evolving, reflecting that NMA is still a very new method.

## USING THE GUIDE

**R**eturning to our opening scenario, the NMA we identified compared the efficacy of different triptans for the abortive treatment of migraine headaches.<sup>1</sup> Patients of interest included adults 18 to 65 years old who experience migraines, with or without aura. Experimental and control interventions included available oral triptans, placebos, and no-treatment controls. The outcomes of interest were pain-free response at 2 hours and 24 hours after the onset of headache. Patients in the included RCTs met similarly broad diagnostic criteria based on criteria from the International Headache Society and had to experience at least 1 migraine headache every 6 weeks. The outcomes assessed are important to patients, and their definitions were consistent across trials. Moreover, the authors planned to assess dose as a potential effect modifier.

The authors conducted a comprehensive search for published literature and sought unpublished RCTs via contact with industry trialists. Two reviewers conducted the search and extracted data independently, in duplicate. The authors did not rate the confidence in estimates from paired comparisons but provided information that allows conclusions about confidence. The authors reported events as proportions with ORs for treatment effects.

## WHAT ARE THE RESULTS?

---

### What Was the Amount of Evidence in the Treatment Network?

One can gauge the amount of evidence in the treatment network from the number of trials, total sample size, and number of events for each treatment and comparison. Furthermore, the



extent to which the treatments are connected in the network is an important determinant of the confidence we can have in the estimates that emerge from the NMA. Understanding the *geometry of the network* (nodes and links) will permit clinicians to examine the larger picture and see what is compared to what.<sup>24</sup> Authors will generally present the structure of the network (as in the examples in Figure 16-1).

When alternative interventions have been compared with a single common comparator (eg, placebo), we call this a star network (Figure 16-1A). A star network only allows for indirect comparisons among active treatments, which reduces confidence in effects, particularly if there are a limited number of trials, patients, and events.<sup>25</sup> When there are data available that use both direct and indirect evidence of the same interventions, we refer to this as a closed loop (Figure 16-1B). The presence of direct evidence increases our confidence in the estimates of interest.

Often, a treatment network will include a mixture of exclusively indirect links and closed loops (Figure 16-1C and D). Most networks have unbalanced shapes with many trials of some comparisons, but few or none of others.<sup>24</sup> In this situation (and indeed, in many situations, as we have pointed out in our discussion of the need for a confidence rating of each paired comparison), evidence may warrant high confidence for some treatments and comparisons but low confidence for others. The credible intervals around direct, indirect, and NMA estimates provide a helpful index of the amount of information available for each paired comparison.

### Were the Results Similar From Study to Study?

In a traditional meta-analysis of paired treatment comparisons, results often vary from study to study. Investigators can address possible explanations of differences in treatment effects using a subgroup analysis and meta-regression. However, these analyses are limited in the presence of small numbers of trials, and apparent subgroup effects often prove spurious, an issue to which we return in our discussion of applicability.<sup>26-28</sup>

Network meta-analyses, with larger numbers of patients and studies, present opportunities for more powerful exploration of explanations of between-study differences. Indeed, as we have pointed out in a prior section of this chapter—Did the Review Address Possible Explanations of Between-Study Differences in Results?—the search conducted by NMA authors for explanations for heterogeneity may be informative.

Nevertheless, as is true for conventional meta-analyses, NMA is vulnerable to unexplained differences in results from study to study. Ideally, NMA authors will, in summarizing the results of each paired comparison, alert you to the extent of inconsistency in results in both the direct and indirect comparisons and the extent to which confidence in estimates decreases accordingly (see Chapter 15, Understanding and Applying the Results of a Systematic Review and Meta-analysis).

### Were the Results Consistent in Direct and Indirect Comparisons?

Direct comparisons of treatments are generally more trustworthy than indirect comparisons. However, these head-to-head trials can also yield misleading estimates (eg, when conflicts of interest influence the choice of comparators used or result in selective reporting). Therefore, indirect comparisons may on occasion provide more trustworthy estimates.<sup>29</sup>

Deciding what estimates are most trustworthy (direct, indirect, or network) requires assessing whether the direct and indirect estimates are consistent or discrepant. One can assess whether direct and indirect estimates yield similar effects whenever there is a closed loop in the network (as in Figure 16-2B). Statistical methods exist for checking this type of inconsistency, typically called a test for *incoherence*.<sup>30,31</sup>

A group of investigators applied a test of incoherence to 112 interventions in which direct and indirect evidence was available. They found that the results were statistically inconsistent 14% of the time.<sup>9</sup> This same evaluation found that comparisons

with smaller number of trials and measuring subjective outcomes had a greater risk of incoherence.

Authors' presentation of direct and indirect estimates for each paired comparison will allow you to easily examine the extent of incoherence between direct and indirect estimates. Authors can perform statistical tests to determine whether chance can explain the difference between direct and indirect estimates. Often, however, the amount of data is limited and not sufficient, and important differences may still exist in the absence of a statistically significant difference.

When incoherence is present, there are many explanations for the authors—and for you—to consider (Box 16-2). Just as unexplained heterogeneity in any direct paired comparison decreases confidence in the pooled estimate, unexplained incoherence reduces confidence in the estimate that arises from the network. Indeed, when large incoherence is present, the more credible estimate may come from either the direct (usually) or indirect (seldom) comparison rather than from the network.

## BOX 16-2

### Potential Reasons for Incoherence Between the Results of Direct and Indirect Comparisons

#### Chance

#### Genuine differences in results

Differences in enrolled participants (eg, entry criteria, clinical setting, disease spectrum, baseline risk, selection based on prior response)

Differences in the interventions (eg, dose, duration of administration, prior administration [second-line treatment])

Differences in background treatment and management (eg, evolving treatment and management in more recent years)

Differences in definition or measurement of outcomes

**Bias in head-to-head (direct) comparisons**

Optimism bias with unconcealed analysis

Publication bias

Selective reporting of outcomes and of analyses

Inflated effect size in *stopped early trials* and in early evidence

Limitations in allocation concealment, blinding, loss to follow-up, analysis as randomized

**Bias in indirect comparisons**

Each of the biasing issues above can affect the results of the direct comparisons on which the indirect comparisons are based

For example, a meta-analysis examining the analgesic efficacy of paracetamol plus codeine in surgical pain found a direct comparison that indicated the intervention was more efficacious than paracetamol alone (mean difference in pain intensity change, 6.97; 95% CI, 3.56-10.37). The *adjusted indirect comparison* did not find a significant difference between paracetamol plus codeine and paracetamol alone (-1.16; 95% CI, -6.95 to 4.64).<sup>32</sup> In this example, the direct and indirect evidence was statistically significantly incoherent ( $P = .02$ ). The explanation for incoherence may be that the direct trials included patients with lower pain intensity at baseline, and such patients may be more responsive to the addition of codeine.

## How Did Treatments Rank and How Confident Are We in the Ranking?

Besides presenting treatment effects, authors may also present the probability that each treatment is superior to all other treatments, allowing ranking of treatments.<sup>33,34</sup> Although this

approach is appealing, it may be misleading because of fragility in the rankings, because differences among the ranks may be too small to be important, or because of other limitations in the studies (eg, risk of bias, inconsistency, indirectness).

We have already provided one example of such a misleading ranking: in an NMA of drug treatments to prevent fragility hip fractures, the authors' conclusion that teriparatide had the highest probability of being ranked first across 10 treatments<sup>24</sup> was misleading because comparison of teriparatide with all other agents, including placebo, warranted only low or very low confidence.

In another example, an NMA that examined direct-acting agents for hepatitis C found no statistical difference for sustained virologic response between telaprevir and boceprevir (OR, 1.42; 95% credible interval, 0.89-2.25); on the basis of these results, the probability of being the best favors teleprevir by far (93%) over boceprevir (7%).<sup>35,36</sup> However, this 93% probability provides a misleadingly strong endorsement for teleprevir. The lower boundary of the credible interval tells us that our confidence in substantial superiority of teleprevir is very low.

Examination of the confidence in estimates from each paired comparisons provides insight into the trustworthiness of any rankings, and reveals the importance of providing such ratings.

### **Were the Results Robust to Sensitivity Assumptions and Potential Biases?**

Given the complexity of some NMAs, authors may assess the robustness of their study findings by applying sensitivity

analyses that reveal how the results change if some criteria or assumptions change. Sensitivity analyses may include restricting the analyses to trials with low risk of bias only or examining different but related outcomes. The Cochrane Handbook provides a discussion of sensitivity analyses.<sup>37</sup>

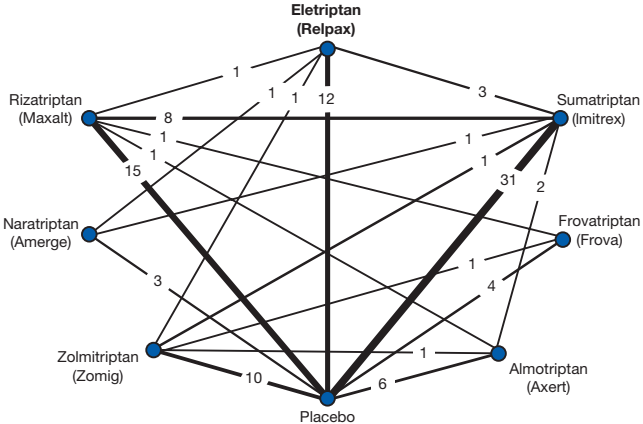
For example, in an NMA on prevention of chronic obstructive pulmonary disease (COPD) exacerbations, the authors used the incidence rate as the primary outcome. However, there is some debate on whether incidence rates should be used in COPD trials,<sup>38</sup> and so the authors conducted sensitivity analyses with the binary outcome of ever having an exacerbation. The results were sufficiently similar to consider the analyses robust.<sup>39</sup>

## USING THE GUIDE

**R**eturning to our clinical scenario, [Figure 16-3](#) displays the network of considered treatments for pain-free response at 2 hours. The authors included 74 RCTs that examined triptans for the treatment and prevention of migraine attacks. Placebo was compared with eletriptan, sumatriptan, rizatriptan, zolmitriptan, almotriptan, naratriptan, and frovatriptan in 15, 30, 16, 5, 9, 5, and 4 trials, respectively. The amount of evidence varied across these comparisons. For example, naratriptan had only been compared with placebo in 2 trials; therefore, confidence in these estimates is likely to be low. Evidence for sumatriptan and rizatriptan was based on a larger amount of evidence from both direct and indirect comparisons. Sumatriptan ( $n = 30$ ), rizatriptan ( $n = 20$ ), and eletriptan ( $n = 16$ ) had the most links, whereas placebo was the most connected node ( $n = 68$ ). The most common

FIGURE 16-3

**Treatment Network for the Drugs Considered in the Example Network Meta-analysis on Triptans for the Abortive Treatment of Migraine for Pain-Free Response at 2 Hours**



The lines between treatment nodes indicate the comparisons made throughout randomized clinical trials (RCTs). The numbers on the lines indicate the number of RCTs informing a particular comparison.

direct comparisons ( $n = 4$  trials) were between sumatriptan and rizatriptan (the 2 most commonly tested treatments). Of these, 15 comparisons were informed direct evidence, but 7 of the direct connections had only 1 trial, and several of the comparisons were informed only by indirect evidence. Frovatriptan was poorly connected to other treatments, and all comparisons that involved this agent warranted, therefore, only moderate confidence at best.

Sixty-three trials reported the outcome of pain-free response at 2 hours, and 25 reported 24 hours of sustained pain-free response. The authors used the  $I^2$  value to assess heterogeneity in pairwise meta-analysis before conducting their NMA; however, they did not report the specific values. They checked the coherence between direct and indirect comparisons from closed loops and provided this information as a supplemental appendix online. Direct and indirect evidence were consistently similar, with no statistical evidence of incoherence (Table 16-1). The authors also conducted several sensitivity analyses to assess the role of dose.

Figure 16-4 displays the results of the NMA of triptans vs placebo. For pain-free response at 2 hours, the authors found that eletriptan, sumatriptan, and rizatriptan exhibited the largest treatment effects against placebo. The results were largely similar for pain-free response at 24 hours.

When the authors examined the comparative effectiveness of each triptan vs the other triptans, evidence warranted at least moderate confidence for some differences among triptans. For example, eletriptan was superior in pain-free response at 2 hours compared with sumatriptan (OR, 1.53; 95% CI, 1.16-2.01), almotriptan (OR, 2.03; 95% CI, 1.38-2.96), zolmitriptan (OR, 1.46; 95% CI, 1.02-2.09), and naratriptan (OR, 2.95; 95% CI, 1.78-4.90).

For all but naratriptan, we have at least moderate confidence in treatment effects vs placebo at 2 and 24 hours. Eletriptan was associated with the largest probability (68%) of being the best treatment for pain-free response at 2 and 24 hours (54.1%). The only other drug that ranked favorably was rizatriptan (22.6% at 2 hours and 9.2% at 24 hours). Given that comparisons between eletriptan and a number of other agents warrant at least moderate confidence, the first rank of eletriptan carried considerable weight.



TABLE 16-1

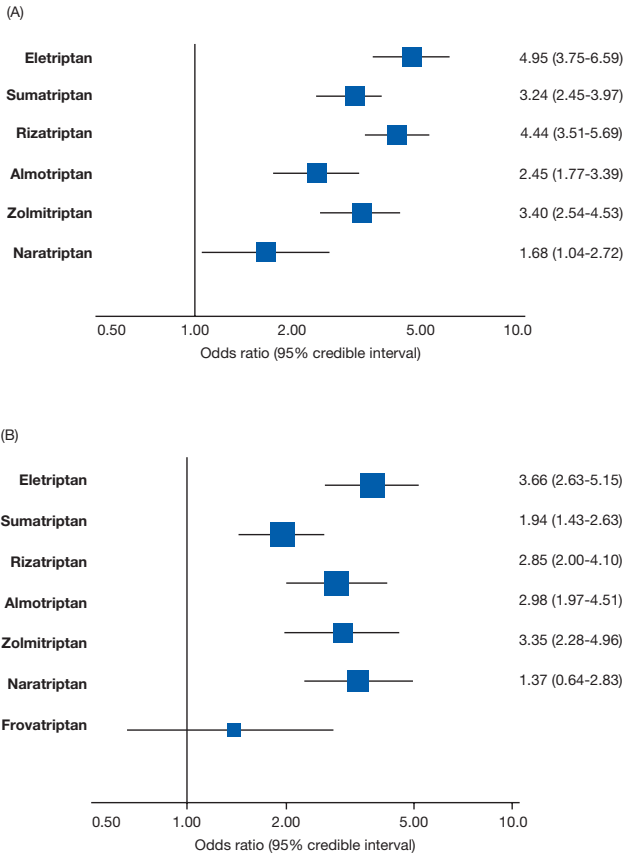
**Consistency Check for a Pain-Free Response at 2 Hours With Triptan in Usual Doses**

Comparison	No. of Trials	Direct Estimate <sup>a</sup>	Indirect Estimate <sup>a</sup>
<b>Three-Treatment Loops Where Inconsistency Can be Checked</b>			
Eletriptan (40 mg) vs sumatriptan (50 mg)	2	1.48 (1.14-2.79)	1.58 (0.60-5.87)
Eletriptan (40 mg) vs zolmitriptan (12.5 mg)	2	1.52 (0.96-1.81)	1.21 (0.35-3.55)
Eletriptan (40 mg) vs naratriptan (2.5 mg)	1	2.46 (1.53-3.98)	2.75 (0.37-19.8)
Sumatriptan (50 mg) vs almotriptan (2.5 mg)	1	1.49 (1.12-1.98)	1.07 (0.63-1.76)
Sumatriptan (50 mg) vs zolmitriptan (12.5 mg)	1	1.12 (0.87-1.45)	0.72 (0.42-1.29)
Sumatriptan (50 mg) vs frovatriptan (2.5 mg)	1	1.07 (0.56-2.04)	0.64 (0.35-1.15)
Almotriptan (2.5 mg) vs zolmitriptan (12.5 mg)	1	0.89 (0.69-1.15)	0.70 (0.41-1.19)
Zolmitriptan (12.5 mg) vs frovatriptan (2.5 mg)	1	0.73 (0.52-1.02)	0.86 (0.47-1.62)
Naratriptan (12.5 mg) vs frovatriptan (2.5 mg)	1	0.82 (0.51-1.20)	0.90 (0.49-1.79)

<sup>a</sup>Odds ratio estimates and 95% confidence intervals for all treatment comparisons from the direct pairwise meta-analysis of head-to-head trials and indirect comparison meta-analysis using placebo as the common comparator.

FIGURE 16-4

### Forest Plot of the Primary Multiple-Treatment Comparison Meta-analysis Results, Triptans vs Placebo



A, Pain-free response at 2 hours; B, 24 hours of sustained pain-free response.

## HOW CAN I APPLY THE RESULTS TO PATIENT CARE?

---

### Were All Patient-Important Outcomes Considered?

Many NMAs report only 1 or a few outcomes of interest. For example, a recent NMA that compared the efficacy of antihypertensive treatments reported only heart failure and mortality,<sup>40</sup> whereas an older NMA of antihypertensive treatments also considered coronary heart disease and stroke.<sup>41</sup> Adverse events are infrequently assessed in meta-analysis and in NMAs, reflecting poor reporting in the *primary studies*.<sup>42,43</sup> Network meta-analyses conducted in the context of health technology assessment submissions and *evidence-based practice* reports are more likely to include multiple outcomes and assessments of harms than the less lengthy NMAs published in clinical medical journals.<sup>20</sup>

### USING THE GUIDE

**T**he authors assessed outcomes (pain-free response at 2 and 24 hours) that are important to patients. The major omission is adverse events—if triptans differed substantially in adverse events, this would be an important consideration for patients. Fortunately, the drug that appears as or more effective than other triptans, eletriptan, also appears to be at least as well tolerated as other triptans.<sup>44</sup>

### Were All Potential Treatment Options Considered?

Network meta-analyses may place restrictions on what treatments are examined. For example, for irritable bowel syndrome, an NMA may focus on pharmacologic agents, neglecting RCTs of diet, peppermint oil, and counseling.<sup>45</sup> Decisions to focus on subclasses of drugs may also be problematic. For example, in

rheumatoid arthritis, biologics are used for patients in whom conventional drugs fail. Five of the 9 available biologics are anti-tumor necrosis factor (TNF) agents. One recent NMA only considered anti-TNF agents and excluded other biologics.<sup>46</sup> To the extent that the other biologic agents are equivalent or superior to the anti-TNF agents, their exclusion risks misleading clinicians regarding the best biologic agents.

### Are Any Postulated Subgroup Effects Credible?

There are very few situations in which investigations have convincingly established important differences in the relative effect of treatment according to patient characteristics.<sup>47</sup> Criteria exist for determining the credibility of subgroup analyses.<sup>47</sup> These criteria include whether the comparisons are within-study (subgroup A and subgroup B both participated in the same study, the stronger comparison) or between-study (one study enrolled subgroup A and another subgroup B, the weaker comparison), chance is an unlikely explanation of the differences in effect between subgroups, and the investigators made a small number of a priori subgroup hypotheses with an accurately specified direction. Network meta-analyses allow a greater number of RCTs to be evaluated and may offer more opportunities for subgroup analysis—but with due skepticism and respect for credibility criteria.

For example, in an NMA that examined inhaled drugs for COPD, the authors examined whether severity of airflow obstruction measured by forced expiratory volume in 1 second ( $FEV_1$ ) influenced patients' response.<sup>48</sup> If the  $FEV_1$  was 40% or less of predicted, long-acting anticholinergics, inhaled corticosteroids, and combination treatment, including inhaled corticosteroids, reduced exacerbations significantly compared with long-acting  $\beta$ -agonists alone

but not if the FEV<sub>1</sub> was greater than 40% of predicted. This difference was significant for inhaled corticosteroids ( $P = .02$  for interaction) and combination treatment ( $P = .01$ ) but not for long-acting anticholinergics ( $P = .46$ ). The fact that these analyses were based on an a priori hypothesis, including a correctly hypothesized direction with a strong biologic rationale (greater inflammation in more severe airway disease) and a low  $P$  value for the test of interaction (ie, chance is an unlikely explanation), strengthens the credibility of the subgroup effect. It is, however, based on a between-group comparison. A reasonable judgment would be moderate to high credibility of the subgroup effect, and a clinical policy of restricting inhaled corticosteroid use to patients with more severe airflow obstruction.

## CLINICAL SCENARIO RESOLUTION

**Y**ou conclude that there is convincing evidence for the role of triptans in aborting migraine headaches at 2 and 24 hours. However, because triptans are a class of drugs you choose to assess whether this *class effect* is real or not. There are data available from direct and indirect comparisons that suggest that eletriptan is superior to several other triptans. You opt to discuss with the patient the benefits of starting treatment with eletriptan and will seek evidence for adverse events.

## CONCLUSION

---

Although an NMA can provide extremely valuable information in choosing among multiple treatments offered for the same condition, it is important to determine the confidence one can

place in the estimates of effect of the treatments considered and the extent to which that confidence differs across comparisons. If authors provide these confidence ratings themselves using criteria such as those suggested by *GRADE (Grading of Recommendations Assessment, Development and Evaluation)*, the task is straightforward—simply survey the confidence ratings. Those rated as high or moderate are trustworthy and those rated low or very low much less so. If the authors do not provide these ratings themselves, you need to make your own assessments, which can be challenging.

The confidence for any comparison will be greater if individual studies are at low risk of bias and publication bias is unlikely; results are consistent in individual direct comparisons and individual comparisons with no-treatment controls and also consistent between direct and indirect comparisons; sample size is large and CIs are correspondingly narrow; and most comparisons have some direct evidence. If all of these hallmarks are present and the differences in effect sizes are large, high confidence in estimates may be warranted. However, in most cases, confidence in some key estimates is likely to warrant only moderate or low confidence. Most concerning, if authors do not provide the necessary information, it is difficult to judge which comparisons are trustworthy and which less so—and in such cases, clinicians may be best served by reviewing systematic reviews and meta-analyses of the direct comparisons and using these to guide their patient management.

## References

1. Thorlund K, Mills EJ, Wu P, et al. Comparative efficacy of triptans for the abortive treatment of migraine: a multiple treatment comparison meta-analysis. *Cephalalgia*. 2014;34(4):258-267.
2. Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. *Lancet*. 1998;351(9096):123-127.
3. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. *N Engl J Med*. 1987;316(8):450-455.

4. Estellat C, Ravaud P. Lack of head-to-head trials and fair control arms: randomized controlled trials of biologic treatment for rheumatoid arthritis. *Arch Intern Med*. 2012;172(3):237-244.
5. Lathyris DN, Patsopoulos NA, Salanti G, Ioannidis JP. Industry sponsorship and selection of comparators in randomized clinical trials. *Eur J Clin Invest*. 2010;40(2):172-182.
6. Salanti G, Higgins JP, Ades AE, Ioannidis JP. Evaluation of networks of randomized trials. *Stat Methods Med Res*. 2008;17(3):279-301.
7. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. *Stat Med*. 2004;23(20):3105-3124.
8. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J Clin Epidemiol*. 1997;50(6):683-691.
9. Song F, Xiong T, Parekh-Bhurke S, et al. Inconsistency between direct and indirect comparisons of competing interventions: meta-epidemiological study. *BMJ*. 2011;343:d4909.
10. Mills EJ, Bansback N, Ghement I, et al. Multiple treatment comparison meta-analyses: a step forward into complexity. *Clin Epidemiol*. 2011;3:193-202.
11. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Ann Intern Med*. 2009;151(4):W65-W94.
12. Sutton A, Ades AE, Cooper N, Abrams K. Use of indirect and mixed treatment comparisons for technology assessment. *Pharmacoeconomics*. 2008;26(9):753-767.
13. Kyrgiou M, Salanti G, Pavlidis N, Paraskevaidis E, Ioannidis JP. Survival benefits with diverse chemotherapy regimens for ovarian cancer: meta-analysis of multiple treatments. *J Natl Cancer Inst*. 2006;98(22):1655-1663.
14. Mills EJ, Kanters S, Thorlund K, Chaimani A, Veroniki AA, Ioannidis JP. The effects of excluding treatments from network meta-analyses: survey. *BMJ*. 2013;347:f5195.
15. Cipriani A, Furukawa TA, Salanti G, et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet*. 2009;373(9665):746-758.
16. Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. *N Engl J Med*. 2008;358(3):252-260.
17. Ioannidis JP. Effectiveness of antidepressants: an evidence myth constructed from a thousand randomized trials? *Philos Ethics Humanit Med*. 2008;3:14.
18. Higgins JP, Whitehead A. Borrowing strength from external trials in a meta-analysis. *Stat Med*. 1996;15(24):2733-2749.
19. Ioannidis JP. Ranking antidepressants. *Lancet*. 2009;373(9677):1759-1760, author reply 1761-1762.
20. Gartlehner G, Hansen RA, Morgan LC, et al. Comparative benefits and harms of second-generation antidepressants for treating major depressive disorder: an updated meta-analysis. *Ann Intern Med*. 2011;155(11):772-785.

21. Nixon RM, Bansback N, Brennan A. Using mixed treatment comparisons and meta-regression to perform indirect comparisons to estimate the efficacy of biologic treatments in rheumatoid arthritis. *Stat Med*. 2007;26(6):1237-1254.
22. Mills EJ, Wu P, Chong G, et al. Efficacy and safety of statin treatment for cardiovascular disease: a network meta-analysis of 170,255 patients from 76 randomized trials. *QJM*. 2011;104(2):109-124.
23. Mills EJ, Wu P, Lockhart I, Thorlund K, Puhan M, Ebbert JO. Comparisons of high-dose and combination nicotine replacement therapy, varenicline, and bupropion for smoking cessation: a systematic review and multiple treatment meta-analysis. *Ann Med*. 2012;44(6):588-597.
24. Salanti G, Kavvoura FK, Ioannidis JP. Exploring the geometry of treatment networks. *Ann Intern Med*. 2008;148(7):544-553.
25. Mills EJ, Ghement I, O'Regan C, Thorlund K. Estimating the power of indirect comparisons: a simulation study. *PLoS One*. 2011;6(1):e16237.
26. Davey-Smith G, Egger MG. Going beyond the grand mean: subgroup analysis in meta-analysis of randomised trials. In: *Systematic Reviews in Health Care: Meta-analysis in context*. 2nd ed. London, England: BMJ Publishing Group; 2001:143-156.
27. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? *Stat Med*. 2002;21(11):1559-1573.
28. Jansen J, Schmid C, Salanti G. When do indirect and mixed treatment comparisons result in invalid findings? A graphical explanation. 19th Cochrane Colloquium Madrid, Spain October 19-22, 2011. 2011:P3B379.
29. Song F, Harvey I, Lilford R. Adjusted indirect comparison may be less biased than direct comparison for evaluating new pharmaceutical interventions. *J Clin Epidemiol*. 2008;61(5):455-463.
30. Lu G, Ades A. Assessing evidence inconsistency in mixed treatment comparisons. *J Am Stat Assoc*. 2006;101(474):447-459.
31. Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. *Stat Med*. 2010;29(7-8):932-944.
32. Zhang WY, Li Wan Po A. Analgesic efficacy of paracetamol and its combination with codeine and caffeine in surgical pain—a meta-analysis. *J Clin Pharm Ther*. 1996;21(4):261-282.
33. Salanti G, Ades AE, Ioannidis JP. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *J Clin Epidemiol*. 2011;64(2):163-171.
34. Goufopoulos V, Salanti G, Pavlidis N, Ioannidis JP. Survival and disease-progression benefits with treatment regimens for advanced colorectal cancer: a meta-analysis. *Lancet Oncol*. 2007;8(10):898-911.
35. Diels J, Cure S, Gavart S. The comparative efficacy of telaprevir versus boceprevir in treatment-naive and treatment experienced patients with genotype 1 chronic hepatitis C virus infection: a mixed treatment comparison analysis. Paper presented at: 14th Annual International Society for Pharmaceutical Outcomes Research (ISPOR) European Congress; November 5-8, 2011; Madrid, Spain.



36. Diels J, Cure S, Gavart S. The comparative efficacy of telaprevir versus boceprevir in treatment-naive and treatment-experienced patients with genotype 1 chronic hepatitis. *Value Health*. 2011;14(7):A266.
37. Higgins JP, Green S. Analysing data and undertaking meta-analyses. In: *Cochrane Handbook for Systematic Reviews of Interventions*. Oxford: Wiley & Sons; 2008.
38. Aaron SD, Fergusson D, Marks GB, et al; Canadian Thoracic Society/Canadian Respiratory Clinical Research Consortium. Counting, analysing and reporting exacerbations of COPD in randomised controlled trials. *Thorax*. 2008;63(2):122-128.
39. Mills EJ, Druyts E, Ghement I, Puhan MA. Pharmacotherapies for chronic obstructive pulmonary disease: a multiple treatment comparison meta-analysis. *Clin Epidemiol*. 2011;3:107-129.
40. Sciarretta S, Palano F, Tocci G, Baldini R, Volpe M. Antihypertensive treatment and development of heart failure in hypertension: a Bayesian network meta-analysis of studies in patients with hypertension and high cardiovascular risk. *Arch Intern Med*. 2011;171(5):384-394.
41. Psaty BM, Lumley T, Furberg CD, et al. Health outcomes associated with various antihypertensive therapies used as first-line agents: a network meta-analysis. *JAMA*. 2003;289(19):2534-2544.
42. Hernandez AV, Walker E, Ioannidis JP, Kattan MW. Challenges in meta-analysis of randomized clinical trials for rare harmful cardiovascular events: the case of rosiglitazone. *Am Heart J*. 2008;156(1):23-30.
43. Ioannidis JP, Evans SJ, Gøtzsche PC, et al; CONSORT Group. Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Ann Intern Med*. 2004;141(10):781-788.
44. Bajwa Z, Sabahat A. Acute treatment of migraine in adults. UpToDate website. <http://www.uptodate.com/contents/acute-treatment-of-migraine-in-adults>. Accessed August 4, 2014.
45. Ford AC, Talley NJ, Spiegel BM, et al. Effect of fibre, antispasmodics, and peppermint oil in the treatment of irritable bowel syndrome: systematic review and meta-analysis. *BMJ*. 2008;337:a2313.
46. Schmitz S, Adams R, Walsh CD, Barry M, FitzGerald O. A mixed treatment comparison of the efficacy of anti-TNF agents in rheumatoid arthritis for methotrexate non-responders demonstrates differences between treatments: a Bayesian approach. *Ann Rheum Dis*. 2012;71(2):225-230.
47. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ*. 2010;340:c117.
48. Puhan MA, Bachmann LM, Kleijnen J, Ter Riet G, Kessels AG. Inhaled drugs to reduce exacerbations in patients with chronic obstructive pulmonary disease: a network meta-analysis. *BMC Med*. 2009;7:2.

# How to Use a Patient Management Recommendation: Clinical Practice Guidelines and Decision Analyses

Ignacio Neumann, Elie A. Akl, Per Olav Vandvik, Thomas Agoritsas, Pablo Alonso-Coello, David M. Rind, Nancy Santesso, Paul Elias Alexander, Reem A. Mustafa, Kameshwar Prasad, Shannon M. Bates, Holger J. Schünemann, and Gordon Guyatt

## IN THIS CHAPTER

---

### Clinical Scenario

#### Developing Recommendations

Practice Guidelines

Decision Analysis

#### Assessing Recommendations

Is the Clinical Question Clear and Comprehensive?

(continued on following page)

Were the Recommendations Based on the Current Best Evidence?

Are Values and Preferences Appropriately Specified for Each Outcome?

Do the Authors Indicate the Strength of Their Recommendations?

Is the Evidence Supporting the Recommendations Easily Understood?

Was the Influence of Conflict of Interests Minimized?

**How Should You Use Recommendations?**

Strong Recommendations

Weak Recommendations

**Clinical Scenario Resolution**

---

## CLINICAL SCENARIO

**Y**ou are an obstetrician seeing a 31-year-old pregnant woman who had an unprovoked deep venous thrombosis of the leg 5 years ago that was treated with warfarin for 6 months without complication. She is no longer using antithrombotic medication and is otherwise healthy. Given a possible increased risk of thrombosis with pregnancy, you are considering discussing the possibility of low-molecular-weight heparin (LMWH) prophylaxis for the rest of the pregnancy.

To inform your discussion, you search first for an evidence-based recommendation and find the following recommendation from a practice guideline<sup>1</sup>: “For pregnant women at moderate to high risk of recurrent venous thromboembolism (VTE) (single unprovoked VTE, pregnancy- or estrogen-related VTE, or multiple prior unprovoked VTE not receiving long-term anticoagulation), we suggest antepartum prophylaxis with prophylactic- or intermediate-dose LMWH rather than clinical vigilance or routine care (weak recommendation, based on low confidence in effect estimates).”

The statement “weak recommendation, based on low confidence in effect estimates” leaves you uncomfortable. You decide to read further to understand the recommendation and its rationale.

## DEVELOPING RECOMMENDATIONS

In general, patient management recommendations are developed in the context of *clinical practice guidelines* (see Chapter 4, Finding Current Best Evidence). However, you also may find guidance originating from a *decision analysis*. Similar criteria of credibility apply to both approaches.<sup>2-5</sup>

### Practice Guidelines

Practice guidelines are statements that include recommendations intended to optimize patient care. They are, ideally,

informed by a *systematic review of evidence* and an assessment of the benefits and *harms* of alternative care options.<sup>2</sup> To make a recommendation, guideline panelists must define clinical questions, select the relevant *outcome variables*, retrieve and synthesize all of the relevant evidence, rate the confidence in the effect estimates, and, relying on a systematic approach but ultimately also on consensus, move from evidence to recommendations.<sup>6</sup> To fully inform their audience, guideline panels should provide not only their recommendations but also the key information on which their recommendations are based.

## Decision Analysis

Decision analysis is a formal method that integrates the evidence regarding the beneficial and harmful effects of treatment options with the *values or preferences* associated with those effects. Clinical decision analyses are built as structured approaches (*decision trees*), and authors will usually include 1 or more diagrams showing the structure of the decision trees used for the analysis.

Figure 17-1 shows a simplified decision tree for the scenario of the pregnant woman considering thromboprophylaxis. The patient has 2 options: to use or not use prophylaxis with LMWH. The decision is represented by a square, termed “decision node.” The lines that emanate from the decision node represent the clinical strategies under consideration.

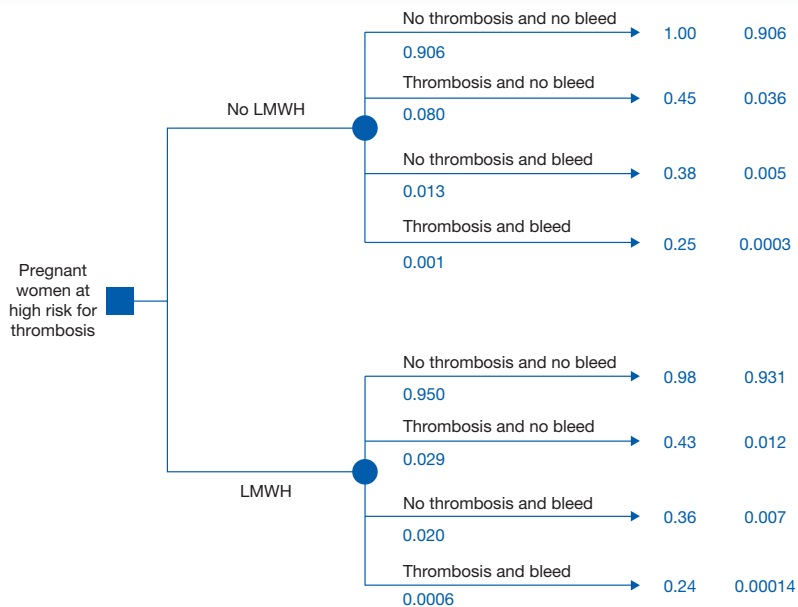
Circles, called “chance nodes,” symbolize the different events that can occur after each clinical strategy. Patients may or may not develop a thrombotic or bleeding event, and the decision analysis requires estimates of the probability of both events. Triangles or rectangles identify outcome states.

The decision analysis also addresses the extent to which each of the outcome events is desirable (no bleeding or thrombotic event) or undesirable (either adverse event) (in technical language, the *utility*). The combination of the probabilities and utilities allows the decision analyst to determine the relative value of each management option.



**FIGURE 17-1**

**Diagram of a Simplified Decision Tree**



Abbreviation: LMWH, low-molecular-weight heparin.



The process of decision analysis makes fully explicit all of the elements of the decision so that they are open for debate and modification.<sup>7</sup> When a decision analysis includes costs among the outcomes, it becomes an *economic analysis* and summarizes trade-offs between health changes and resource expenditure.

## EXAMPLE OF A DECISION TREE

---

Returning to Figure 17-1, each arm of the decision (no prophylaxis vs LMWH) has 1 chance node at which 4 possible outcomes could occur (the 4 possible combinations arising from bleeding or not bleeding and from having a thrombosis or not having a thrombosis). The figure depicts the probabilities associated with the decision. In the no-prophylaxis strategy, patients would have a probability of bleeding and having a thrombosis of 0.1%, a probability of bleeding and not having a thrombosis of 1.3%, a probability of not bleeding but having a thrombosis of 8%, and a probability of not bleeding and not having a thrombosis of 90.6%. With the LMWH prophylaxis strategy, the probability of bleeding and having a thrombosis is 0.06%, the probability of bleeding and not having a thrombosis is 2%, the probability of not bleeding but having a thrombosis is 2.9%, and the probability of not bleeding and not having a thrombosis is 95%.<sup>1,8</sup>

The figure also presents the values associated with each health state on a scale of 0 to 1, with 1 representing the utility of full health and 0 representing the utility of death. In the no-prophylaxis strategy, the health state without any negative outcome (no thrombosis or bleeding) represents full health, a utility of 1.0. The occurrence of a thrombosis or bleeding event decreases the value of

the health state to 0.45 in the case of thrombosis and to 0.38 in the case of bleeding. When both negative outcomes occur at the same time, the corresponding utility is even lower: 0.25. In the LMWH arm, the addition of the *burden of treatment* slightly decreases the utility of the 4 health states.

The final step in the decision analysis is to calculate the total expected value—the sum of the probabilities and utilities associated with each outcome—for each possible course of action. Given the particular set of probabilities and utilities we have presented, the estimated value of the no-prophylaxis branch would be  $(0.906 \times 1.0) + (0.080 \times 0.45) + (0.013 \times 0.38) + (0.001 \times 0.25)$ , which is 0.947. The value of the LMWH branch would be  $(0.950 \times 0.98) + (0.029 \times 0.43) + (0.020 \times 0.36) + (0.0006 \times 0.24)$ , which is 0.950. In this example, the prophylaxis strategy is more desirable, but the difference in the expected values between the 2 options—called “relative utility”—is relatively small.

The model presented in Figure 17-1 is oversimplified in a number of ways. For example, it does not take into account the possibility of fatal events or potential long-term morbidity (eg, after an intracranial bleeding or the development of postthrombotic syndrome). Also, it does not consider the time in the health states. For instance, having a major bleeding without any complication may appreciably reduce the utility during the episode, but almost all patients will return to a perfect health state relatively quickly. *Multistate transition models* using simulation—termed *Markov models*—permit analyses that are closer to real life. For example, an analysis using multistate transition models concluded that for patients like the one presented in the opening scenario and in the decision tree (high risk for VTE recurrence), antepartum prophylaxis with LMWH is a cost-effective use of resources.<sup>9</sup>



## ASSESSING RECOMMENDATIONS

---

Box 17-1 presents our guidance for determining the extent to which a guideline or decision analysis will provide trustworthy recommendations.

### Is the Clinical Question Clear and Comprehensive?

The most useful patient management recommendations from guidelines and decision analyses will use a standardized format that details precisely the recommended actions, the alternatives with which they are compared, to whom they apply, and under what circumstances.

#### BOX 17-1

### Users' Guides for Assessing Treatment Recommendations

#### Is the clinical question clear and comprehensive?

Is the recommended intervention clear and actionable?

Is the alternative clear?

Were all of the relevant outcomes important to patients explicitly considered?

#### Was the recommendation based on the best current evidence?

Are values and preferences associated with the outcomes appropriately specified?

Do the authors indicate the strength of their recommendations?

Is the evidence supporting the recommendation easily understood?

For strong recommendations, is the strength appropriate?

For weak recommendations, does the information provided facilitate shared decision making?

Was the influence of conflict of interests minimized?

### **Is the Recommended Intervention Clear and Actionable?**

Recommendations are sometimes too vague to be helpful. Consider, for instance, this recommendation from a clinical practice guideline<sup>10</sup>: “For both outpatients and inpatients with diabetic foot infection, clinicians should attempt to provide a well-coordinated approach by those with expertise in a variety of specialties, preferably by a multidisciplinary diabetic foot care team.” What remains unclear in this recommendation is the level of obligation in the “attempt,” what is involved in making care “well-coordinated,” and which specialties are included in the “variety.”

In contrast, another guideline from the National Foundation for Health Care Excellence<sup>11</sup> makes clear what is being recommended: “We recommend that a multidisciplinary foot care team manage the care of patients with diabetic foot problems who require inpatient care. The multidisciplinary foot care team should include a diabetologist, a surgeon with the relevant expertise, a diabetes nurse specialist, a podiatrist and a tissue viability nurse.”

### **Is the Alternative Clear?**

When guideline panelists develop recommendations, they choose a specific course of action over others. If the alternative is not clear, the significance of the recommendation will remain obscure. For example, in the recommendation “Uterine massage is recommended for the treatment of postpartum hemorrhage,”<sup>12</sup> the absence of an explicit alternative may introduce challenges in the interpretation. Are the panelists suggesting performing uterine massage as a first-line treatment in preference to other therapeutic measures, or are they recommending it in addition to other concomitant measures? By comparing the recommendation with others within the guideline, it is possible to infer that panelists meant that uterine massage should be used in addition to other measures and not as a single intervention, but recommendation statements should be clear enough to be interpreted without having to read the full guideline.

In contrast, the recommendation “We recommend isotonic crystalloids ... in preference to ... colloids for the initial intravenous fluid resuscitation of women with postpartum hemorrhage”<sup>12</sup> offers a clearer message by making the alternative explicit.

As you may have noticed, in both recommendations regarding the management of diabetic foot problems presented in the previous section, the *control group* is not clearly defined. Although the option of “no foot care team” seems to be the implicit comparator, it is not clear what this management strategy entails.

Clinicians who use a decision analysis will not face the problem of ambiguous alternatives because the options in comparison are explicit.

### Were All of the Relevant Outcomes Important to Patients Explicitly Considered?

The balance between the benefits and the harms of the interventions will depend on what outcomes are considered. Clinicians should judge whether the guideline panel or the decision analysts included all *patient-important outcomes*.

For example, the eighth edition of the antithrombotic guidelines (AT8) of the American College of Chest Physicians (ACCP) recommended the use of elastic stockings for patients with stroke who have contraindications to anticoagulants.<sup>13</sup> The 9th edition of the antithrombotic guidelines (AT9) suggested against its use.<sup>14</sup> Both guideline panels considered the outcomes of mortality, pulmonary embolism, and symptomatic deep venous thrombosis, but AT9 panelists also considered that elastic stockings produce a 4-fold increase in the risk of skin complications: 39 more per 1000 patients treated for 1 month (95% *confidence interval* [CI], 17-77 more per 1000).<sup>15</sup> The additional consideration of skin complications is responsible for the change in recommendations.

Outcomes typically considered as patient important include mortality, morbidity (eg, major bleeding, acute exacerbation of a chronic disease, hospital admission), and patient-reported outcomes (eg, quality of life, functional status). *Surrogate outcomes* (eg, lipid levels, bone density, cognitive function tests) are variably associated with patient-important outcomes but are never important in and of themselves.

In addition, AT8 suggested international normalized ratio (INR) monitoring at an interval of no longer than every 4 weeks in patients treated with vitamin K antagonists.<sup>16</sup> This recommendation was primarily based on studies that found that frequent monitoring increased the time in therapeutic INR range—a surrogate outcome. However, AT9 suggested an INR testing frequency of up to 12 weeks rather than every 4 weeks.<sup>17</sup> This recommendation was based on studies that found no increase in thrombotic events or major bleeding with monitoring every 12 weeks. Both recommendations were based on explicitly defined outcomes. However, the outcomes were surrogate in the first case and—more appropriately—patient important in the second.

Outcomes not plausibly influenced by the intervention are typically not relevant for decision making and therefore may not be considered. For example, mortality is a very important outcome; however, it is not relevant for the decision of whether to use intranasal antihistamines for the treatment of allergic rhinitis because the intervention does not plausibly affect the probability of dying.

### **Were the Recommendations Based on the Current Best Evidence?**

Guideline panelists and decision analysts should base their estimates of the benefits and harms of the intervention and their

evaluation of the associated confidence in effect estimates on current or updated systematic reviews, preferably those that include meta-analysis. In the absence of such meta-analytic systematic reviews, guideline panelists may conduct their own reviews or provide less systematic evidence summaries. Clinicians should look for a description of the process used to identify and summarize the relevant evidence and should judge to what extent this process is credible. Clinicians also should check the date on which the literature search was conducted (see Chapter 14, The Process of a Systematic Review and Meta-analysis).

Recommendations that do not use the best current evidence risk promoting suboptimal or even harmful care. For example, for several years guideline panels ignored a substantial body of evidence that suggested the effectiveness of prophylaxis with quinolones in patients with postchemotherapy neutropenia.<sup>18</sup> Only in its 2010 guidelines did the Infectious Diseases Society of America suggest the prophylactic use of antibiotics in this population.<sup>19</sup> This highlights the necessity for rapid and sometimes frequent updating of guidelines in areas under active investigation (see Chapter 4, Finding Current Best Evidence).

### **Are Values and Preferences Appropriately Specified for Each Outcome?**

Assessing treatment effects on outcomes is largely a question of measurement and a matter of science. Assigning preferences to outcomes is a matter of values. Consider, for example, the outcomes associated with routine mammographic screening in women aged 40 to 49 years: there is a very small and questionable reduction of breast cancer mortality and a relatively high probability of a false-positive result (which typically leads to unnecessary follow-up testing and sometimes to unnecessary biopsy of the breast).<sup>20</sup> A guideline panel must consider the value attached to each of these 2 outcomes when trading them off to develop a recommendation. A panel that assigns a higher value to the very small reduction in cancer mortality would support the screening, whereas a panel that assigns a higher value

to avoiding unnecessary procedures would not. Consequently, clinicians should look for explicit statements regarding the values and preferences used to inform the recommendation.

Whose values should drive recommendations? Under ideal circumstances, recommendations should be based on a systematic review of relevant studies exploring patients' values and preferences<sup>21</sup>; unfortunately, such evidence is still rare. In the absence of a body of empirical evidence about patients' values and preferences, guideline panels or decision analysts may fall back on the experience of clinicians who regularly engage in shared decision making. Another alternative is the involvement of representative patients and consumers in the recommendation development process.<sup>22</sup> However, ensuring that those involved—clinicians or patients—will be able to represent typical patients is challenging and perhaps only partly achievable.

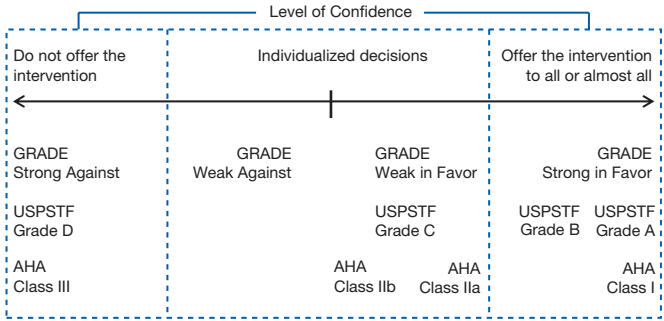
Whatever the source of values and preferences, it is possible to make them explicit and transparent. Unfortunately, failure to do so remains the most common serious deficit in current practice guidelines. In contrast, decision analysis requires explicit and quantitative specification of values because each outcome is assigned a given health utility. However, although the values and preferences in a decision analysis may be explicit, their source may be problematic. For example, a systematic review of 54 *cost-utility analyses* (including 45 decision analyses) in child health found that the source used for valuing health states was the authors' own judgment in 35% of the analyses, and in another 11% the source of values and preferences was not stated.<sup>23</sup>

### Do the Authors Indicate the Strength of Their Recommendations?

Trustworthy recommendations should specify the strength of the recommendations and also a rating of the confidence in effect estimates that support the recommendations (also known as quality of evidence).<sup>2</sup> *Sensitivity analyses* are used to explore the strength of the conclusions that arise from a decision analysis.

**FIGURE 17-2**

**Direction and Strength of Recommendations in Different Grading Systems**



Abbreviations: AHA, American Heart Association; GRADE, Grading of Recommendations Assessment, Development and Evaluation; USPSTF, US Preventive Services Task Force.

**Grades of Recommendation**

There are dozens of grading systems for recommendations.<sup>24</sup> However, the 3 most commonly used approaches are *GRADE* (*Grading of Recommendations Assessment, Development and Evaluation*)<sup>25</sup> and those used by the American Heart Association (AHA)<sup>26</sup> and the US Preventive Services Task Force (USPSTF).<sup>27</sup> A detailed discussion of the differences among these systems is beyond the scope of this chapter; we will, however, mention 2 important similarities.

The 3 systems feature a rating for confidence in effect estimates (ie, quality of evidence). Confidence in the effect estimates represents the extent to which the estimates are sufficiently credible to support a particular recommendation (Figure 17-2). The GRADE approach specifies 4 levels of confidence: high, moderate, low, and very low (see Chapter 15, Understanding and Applying the Results of a Systematic Review and Meta-analysis).

The AHA and USPSTF systems specify 3 levels of confidence: A, B, and C in the AHA approach and high, moderate, and low in the USPSTF approach.

The 3 systems share another critical feature: they differentiate between recommendations that should be applied (or avoided) in all, or almost all, patients (ie, strong recommendations) from those that require individualization to the patient's values, preferences, and circumstances (ie, weak recommendations) (Figure 17-2).

### **Sensitivity Analysis**

Decision analysts use sensitivity analyses, the systematic exploration of the uncertainty in the data, to vary estimates for downsides, benefits, and values and to determine the impact of these varying estimates on expected outcomes. Sensitivity analysis asks the question: to what extent is the relative utility of the alternatives affected by the uncertainties in the estimates of the likelihood or value of the outcomes? To the extent that the result of the decision analysis does not change with varying probability estimates and varying values, clinicians can consider the recommendation a strong one. When the final decision shifts with different plausible values of probabilities or values, the conclusion becomes much weaker: the right choice may differ given the true probabilities, and patients' choices are likely to vary according to their preferences.

## **Is the Evidence Supporting the Recommendations Easily Understood?**

### **For Strong Recommendations, Is the Strength Appropriate?**

The message to the clinician from strong recommendations is "just do it." Recommendations that are inappropriately graded as strong may therefore have substantial undesirable consequences.

High confidence in the effect estimates will support a strong recommendation if the desirable consequences considerably outweigh the undesirable consequences, if there is reasonable confidence and limited variability in patients' values and preferences,



and if the benefits of the proposed course of action justify its cost. When there is substantial uncertainty regarding the effects of the intervention (low confidence in the effect estimates), clinicians should generally expect weak recommendations.

Sometimes, guideline panels can appropriately offer strong recommendations despite low or very low confidence in effect estimates. [Table 17-1](#) presents 5 paradigmatic situations in which this can occur. Clinicians should carefully examine a strong recommendation based on low or very low confidence. If it does

TABLE 17-1

### Five Paradigmatic Situations That Justify Strong Recommendations Based on Low or Very Low Confidence

Paradigmatic Situation	Confidence in Effect Estimates for Health Outcomes (Quality of Evidence)		Balance of Benefits and Harms	Values and Preferences
	Benefits	Harms		
Life-threatening situation	Low or very low	Immaterial (very low to high)	Intervention may reduce mortality in a life-threatening situation; adverse events not prohibitive	A very high value is placed on an uncertain but potentially life-preserving benefit
Uncertain benefit, certain harm	Low or very low	High or moderate	Possible but uncertain benefit; substantial established harm	A much higher value is placed on the adverse events in which we are confident than in the benefit, which is uncertain

not correspond to any of the situations listed in Table 17-1, it is likely that the recommendation was inappropriately graded.

For example, a systematic survey of the Endocrine Society guidelines issued between 2005 and 2011 found that 121 of the total of 357 recommendations identified were strong recommendations based on low or very low confidence in effect estimates. Of these 121, only 35 (29%) were consistent with one of the situations presented in Table 17-1 and thus clearly appropriate.<sup>31</sup> This result highlights the need for caution when facing strong recommendations based on low or very low confidence in effect estimates.

<b>Resource Considerations</b>	<b>Recommendation</b>	<b>Example</b>
Small incremental cost (or resource use) relative to the benefits justify the intervention	Strong recommendation in favor	Indirect evidence from seasonal influenza suggests that patients with avian influenza may benefit from the use of oseltamivir (low confidence in effect estimates). Given the high mortality of the disease and the absence of effective alternatives, the WHO made a strong recommendation in favor of the use of oseltamivir rather than no treatment in patients with avian influenza. <sup>28</sup>
High incremental cost (or resource use) relative to the benefits may not justify the intervention	Strong recommendation against	In patients with idiopathic pulmonary fibrosis, treatment with azathioprine plus prednisone offers a possible but uncertain benefit in comparison with no treatment. The intervention, however, is associated with a substantial established harm. An international guideline made a recommendation against the combination of corticosteroids plus azathioprine in patients with idiopathic pulmonary fibrosis. <sup>28</sup>

(Continued)

TABLE 17-1

**Five Paradigmatic Situations That Justify Strong Recommendations Based on Low or Very Low Confidence (Continued)**

Paradigmatic Situation	Confidence in Effect Estimates for Health Outcomes (Quality of Evidence)		Balance of Benefits and Harms	Values and Preferences
	Benefits	Harms		
Potential equivalence, one option clearly less risky or costly	Low or very low	High or moderate	Magnitude of benefit apparently similar—though uncertain—for alternatives; we are confident of less harm or cost for one of the competing alternatives	A high value is placed on the reduction in harm
High confidence in similar benefits, one option potentially more risky or costly	High or moderate	Low or very low	Established that magnitude of benefit is similar for alternative management strategies; best (though uncertain) estimate is that one alternative has appreciably greater harm	A high value is placed on avoiding the potential increase in harm
Potential catastrophic harm	Immaterial (very low to high)	Low or very low	Potential important harm of the intervention, magnitude of benefit is variable	A high value is placed on avoiding potential increase in harm

Abbreviations: AT9, 9th edition of the antithrombotic guidelines; CI, confidence interval; MALT, mucosa-associated lymphoid tissue; WHO, World Health Organization.

Resource Considerations	Recommendation	Example
High incremental cost (or resource use) relative to the benefits may not justify one of the alternatives	Strong recommendation for less harmful/less expensive	Low-quality evidence suggests that initial <i>Helicobacter pylori</i> eradication in patients with early stage extranodal marginal zone (MALT) B-cell lymphoma results in similar rates of complete response in comparison with the alternatives of radiation therapy or gastrectomy, but with high confidence of less harm, morbidity, and cost. Consequently, UpToDate made a strong recommendation in favor of <i>H pylori</i> eradication rather than radiotherapy in patients with MALT lymphoma. <sup>29</sup>
High incremental cost (or resource use) relative to the benefits may not justify one of the alternatives	Strong recommendation against the intervention with possible greater harm	In women requiring anticoagulation and planning conception or in pregnancy, high confidence estimates suggest similar effects of different anticoagulants. However, indirect evidence (low confidence in effect estimates) suggests potential harm to the unborn infant with oral direct thrombin (eg, dabigatran) and factor Xa inhibitors (eg, rivaroxaban, apixaban). The AT9 guidelines recommended against the use of such anticoagulants in women planning conception or in pregnancy. <sup>1</sup>
High incremental cost (or resource use) relative to the benefits may not justify the intervention	Strong recommendation against the intervention	In males with androgen deficiency, testosterone supplementation likely improves quality of life. Low-confidence evidence suggests that testosterone increases cancer spread in patients with prostate cancer. The US Endocrine Society made a recommendation against testosterone supplementation in patients with prostate cancer. <sup>30</sup>

In decision analysis, the parallel to strong recommendations occurs when the relative utility of the management options changes little and the preferred alternative does not change, after varying probability estimates and varying values. Clinicians should look for a table that lists which variables were included in their sensitivity analyses, what range of values they used for each variable, and which variables, if any, altered the relative desirability of the management strategies under consideration.

Ideally, decision analysts will subject all of their probability estimates to a sensitivity analysis. The range over which they will test should depend on the source of the data. If the estimates come from large *randomized trials* with low *risk of bias* and narrow CIs, the range of estimates tested can be narrow. When risk of bias is greater or estimates of benefits and downsides less precise, sensitivity analyses testing a wide range of values become appropriate. Decision analysts also should test utility values with sensitivity analyses, with the range of values again determined by the source of the data. If large numbers of patients or knowledgeable and representative members of the general public gave similar ratings to the outcome states, investigators can use a narrow range of utility values in the sensitivity analyses. If the ratings came from a small group of raters or if the individuals provided widely varying estimates of typical utilities, then investigators should use a wider range of utility values in the sensitivity analyses.

### **For Weak Recommendations, Does the Information Facilitate Shared Decision Making?**

Recommendations—in particular, weak recommendations—should explicitly provide the key underlying information necessary to act on the recommendation. In guidelines, this information is typically found in the remarks section, in the recommendation rationale, or in tables that accompany the recommendation. The GRADE Working Group, in collaboration with the Cochrane Collaboration, has designed a specific table for this purpose: the *summary-of-findings table*. This table provides

the confidence ratings for all important outcomes and the associated estimates of relative and absolute effects. Table 17-2 shows a summary-of-findings table relevant for the clinical scenario presented at the beginning of this chapter. As we discuss later, summary-of-findings tables can facilitate shared decision making.<sup>33</sup> The absolute measures of effect you will find in GRADE summary-of-findings tables are typically presented within the decisions trees in decision analyses.

### **Was the Influence of Conflict of Interests Minimized?**

The judgments involved in the interpretation of the evidence and the decision on the final recommendation may be vulnerable to *conflicts of interest*. In medicine, guideline panelists frequently—and decision-analyst authors sometimes—report financial ties with the pharmaceutical industry.<sup>34-36</sup> Nonfinancial conflicts of interests are also common and may have even greater effect than financial conflicts.<sup>37,38</sup> These conflicts include intellectual conflicts (eg, previous publication of studies relevant to a recommendation) and professional conflicts (eg, radiologists making recommendations about breast cancer screening or urologists recommending prostate cancer screening).<sup>39,40</sup>

Clinicians can check the conflict of interest statements of the guideline panelists or decision analysts, usually found at the beginning or end of a publication or in a supplementary file. Just as important, clinicians should check what strategies were implemented to manage these conflicts of interest. Guidelines or decision analyses with a large representation of panelists without conflicts of interest, that have placed nonconflicted participants in positions of authority, or that have implemented rules to limit the influence of both financial and nonfinancial conflicts of interest are more credible than those that have not. Guidelines that excluded conflicted experts are likely to have limited the influence of conflicts of interest but may have compromised the credibility of the guidelines and possibly threatened their acceptability. Clinicians also can check whether recommendations

were collected and managed for the whole guideline or on a recommendation-by-recommendation basis. The influence of potential conflicts of interest may be diminished with the latter approach.

The AT9 guidelines provide an example of implementation of a number of these strategies.<sup>38</sup> A nonconflicted methodologist was chosen as the chair of each of the 14 panels making recommendations and was primarily responsible for that chapter. The chair and 2 other members of the executive committee ultimately responsible for the whole guideline were nonconflicted methodologists. Both financial and intellectual conflicts of interest were assessed on a recommendation-by-recommendation basis. Panelists with major conflicts were in principle excluded from participation in decision making. Challenges in implementing this approach highlight the efforts required to arrive at an optimal strategy for managing conflict of interest.<sup>41,42</sup>

## HOW SHOULD YOU USE RECOMMENDATIONS?

---

### Strong Recommendations

If the panel's assessment is astute, clinicians can apply strong recommendations to all or almost all of the patients in all or almost all circumstances without thorough—or even cursory—review of the underlying evidence and without a detailed discussion with the patient. The same is true for decision analysis when the utility of one alternative is substantially greater than the other and this relative utility is robust to sensitivity analyses. Whether



TABLE 17-2

### Summary-of-Findings Table: Antepartum and Postpartum Prevention of VTE With Prophylactic Dose of Low-Molecular-Weight Heparin vs No Prophylaxis in Pregnant Women With Prior VTE<sup>1</sup>

Outcome	RR (95% CI)	Anticipated Absolute Effects During Pregnancy		Confidence in the Estimates of the Effect
		Risk Without Prophylaxis	Risk Difference With LMWH	
Symptomatic VTE	0.36 (0.20-0.67)	<b>Low Risk</b>		Low due to indirectness <sup>b</sup> and imprecision <sup>c</sup>
		20 VTE per 1000	13 fewer VTE per 1000 (from 16 to 7 fewer)	
		<b>Intermediate and High Risk<sup>a</sup></b>		
		80 VTE per 1000	51 fewer VTE per 1000 (from 65 to 30 fewer)	
Major bleeding	1.57 (1.32-1.87) <sup>d</sup>	<b>Antepartum Period</b>		Low due to indirectness <sup>a</sup> and imprecision <sup>f</sup>
		3 bleeds per 1000	1 more bleed per 1000 (from 1 to 3 more) <sup>e</sup>	
		<b>Postpartum Period</b>		
		10 bleeds per 1000	6 more bleeds per 1000 (from 3 to 8 more) <sup>d</sup>	

(Continued)





TABLE 17-2

**Summary-of-Findings Table: Antepartum and Postpartum Prevention of VTE With Prophylactic Dose of Low-Molecular-Weight Heparin vs No Prophylaxis in Pregnant Women With Prior VTE (Continued)**

Outcome	RR (95% CI)	Anticipated Absolute Effects During Pregnancy		Confidence in the Estimates of the Effect
		Risk Without Prophylaxis	Risk Difference With LMWH	
Burden of treatment	...	No incremental burden	Daily injections	High

Abbreviations: CI, confidence interval; LMWH, low-molecular-weight heparin; RR, relative risk; VTE, venous thromboembolism.

<sup>a</sup>Single unprovoked VTE, pregnancy-related or estrogen-related VTE, or multiple prior unprovoked VTE not receiving long-term anticoagulation.

<sup>b</sup>Population is indirect (ie, did not include pregnant women).

<sup>c</sup>95% confidence interval includes marginal benefit.

<sup>d</sup>Relative effect estimate based on the systematic review by Collins et al.<sup>32</sup>

<sup>e</sup>Absolute risk estimates for major bleeding in women using LMWH based on the systematic review by Greer et al.<sup>8</sup>

<sup>f</sup>95% confidence interval includes marginal harm.

Adapted from Bates et al.<sup>1</sup>

## USING THE GUIDE

### **Is the Clinical Question Clear and Comprehensive?**

The recommendation presented at the beginning of this chapter clearly specifies what is being proposed (“antepartum prophylaxis with prophylactic- or intermediate-dose LMWH”) and what was the comparison (“rather than clinical vigilance or routine care”).<sup>1</sup>

As we can see in [Table 17-2](#), guideline panelists considered the outcomes of symptomatic thromboembolism, major bleeding, and burden of treatment—the outcomes likely important to patients.

### **Was the Recommendation Based on the Best Current Evidence?**

In the methods section of the published AT9 guidelines, we find the following description: “To identify the relevant evidence, a team ... conducted literature searches of Medline, the Cochrane Library, and the Database of Abstracts of Reviews of Effects ... for systematic reviews and another for original studies” and “The quality of reviews was assessed ... and wherever possible, current high-quality systematic reviews were used as the source of summary estimates.”<sup>43</sup> This strategy ensured that estimates were based on best current evidence at the time the recommendation was issued.

### **Are Values and Preferences Associated With Outcomes Appropriately Specified?**

Guideline authors noted that a systematic review of patient preferences for antithrombotic treatment did not identify any studies of pregnant women. A rating exercise of different outcomes among experienced clinicians participating on the guideline suggested that 1 episode of VTE (deep venous thrombosis or pulmonary embolism) is more or less equivalent to 1 major extracranial bleed. Panelists’ clinical experience suggested that most women, but not all, would choose long-term prophylaxis when confronted with the burden of self-injecting with LMWH for several months, suggesting a relatively high value on preventing VTE

and a relatively high tolerance for self-injection. These values and preferences were used to develop the recommendation.

**Do the Authors Indicate the Strength of Their Recommendations?**

The recommendation was classified as “weak” using the GRADE approach.

**Is the Evidence Supporting the Recommendation Easily Understood?**

The recommendation was accompanied by a summary-of-findings table (Table 17-2) that provides absolute estimates for the outcomes important to patients. We discuss subsequently how this information can help with shared decision making.

**Was the Influence of Conflict of Interests Minimized?**

As we described earlier, the AT9 guidelines implemented a number of the strategies to diminish the influence of conflict of interest on recommendations.

discussion of the evidence with patients might sometimes still be helpful in such circumstances—for instance, whether it may increase adherence to treatment—remains uncertain.

For example, the Allergic Rhinitis and its Impact on Asthma guideline recommended intranasal glucocorticoids rather than intranasal antihistamines for treatment of allergic rhinitis in adults (strong recommendation).<sup>44</sup> This recommendation was based on an important reduction of *symptoms* with glucocorticoids (rhinorrhea, nasal blockage and itching) with no important adverse events. The effect estimates came from a systematic review of randomized trials with low risk of bias, consistent results across trials, precise effects (narrow CIs), and results applicable to the population. The guideline panel’s inference that all, or almost all, informed patients would

choose the glucocorticoids is eminently reasonable. Therefore, a detailed discussion with the patients about the benefits and potential harms of intranasal glucocorticoids over intranasal antihistamines will not be necessary.

There will always be idiosyncratic circumstances in which clinicians should not adhere to even strong recommendations. For instance, aspirin in the context of myocardial infarction warrants a strong recommendation, but it would be a mistake to administer the treatment to a patient who is allergic to aspirin. Such idiosyncratic situations are, fortunately, unusual.

## **Weak Recommendations**

With careful consideration of the evidence, as well as of patient's values and preferences, many recommendations are weak, even in clinical fields with a large body of randomized trials and systematic reviews. For instance, two-thirds of more than 600 recommendations issued in AT9 were weak.<sup>17</sup>

Because weak recommendations are typically sensitive to patients' values and preferences, a shared decision-making approach that involves a discussion with the patient addressing the potential benefits and harms of the proposed course of action is the optimal way to ensure that decisions reflect both the best evidence and patients' values and preferences (see Chapter 18, Decision Making and the Patient). To use weak recommendations, clinicians need to understand the underlying evidence.

For example, the American College of Physicians suggested the use of cholinesterase inhibitors or memantine in patients with dementia (weak recommendation).<sup>45</sup> This recommendation is based on evidence from randomized trials warranting high confidence in a small benefit of the drugs in slowing the deterioration of cognition and global function. Guideline panelists pointed out that, if quality of life is judged as poor—in particular, with more advanced dementia—family members may not view the limited slowing of dementia progression as a desirable goal. Moreover, the magnitude of the effect is small, and

there are adverse effects associated with the drugs. The panel then reasonably expected that informed patients (or their families) would make different choices.

## CLINICAL SCENARIO RESOLUTION

**A**fter reviewing this guide, and specifically the information in Table 17-2, you decide that the recommendation is trustworthy and you plan to engage patients like the one presented in the opening scenario in shared decision making. When you meet with the patient, you start by discussing the benefits of LMWH during pregnancy vs no treatment (51 fewer cases of symptomatic VTE per 1000 women), followed by information about adverse effects (7 more maternal bleeds per 1000 women followed up during the pregnancy and post partum), and you mention the potential burden of treatment that daily injections for several months will represent (low confidence in effect estimates for all outcomes aside from the burden of injections). If the guideline panel is correct, most patients will place a higher value in lowering the risk of a thrombotic event and less on the uncertain small increase in the risk of bleeding and the certain burden of treatment. Such patients will choose prophylaxis. If the panel is correct, however, some patients will decline therapy.

Thus, shared decision making is required to ensure the patient understands the best evidence available and the decision is consistent with the patient's values and preference. You are not surprised when the patient chooses VTE prophylaxis.

## References

1. Bates SM, Greer IA, Middeldorp S, Veenstra DL, Prabulos AM, Vandvik PO. VTE, thrombophilia, antithrombotic therapy, and pregnancy: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest

- Physicians Evidence-Based Clinical Practice Guidelines. *Chest*. 2012;141(2 suppl):e691S-736S.
- Graham R, Mancher M, Wolman DM, Greenfield S, Steinberg E, eds. *Clinical Practice Guidelines We Can Trust*. Washington, DC: National Academies Press; 2011.
  - Laine C, Taichman DB, Mulrow C. Trustworthy clinical guidelines. *Ann Intern Med*. 2011;154(11):774-775.
  - Qaseem A, Forland F, Macbeth F, Ollenschläger G, Phillips S, van der Wees P; Board of Trustees of the Guidelines International Network. Guidelines International Network: toward international standards for clinical practice guidelines. *Ann Intern Med*. 2012;156(7):525-531.
  - Shekelle P, Woolf S, Grimshaw JM, Schünemann HJ, Eccles MP. Developing clinical practice guidelines: reviewing, reporting, and publishing guidelines; updating guidelines; and the emerging issues of enhancing guideline implementability and accounting for comorbid conditions in guideline development. *Implement Sci*. 2012;7:62.
  - Schünemann HJ, Wiercioch W, Etzeandía I, et al. Guidelines 2.0: systematic development of a comprehensive checklist for a successful guideline enterprise. *CMAJ*. 2014;186(3):E123-E142.
  - Kassirer JP, Moskowitz AJ, Lau J, Pauker SG. Decision analysis: a progress report. *Ann Intern Med*. 1987;106(2):275-291.
  - Greer IA, Nelson-Piercy C. Low-molecular-weight heparins for thromboprophylaxis and treatment of venous thromboembolism in pregnancy: a systematic review of safety and efficacy. *Blood*. 2005;106(2):401-407.
  - Johnston JA, Brill-Edwards P, Ginsberg JS, Pauker SG, Eckman MH. Cost-effectiveness of prophylactic low molecular weight heparin in pregnant women with a prior history of venous thromboembolism. *Am J Med*. 2005;118(5):503-514.
  - Lipsky BA, Berendt AR, Cornia PB, et al; Infectious Diseases Society of America. 2012 Infectious Diseases Society of America clinical practice guideline for the diagnosis and treatment of diabetic foot infections. *Clin Infect Dis*. 2012;54(12):e132-e173.
  - National Institute for Health and Care Excellence. *Diabetic foot problems: Inpatient management of diabetic foot problems (CG119)*. London, England: National Institute for Health and Care Excellence; 2011.
  - World Health Organization. WHO recommendations for the prevention and treatment of postpartum haemorrhage. Geneva, Switzerland: World Health Organization; 2012.
  - Albers GW, Amarenco P, Easton J, Sacco RL, Teal P. Antithrombotic and thrombolytic therapy for ischemic stroke: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines (8th Edition). *Chest*. 2008;133(6 suppl):630S-669S.
  - Lansberg MG, O'Donnell MJ, Khatri P, et al. Antithrombotic and thrombolytic therapy for ischemic stroke: antithrombotic therapy and prevention of thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest*. 2012;141(2 suppl):e601S-e636S

15. Dennis M, Sandercock PA, Reid J, et al; CLOTS Trials Collaboration. Effectiveness of thigh-length graduated compression stockings to reduce the risk of deep vein thrombosis after stroke (CLOTS trial 1): a multicentre, randomised controlled trial. *Lancet*. 2009;373(9679):1958-1965.
16. Ansell J, Hirsh J, Hylek E, Jacobson A, Crowther M, Palareti G; American College of Chest Physicians. Pharmacology and management of the vitamin K antagonists: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines (8th Edition). *Chest*. 2008;133(6 suppl):160S-198S.
17. Holbrook A, Schulman S, Witt DM, et al; American College of Chest Physicians. Evidence-based management of anticoagulant therapy: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest*. 2012;141(2 suppl):e152S-84S.
18. Hughes WT, Armstrong D, Bodey GP, et al. 2002 guidelines for the use of antimicrobial agents in neutropenic patients with cancer. *Clin Infect Dis*. 2002;34(6):730-751.
19. Freifeld AG, Bow EJ, Sepkowitz KA, et al; Infectious Diseases Society of America. Clinical practice guideline for the use of antimicrobial agents in neutropenic patients with cancer: 2010 update by the Infectious Diseases Society of America. *Clin Infect Dis*. 2011;52(4):e56-e93.
20. US Preventive Services Task Force. Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. *Ann Intern Med*. 2009;151(10):716-726, W-236.
21. MacLean S, Mulla S, Akl EA, et al; American College of Chest Physicians. Patient values and preferences in decision making for antithrombotic therapy: a systematic review: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest*. 2012;141(2 suppl):e1S-e23S.
22. Nilsen ES, Myrhaug HT, Johansen M, Oliver S, Oxman AD. Methods of consumer involvement in developing healthcare policy and research, clinical practice guidelines and patient information material. *Cochrane Database Syst Rev*. 2006;(3):CD004563.
23. Griebisch I, Coast J, Brown J. Quality-adjusted life-years lack quality in pediatric care: a critical review of published cost-utility studies in child health. *Pediatrics*. 2005;115(5):e600-e614.
24. Atkins D, Eccles M, Flottorp S, et al; GRADE Working Group. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches *BMC Health Serv Res*. 2004;4(1):38.
25. Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. *J Clin Epidemiol*. 2011;64(4):380-382.
26. American College of Cardiology Foundation and American Heart Association. Methodology Manual and Policies From the ACCF/AHA Task Force on Practice Guidelines (2010). <http://my.americanheart.org/professional/StatementsGuidelines/PoliciesDevelopment/Development/>

- Methodologies-and-Policies-from-the-ACCAHA-Task-Force-on-Practice-Guidelines\_UCM\_320470\_Article.jsp. Accessed August 4, 2014.
27. US Preventive Services Task Force. Grade definitions. <http://www.uspreventiveservicestaskforce.org/uspstf/grades.htm>. Accessed August 4, 2014.
  28. Schünemann HJ, Hill SR, Kakad M, et al; WHO Rapid Advice Guideline Panel on Avian Influenza. WHO Rapid Advice Guidelines for pharmacological management of sporadic human infection with avian influenza A (H5N1) virus. *Lancet Infect Dis*. 2007;7(1):21-31.
  29. Freedman AS, Lister A, Connor RF. Management of gastrointestinal lymphomas. UpToDate. <http://www.uptodate.com>. Accessed March 27, 2014.
  30. Bhasin S, Cunningham GR, Hayes FJ, et al; Task Force, Endocrine Society. Testosterone therapy in men with androgen deficiency syndromes: an Endocrine Society clinical practice guideline. *J Clin Endocrinol Metab*. 2010;95(6):2536-2559.
  31. Brito JP, Domecq JP, Murad MH, Guyatt GH, Montori VM. The Endocrine Society guidelines: when the confidence cart goes before the evidence horse. *J Clin Endocrinol Metab*. 2013;98(8):3246-3252.
  32. Collins R, Scrimgeour A, Yusuf S, Peto R. Reduction in fatal pulmonary embolism and venous thrombosis by perioperative administration of subcutaneous heparin. Overview of results of randomized trials in general, orthopedic, and urologic surgery. *N Engl J Med*. 1988;318(18):1162-1173.
  33. Treweek S, Oxman AD, Alderson P, et al; DECIDE Consortium. Developing and Evaluating Communication Strategies to Support Informed Decisions and Practice Based on Evidence (DECIDE): protocol and preliminary results. *Implement Sci*. 2013;8:6.
  34. Norris SL, Holmer HK, Ogden LA, Burda BU. Conflict of interest in clinical practice guideline development: a systematic review. *PLoS One*. 2011;6(10):e25153.
  35. Neuman J, Korenstein D, Ross JS, Keyhani S. Prevalence of financial conflicts of interest among panel members producing clinical practice guidelines in Canada and United States: cross sectional study. *BMJ*. 2011;343:d5621.
  36. Choudhry NK, Stelfox HT, Detsky AS. Relationships between authors of clinical practice guidelines and the pharmaceutical industry. *JAMA*. 2002;287(5):612-617.
  37. Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124.
  38. Guyatt G, Akl EA, Hirsh J, et al. The vexing problem of guidelines and conflict of interest: a potential solution. *Ann Intern Med*. 2010;152(11):738-741.
  39. Norris SL, Burda BU, Holmer HK, et al. Author's specialty and conflicts of interest contribute to conflicting guidelines for screening mammography. *J Clin Epidemiol*. 2012;65(7):725-733.
  40. Dahm P, Kunz R, Schünemann H. Evidence-based clinical practice guidelines for prostate cancer: the need for a unified approach. *Curr Opin Urol*. 2007;17(3):200-207.



41. Neumann I, Karl R, Rajpal A, Akl EA, Guyatt GH. Experiences with a novel policy for managing conflicts of interest of guideline developers: a descriptive qualitative study. *Chest*. 2013;144(2):398-404.
42. Neumann I, Akl EA, Valdes M, et al. Low anonymous voting compliance with the novel policy for managing conflicts of interest implemented in the 9th version of the American College of Chest Physicians antithrombotic guidelines. *Chest*. 2013;144(4):1111-1116.
43. Guyatt GH, Norris SL, Schulman S, et al. Methodology for the development of antithrombotic therapy and prevention of thrombosis guidelines: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest*. 2012;141(2 suppl):53S-70S.
44. Brozek JL, Bousquet J, Baena-Cagnani CE, et al; Global Allergy and Asthma European Network; Grading of Recommendations Assessment, Development and Evaluation Working Group. Allergic Rhinitis and its Impact on Asthma (ARIA) guidelines: 2010 revision. *J Allergy Clin Immunol*. 2010;126(3):466-476.
45. Qaseem A, Snow V, Cross JT Jr, et al; American College of Physicians/American Academy of Family Physicians Panel on Dementia. Current pharmacologic treatment of dementia: a clinical practice guideline from the American College of Physicians and the American Academy of Family Physicians. *Ann Intern Med*. 2008;148(5):370-378.

# Decision Making and the Patient

Victor M. Montori, Glyn Elwyn, PJ Devereaux,  
Sharon E. Straus, R. Brian Haynes, and  
Gordon Guyatt

## IN THIS CHAPTER

---

### Introduction

What Approaches to Decision Making Are Available?

Paternalistic Approach

Clinician-as-Perfect-Agent Approach

Informed Decision-Making Approach

Shared Decision-Making Approach

What Decision-Making Approach Should I Choose With  
This Patient?

### What Tools Can I Use in Making Challenging Decisions With This Patient?

Patient Decision Aids

### Should I Use More Time and Effort in Decision Making With This Patient Now?

Time as a Barrier

Important vs Unimportant Decisions

(continued on following page)

Straightforward vs Difficult Decisions

Misinformed Participants

The Patient With Multiple Chronic Conditions

Other Solutions

Use a Patient Decision Aid

**Conclusion**

---

## INTRODUCTION

---

One of the 3 key principles of *evidence-based medicine* (EBM) is that the *evidence* alone is never sufficient to make a clinical decision (see Chapter 2, What Is Evidence-Based Medicine?). Clinicians require expertise in interpreting the patient dilemma (in its clinical, social, and economic contexts) and in identifying the body of evidence that bears on optimal patient treatment. These considerations, however, are not enough. Evidence-based medicine requires that clinical decisions be consistent with the informed *values and preferences* of the patient.

We use values and preferences as an overarching term that includes patients' perspectives, priorities, beliefs, expectations, values, and goals for health and life. We also use this phrase, more precisely, to mean the processes that individuals use in considering the potential benefits, *harms*, costs, and inconveniences of the management options in relation to one another.

Consideration of patient values and preferences often enables clinicians to understand the patient who declines life-saving treatment and the patient who seeks active treatment even when, from a clinician's perspective, the hope of any gain is lost and palliation may seem a wiser path.

Differences in values and preferences also may explain policy decisions and *practice guidelines* that, despite relying on the same evidence, differ across settings and contexts. Patient values and preferences become more crucial when confidence in the estimates of a beneficial effect is low and when the balance is close between important benefits and similarly important downsides.

### What Approaches to Decision Making Are Available?

**Box 18-1** summarizes decision-making approaches theoretically available to the clinician and patient facing an important decision.

**BOX 18-1****Decision-Making Approaches****Minimal or no attempt to ensure decision consistent with patient values and preferences**

Paternalistic or parental approaches: Clinician makes minimal effort to establish patient values and preferences, makes decision on behalf of patient

**Approaches that attempt to ensure decision consistent with patient values and preferences**

Clinician-as-perfect-agent approach: Clinician ascertains patient's values and preferences, makes decision on behalf of patient

Informed decision making: Clinician provides patient with the information; patient makes the decision

Shared decision making: Patient and clinician both bring information/evidence and values and preferences to the decision

**Paternalistic Approach**

When clinicians offer patients minimal information about the options and make the decision without patient input, a style commonly referred to as a paternalistic or parental approach, they are not considering patient values and preferences. This does not mean that patients do not have an opportunity to express their wishes, but they may do so in a delayed fashion and through actions. For instance, if the treatment choice is not consistent with their values and preferences, patients may not act on the decision or may abandon the plan shortly after the visit with the clinician. Evidence-based medicine requires respecting and incorporating patient values and preferences in the process of decision. Thus, this parental approach, in its violation of patient autonomy, is inconsistent with EBM.

## Clinician-as-Perfect-Agent Approach

In theory, one can ensure that decisions are consistent with patient values and preferences without actively involving the patient in the decision. To do so, clinicians must assess the patient's values and preferences and then place these in the context of the evidence about the benefits and risks of alternative courses of action.

Some experts consider this approach, sometimes called the clinician-as-perfect-agent model, impossible to implement.<sup>1</sup> Their position is based on the absence of effective approaches that would confidently yield a deep understanding of the processes that patients use in considering the potential benefits, harms, costs, and inconveniences of the options in relation to one another.

Other experts offer tools for eliciting patient values and preferences, an approach that relies on what is called expected utility theory. Along with these tools, these experts offer models—*decision analyses*—for eliciting the numerical value (utility) that patients might put on a particular outcome and then integrating these values with a calculation of the likelihood of each important outcome for alternative management strategies (see Chapter 17, How to Use a Patient Management Recommendation: Clinical Practice Guidelines and Decision Analyses). These models are limited in that (1) psychologists have found that patients do not consistently make decisions compatible with the underlying assumptions of decision analyses,<sup>2,3</sup> and (2) the models are difficult to use in day-to-day practice.<sup>4</sup> Moreover, there is limited empirical support for the assumptions supporting these tools,<sup>5</sup> and decisions from these analyses may not be the ones reasonable patients would make even after understanding the issues.

## Informed Decision-Making Approach

In a very different decision-making style, empowered patients may obtain all of the information pertinent to the decision, consider the options, and make a decision with minimal clinician input. This approach, often referred to as the informed

decision-making style, recognizes that patients and clinicians have their own expertise. Patients are experts in their values and preferences and in their personal contexts (personal and social factors—such as working the night shift, lacking a caregiver to help with pill taking and attending laboratory testing, and undisclosed use of alternative medicine agents—that may affect their *adherence* to or tolerance of a treatment or that may affect the effectiveness of a treatment). Clinicians are experts in the technical aspects of the decision (ie, the evidence base informing the pros and cons of each of the options and the experience concerning implementation). The clinician's role with patients choosing this approach is primarily to present information with completeness and clarity.<sup>6</sup>

### Shared Decision-Making Approach

In this approach, patients and clinicians engage in a bidirectional exchange. The clinician shares the evidence from clinical research, and the patient shares the evidence accessible in the “patient space” acquired through personal experience, social interaction, and consultation of lay sources, technical references, or the Internet. The bidirectional interaction also includes personal information (ie, sharing the basis for values and preferences). Both the patient and clinician deliberate about the options, explicitly acknowledging the values and preferences they are using, and together arrive at an agreement about the best course of action. The label offered for this model is the shared decision-making process.<sup>6,7</sup>

There are numerous descriptions of shared decision making.<sup>8</sup> One model uses the idea of clinicians having 3 types of “talk” with their patients: team talk, option talk, and decision talk.<sup>9</sup> [Box 18-2](#) describes the 3 types of talk, and [Figure 18-1](#) illustrates the suggested sequence, supporting the patients in gaining understanding about alternative courses of action and, in so doing, constructing informed preferences and, in due course, coming to good decisions.

## BOX 18-2

**Talk Model of Shared Decision Making****Team Talk**

Team talk facilitates patients' awareness that reasonable options exist and that the clinician will help the patient understand how to consider these options in more detail. Components of team talk include:

**Stepping back.** Summarize and say: "Now that we have identified the problem, it's time for us as a team to think about what to do next."

**Offering the choices.** Be aware that patients can misconstrue the presentation of choice and think that the clinician is incompetent, uninformed, or both. Reduce this risk by saying: "There is good information about how these treatments differ that I'd like to discuss with you so that we can work together to consider them."

**Justifying the choices.** Emphasize the importance of respecting individual values and preferences and the role of uncertainty.

For individual values and preferences, explaining that different issues matter more to some people than to others should be easily grasped. Say: "Treatments have different consequences. Some will matter more to you than to other people."

As for uncertainty, patients are often unaware of the extent of uncertainty in medicine—that evidence may be lacking and that individual outcomes are unpredictable at the individual level. Say: "Treatments are not always effective, and the chances of experiencing adverse effects vary."

**Checking the patient's reaction.** The choice of options may be disconcerting, and some patients may express concern.



Suggested phrases to use: "Shall we go on?" "Shall I tell you about the options?"

**Postponing closure.** Some patients react by asking clinicians to tell them what to do. We suggest postponing or deferring closure if this occurs, reassuring the patient that you are willing to support the process. Say: "I'm happy to share my views and help you get to a good decision. But before I do so, may I describe the options in more detail so that you understand what is at stake?"

### **Option Talk**

Option talk is the act of being clear about reasonable treatment alternatives and helping patients compare them. Components of option talk include:

**Checking the patient's knowledge.** Even well-informed patients may only be partially aware of their options and the associated harms and benefits, or they be misinformed. Check by asking, "What have you heard or read about the treatment of your condition?"

**Listing the options.** Make a clear list of the options because it provides good structure. Jot them down and say: "Let me list the options before we get into more detail." If appropriate, include the option of watchful waiting, or use positive terms such as active surveillance.

**Describing the options.** Generate dialogue and explore values and preferences. Describe the options in practical terms. If there are 2 medical treatments, say: "Both options similarly involve taking medication on a regular basis." Point out when there are clear differences (surgery vs medication), where postponement is possible, and where decisions are reversible. Say: "These options will have different implications for you as compared with other people, so I want to describe...."

**Explaining harms and benefits.** Being clear about the pros and cons of different options is at the heart of shared decision making. Learn about effective risk communication, such as framing effects and the importance of providing risk data in absolute as well as relative terms. Try giving information in chunks and then checking to see whether the patient understands, a process known as “chunking and checking.”

**Providing patient decision aids.** These tools make options visible and may save time. Some are sufficiently concise to use in clinical encounters. Say: “These tools have been designed to help you understand options in more detail and help us make a decision together. Let’s review these together.”

**Summarizing the options.** List the options again and assess understanding by asking for reformulations. This is called the “teach-back” method and is a good check for misconceptions.

### **Decision Talk**

Make an effort here to ask about “what matters most” to the patients, now that they better understand how to compare the alternatives. Help them form their own views, and try to work with patients to see how best to take the next steps—to make a wise and well-considered decision. Components of decision talk include:

**Focusing on values and preferences.** Guide the patient to form preferences. A suggested phrase: “What, from your point of view, matters most to you?” Help patients consider which aspects of the options will lead them to choose one option over another, according to their own priorities.

**Eliciting a preference.** Be prepared with a backup plan by offering more time or being willing to guide patients, if they indicate that this is their wish.

**Moving to a decision.** Try checking for the need to either defer a decision or make a decision. Suggested phrases: “Are you ready to decide?” “Do you want more time? Do you have more questions?” “Are there more things we should discuss?”

**Offering review.** A good point of closure is to remind the patient, when feasible, that decisions may be reviewed.

Some clinicians might interpret shared decision making as requiring clinicians to present their own values and preferences that may then influence the decision. Evidence-based practitioners may find this undesirable for 2 reasons. The first reason is philosophical: although clinicians may experience consequences of these choices through empathy, by experiencing regret when patients experience bad outcomes, or by getting sued, it is patients who endure the treatments and bear the burdens of the outcomes of the choices made. The second reason relates to how patients and clinicians have historically related to each other. Patients may not be willing to reveal their values and preferences if they seem at odds with those the clinician reveals. This concern is made more important by evidence, particularly in preventive care decisions, that patients and clinicians sometimes have values and preferences that differ, although neither party is aware of the differences ([Box 18-3](#)).

In contrast, one might argue that all decision-making approaches incorporate clinician preferences if only to the extent that it is clinicians who decide the range of options that they are willing to offer to the patient. If one takes this position, then shared decision making has the merit of explicitly considering clinicians' values and preferences rather than doing so implicitly. Furthermore, patients appear to be interested in clinicians' preferences. Our guess is that every clinician who has tried to encourage patient autonomy has faced some form

FIGURE 18-1

## Decision-Making Approaches and Evidence-Based Medicine

Approaches	Parental	Clinician as Perfect Agent	Shared Decision Making	Informed
Direction and amount of information flow about options	Clinician ► Patient	Clinician ► Patient	Clinician ◄ ► Patient	Clinician ► Patient
Direction of information flow about values and preferences	Clinician ► Patient	Clinician ◄ Patient	Clinician ◄ ► Patient	Clinician ◄ Patient
Deliberation	Clinician	Clinician	Clinician, patient	Patient
Decider	Clinician	Clinician	Clinician, patient	Patient
Consistent with EBM principles	No, when decision is not purely technical and there are options	Yes	Yes	Yes

Abbreviation: EBM, evidence-based medicine. Modified from Charles et al.<sup>7</sup>

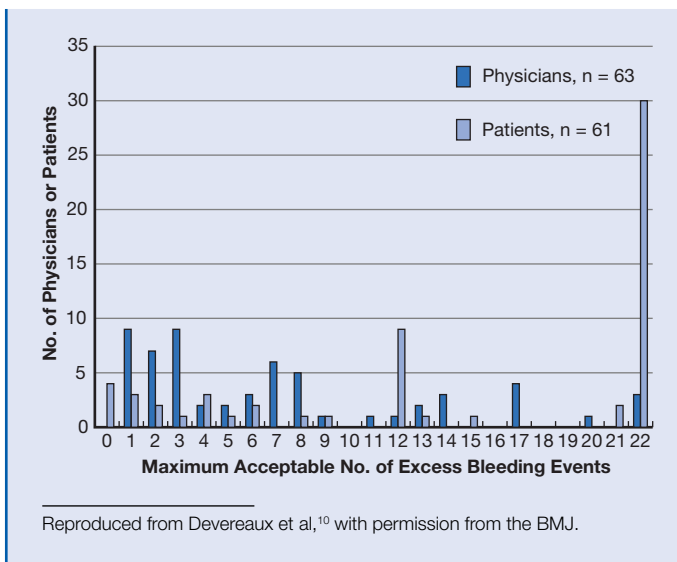
of the question, "What would you do?" Finally, because shared decision making espouses the incorporation of patient values and preferences into the decision-making process, it responds to patients' desires to be cared for by their clinician.

These considerations suggest that for shared decision making to work well, the power gradient between clinicians and patients needs to decrease substantially. Only a minimal gradient will ensure that informed patients can confidently choose an option inconsistent with the clinician's preferences; in reality,

### BOX 18-3

#### **Do Patients and Their Clinicians Share Similar Values and Preferences?**

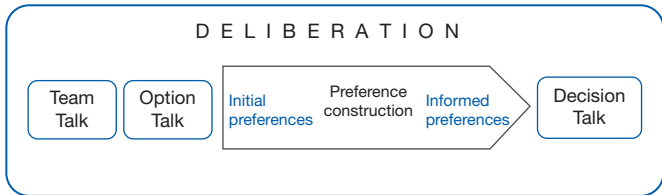
Devereaux et al<sup>10</sup> used a technique called probability tradeoff to determine the relative strength of aversion to stroke and gastrointestinal bleeding in the context of anticoagulation to prevent stroke in 61 at-risk patients and 63 physicians who treated patients with atrial fibrillation. The figure in this box shows the maximum number of excess upper gastrointestinal tract bleeding episodes per 100 patients treated to prevent 8 additional strokes (4 major and 4 minor) that patients and physicians found acceptable. The figure shows the following: (1) there is variability in stroke aversion among patients and among physicians; (2) patients were much more stroke averse than physicians; and (3) physicians seem more averse to adverse outcomes that they "cause" with their prescription (eg, bleeding) than to adverse outcomes that result from clinical course (eg, strokes). If one believes that patient preferences should guide treatment, these data suggest the following: if clinicians fail to incorporate patient values and preferences in the decision-making process, they will recommend against anticoagulation more often than is appropriate and, depending on which physician patients see, they will or will not get the treatment they would prefer.



many report their clinician's opinion as the most important factor that drives their decision to undergo an invasive procedure.<sup>11</sup> Also, there is evidence that even well-educated patients fear conflict that may arise if they were to engage in decision making and prefer an approach distinct from one their clinicians recommend.<sup>12</sup> A reduced power gradient implies that clinicians will act according to patients' informed values and preferences even when the decisions are not those they would have made for themselves (or that will enhance their income).

Figure 18-2 describes our current understanding of decision-making approaches. According to this understanding, clinicians can be aware of clues that patients give during the encounter about their values and preferences for involvement in a decision. All forms of participatory decision making, including its extremes of patient and clinician participation, involve clinicians offering patients evidence-based information about the available options.

FIGURE 18-2

**Team, Option, and Decision Talk: Shared Decision Making Made Easier**

### What Decision-Making Approach Should I Choose With This Patient?

Although surveys consistently reveal that patients are willing to receive information relevant to the decision at hand,<sup>13</sup> many patients prefer clinicians to take decisional responsibility.<sup>14,15</sup> Reasons include intense emotions surrounding the decision, lack of understanding, impaired physical or cognitive function, lack of self-confidence, and the general human tendency to prefer other people to take responsibility. More problematic reasons, however, exist: patients may not participate in decision making because clinicians do not communicate information in ways that are accessible to the patient (ie, use of technical language that requires health literacy and numeracy<sup>16</sup>), they have no experience or expectation of participating, or they fear disappointing or angering their clinician.

These considerations suggest that clinicians should present information about the options and then adapt to the decision approach patients prefer. Furthermore, these considerations suggest the need to exercise a high degree of empathy in determining what approach best accommodates the patient and the need to remain flexible as the patient's wishes change, which may occur even within the same visit and with each decision considered.

Given the variation in patients' values and preferences regarding the extent to which they wish to take responsibility for

management decisions, an empathic, flexible approach within the range of participatory decision-making styles offers advantages. The extent to which clinicians' values and preferences enter the discussion and the extent to which the clinician or patient plays the most active role in the final decision-making process can reflect the patient's preferred decision-making approach. Many clinicians have the impression that poorer or less educated patients, particularly those in low-income countries, are less inclined to participate in decision making. This may be so. It is also possible, however, that if clinicians practice optimal information sharing, listening, and empathy, they will find such patients capable of and interested in participating in making decisions about their care.

In summary, EBM practitioners seeking to incorporate patient values and preferences into clinical decisions should be able to effectively communicate to patients the nature of each of the options, empathically identify and enable the maximum extent of participation that the informed patient wants to have in the decision-making process, and identify and explicitly acknowledge when their own values and preferences are affecting the process of arriving at a decision.

## **WHAT TOOLS CAN I USE IN MAKING CHALLENGING DECISIONS WITH THIS PATIENT?**

---

### **Patient Decision Aids**

To effectively communicate the nature of the options, researchers have devised and tested tools called patient *decision aids*. These tools are an alternative to the use of intuitive approaches of communicating concepts of risk and risk reduction that clinicians may have developed through clinical experience. Decision aids present, in a patient-friendly manner, descriptive and probabilistic information about the disease, treatment options, and potential outcomes.<sup>17-19</sup> A well-constructed decision aid is based



on a *systematic review* of the literature and produces a rigorous summary of the outcomes and their probabilities. Clinicians who doubt that the summary of probabilities is rigorous can review the *primary studies* on which those probabilities are based and, using the principles in the *Users' Guides to the Medical Literature*, determine their accuracy. Furthermore, a well-constructed decision aid offers a tested and effective way of communicating information to patients who may have little background in quantitative decision making. Most commonly, decision aids use visual props, such as icon arrays, to present the proportion of people who experience the outcomes of importance with and without the intervention (Figure 18-3).

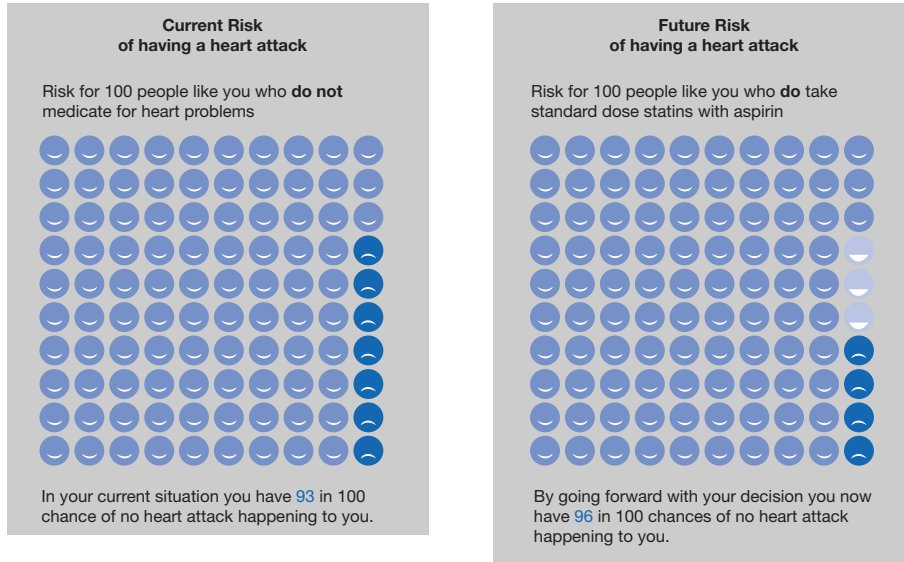
What influence do decision aids have on clinical practice? A Cochrane review identified 86 *randomized trials* of decision aids that support screening and treatment decisions.<sup>18</sup> Compared with usual care, decision aids increased reported patient participation in decision making (*relative risk*, 1.4; 95% *confidence interval* [CI], 1.0-2.3), improved patient knowledge (19/100 points in knowledge surveys; 95% CI, 13-24), and reduced decisional conflict (-9.1/100; 95% CI, -12 to -6). The systematic reviewers concluded that decision aids did not, however, consistently improve satisfaction with the decision-making process, *health outcomes*, or adherence to treatment, or reduce health care use or costs.

Like guidelines, decision aids may be problematic when conflicted developers fail to present the evidence, options, and outcomes fairly. As in guidelines, standards are being developed to try and ensure that decision aids are safe for patient use and do not mislead patients and clinicians.<sup>20,21</sup>

In summary, decision aids increase patient knowledge and improve measures intended to reflect the quality of the decision-making process and its outcome. The use of decision aids in routine clinical practice remains rare, and many implementation barriers exist.<sup>22</sup> Simple decision aids that clinicians can integrate into regular patient care could improve adoption.<sup>23</sup> Randomized trials have found that simple tools for use during the clinical encounter can increase the extent of patient participation

**FIGURE 18-3**

**Sample Decision Aid Developed to Help Patients Decide Whether to Take a Statin to Reduce Their Coronary Risk**



By permission of Mayo Foundation for Medical Education and Research. All rights reserved.

in decision making and, in turn, affect the extent to which informed patients' values determine health care decisions.<sup>24-28</sup> This evidence has not revealed consistent effects of using decision aids on treatment choice, adherence, clinical outcomes, or health care use or costs.

## SHOULD I USE MORE TIME AND EFFORT IN DECISION MAKING WITH THIS PATIENT NOW?

---

### Time as a Barrier

Should clinicians interested in practicing EBM and expecting to make clinical decisions that incorporate the values and preferences of the informed patient use 1 or more of the above approaches for all decisions? The ultimate constraint of clinical practice is time. Many clinicians have more to do in each encounter than they did in the past.<sup>29-31</sup> Attention to the patient's agenda competes with other activities that clinicians ought to do (eg, documentation, routine preventive care<sup>32</sup>) during visits that have not increased in duration to accommodate these additional activities and demands. Thus, it is not surprising that clinicians frequently cite time as a key barrier to patient education about options and to enhanced patient participation in decision making.<sup>33</sup> **Box 18-4** provides some suggestions for what to do when time is limited.

### Important vs Unimportant Decisions

Many of the decisions that patients face are not crucial. Even if the patient-clinician team makes the wrong choice (ie, they do not make the choice that would result from a full discussion), the adverse consequences are minimal or at least limited. Rather than devoting time to these situations, busy clinicians may choose to focus their efforts to ensure that decisions are

**BOX 18-4****Solutions to the Time Problem**

- Make time for discussion of key decisions
- Reserve special follow-up appointments for discussion
- Restrict time-consuming approaches to key decisions
- Reserve time-consuming approaches for important problems
- Reserve time-consuming approaches for difficult decisions
- Get help
- If possible, refer patient to colleagues with time and expertise for decision-making discussions
- Use decision aids

consistent with patients' values and preferences for choices associated with the most important consequences.

What may be unimportant for one patient, however, may be critical for another. Consider a farmer with an irritating but benign lesion on his hand and the rapidity with which the dermatologist would decide to freeze the lesion after obtaining patient consent. Now consider how the same dermatologist would consider treatment approaches for a similar skin lesion, this time in a woman working as a hand model. The dermatologist will have to engage in much more than a cursory consent procedure to care for this patient, who is likely to place a much greater value on avoiding a visible scar than on avoiding costly cosmetic procedures compared with almost all other patients with the same lesion.

**Straightforward vs Difficult Decisions**

When the decision is straightforward (ie, there is an option that almost all informed patients would choose because it is highly effective in achieving *patient-important outcomes*, easy to administer, inexpensive, and safe), decision making can be

expeditious. This is the case for aspirin use in a patient in the emergency department who has an acute coronary syndrome. Under these circumstances, a single sentence explaining the rationale and plan can suffice.

In other situations, the benefits and downsides of an intervention are more closely balanced. For instance, clinicians should have a discussion regarding use of low-dose aspirin for coronary prevention. Use of this agent is associated with bleeding, a risk that increases as the coronary risk increases. This downside must therefore be considered against the potential benefits, including the favorable effects of aspirin on coronary risk and colon cancer.<sup>34</sup>

These 2 situations—a clear decision that virtually all informed patients would endorse vs a close call—should correspond to strong and weak recommendations that guideline panels offer (see Chapter 17, How to Use a Patient Management Recommendation: Clinical Practice Guidelines and Decision Analyses). If guideline panels function appropriately, clinicians can interpret a strong recommendation as “just do it” and a weak recommendation as an invitation to engage patients in shared decision making. Sometimes, clinicians and patients need to spend more time making decisions that, when initially considered, appear straightforward. Some decisions, such as lifestyle and pharmacologic treatments for chronic conditions, require review—and reaffirmation or revision. The need for review may occur every time patients learn about or experience a potential adverse effect, renew the prescription and pay for it, or learn about an alternative solution. Time and resources spent exploring these decisions may help patients remember why they started using these interventions in the first place and enhance their adherence to these treatments (this was the motivation behind the decision aid about statin use in patients with diabetes, described in Figure 18-3).

## Misinformed Participants

Clinicians may have a distorted perception of the evidence. Distortions can be the result of misleading marketing messages

that reach clinicians informally through colleagues or formally through industry-funded continuing medical education and office detailing. Misleading presentations of research findings in primary reports of research can distort clinicians' understanding of the evidence. Panels that develop guidelines may include experts whose recommendations are influenced by their conflicts of interest. This is particularly problematic when adherence to guidelines becomes linked to monetary incentives (ie, pay-for-performance programs). Patients may perceive something amiss when clinicians make treatment recommendations that are too expensive, too invasive, or too new. Such patients, if unable to participate fully, may forgo these treatments after the visit, lose trust in the clinician, or seek attention elsewhere.

Patients also may be misinformed. Distorted evidence reaches patients through advertisements in traditional media, lay medical or health publications, social networks, and the Internet and through misinformed or conflicted clinicians. Consider the more than 75% of patients who received a coronary stent for stable angina who, after receiving this treatment, reported their belief (contradicted by evidence warranting high confidence) that this treatment will reduce their risk of myocardial infarction and death.<sup>35</sup> Patients convinced of what they see in print may feel empowered to request a prescription from their clinician for interventions that they do not need or would not want if they were adequately informed. Given time and skill constraints, patients who seek attention knowing what they want may leave clinician offices with their wishes satisfied, whereas clinicians are left feeling uncomfortable about the course of action chosen.<sup>36</sup>

Clinicians should spend more time with information sources when they suspect their own understanding is limited or inaccurate and more time with their patients when they suspect their patients have a distorted knowledge base. Strategies to calibrate the clinicians' knowledge base may include the review of the evidence that supports claims of effectiveness and strong recommendations from a variety of information sources (see

Chapter 4, Finding Current Best Evidence) using the skills taught in the *Users' Guides to the Medical Literature*. Strategies to calibrate patients' knowledge are less clear but may include involving the patient in such evidence reviews. An alternative approach is, when they are available, to use evidence-based decision aids.

## The Patient With Multiple Chronic Conditions

Straightforward decisions about adding a new prevention or treatment intervention can become challenging in a patient with a chronic condition who is overwhelmed by health care options. This happens most often in patients who have multiple chronic conditions, a situation that is becoming increasingly common at a younger age, particularly among the disadvantaged.<sup>37</sup> For these patients, each option not only involves a set of potential benefits and harms inherent to that treatment but also brings an obligatory set of treatment monitoring and administration tasks that represent an incremental *burden of treatment*. The new intervention will have to compete for patient attention, energy, and time against the patient's established treatment program. The end result of this competition may include optimal or inadequate adherence to the new treatment or discontinuation of an established therapy.

Clinicians need to assess patients' capacity to face treatment burdens. Influences on this capacity include patients' resilience, literacy, physical and mental health, financial solvency, social capital, and level of support in their environment. Clinicians must consider not only the extent to which adding a new treatment is consistent with patient values and preferences but also how feasible the resulting regimen is. Treatments may need to be prioritized, with discontinuation of low-value interventions. Such treatments impose an important burden to the patient (difficult to administer, expensive, disabling adverse effects) in exchange for limited or unclear benefits (improving a biochemical or physiologic measure without a small or uncertain impact on quality of life or *prognosis*). Prioritization of the treatment program is another opportunity for collaborative deliberation between

clinician and patient. This effort, sometimes referred to as *minimally disruptive medicine*, seeks patient goals for health while imposing the smallest possible treatment burden on their lives.<sup>38</sup>

## Other Solutions

Clinicians could consider delaying making a decision and ask that it be considered during another visit, designated for that purpose. This assumes that clinicians are permitted to allot this time in their schedule for these additional focused visits. Another option is to refer the patient to a specialist colleague with time and expertise in shared decision making. Primary care teams may designate members of the team—physicians, nurses, pharmacists, or care managers—to focus on making decisions with patients with whom the team has developed a partnership. In some centers, decision coaches (often nurses or other health care professionals) provide detailed exploration of important decisions.<sup>39</sup>

## Use a Patient Decision Aid

Patients considering important decisions may benefit from educational material that they can take home and review with family, friends, and advisers. They then can return with questions and potentially with a final decision. There are more than 300 such patient decision aids in the Cochrane Inventory found at the Cochrane Decision Aid Registry (<http://decisionaid.ohri.ca/cochinvent.php>). This inventory, kept by investigators at the Ottawa Health Decision Centre, describes the decision aid and its purpose and offers contact information about each tool's developer and availability. Unfortunately, almost 80% of these tools have not been evaluated clinically.<sup>40</sup>

A more promising approach is to use decision aids in the clinical encounter. Such tools are optimally designed (often using user-centered approaches) for the specific context to be time sensitive and efficient. The number and nature of these tools (eg, issue cards, option grids) are now expanding as evidence accumulates of their effectiveness and feasible use in routine



care.<sup>24-28</sup> This evidence suggests that simple tools for use during the clinical encounter add, on average, approximately 3 minutes to a primary care consultation (examples are available at <http://shareddecisions.mayoclinic.org> and <http://www.optiongrid.org>).

## CONCLUSION

---

Evidence-based medicine maintains that patient management decisions should reflect both the best available evidence and the patients' values and preferences (see Chapter 2, What Is Evidence-Based Medicine?). It follows that choices should be those that patients would make in collaboration with clinicians who ensure that both they and their patients are optimally informed and who respect what is most important to patients. Achieving that goal represents a major challenge and a fruitful area for clinical research. Clinicians should be aware of the different approaches to clinical decision making and the need to tailor the approach to the individual patient. They should understand how evidence and preferences fit together in the decision-making process and use the limited evidence available to find the approaches that are right for them and for their patients.

## References

1. Gafni A, Charles C, Whelan T. The physician-patient encounter: the physician as a perfect agent for the patient versus the informed treatment decision-making model. *Soc Sci Med*. 1998;47(3):347-354.
2. Gafni A. When does a competent patient make an irrational choice [letter]? *N Engl J Med*. 1990;323(19):1354.
3. Kahneman D, Tversky A. Prospect theory: an analysis of decisions under risk. *Econometrica*. 1979;47(2):263-292.
4. Elwyn G, Edwards A, Eccles M, Rovner D. Decision analysis in patient care. *Lancet*. 2001;358(9281):571-574.
5. Gafni A, Birch S. Preferences for outcomes in economic evaluation: an economic approach to addressing economic problems. *Soc Sci Med*. 1995;40(6):767-776.

6. Charles C, Gafni A, Whelan T. Decision-making in the physician-patient encounter: revisiting the shared treatment decision-making model. *Soc Sci Med.* 1999;49(5):651-661.
7. Charles C, Gafni A, Whelan T. Shared decision-making in the medical encounter: what does it mean? (or it takes at least two to tango). *Soc Sci Med.* 1997;44(5):681-692.
8. Makoul G, Clayman ML. An integrative model of shared decision making in medical encounters. *Patient Educ Couns.* 2006;60(3):301-312.
9. Elwyn G, Tsulukidze M, Edwards A, Légaré F, Newcombe R. Using a 'talk' model of shared decision making to propose an observation-based measure: Observer OPTION 5 Item. *Patient Educ Couns.* 2013;93(2):265-271.
10. Devereaux PJ, Anderson DR, Gardner MJ, et al. Differences between perspectives of physicians and patients on anticoagulation in patients with atrial fibrillation: observational study. *BMJ.* 2001;323(7323):1218-1222.
11. Mazur DJ, Hickam DH, Mazur MD, Mazur MD. The role of doctor's opinion in shared decision making: what does shared decision making really mean when considering invasive medical procedures? *Health Expect.* 2005;8(2):97-102.
12. Frosch DL, May SG, Rendle KA, Tietbohl C, Elwyn G. Authoritarian physicians and patients' fear of being labeled 'difficult' among key obstacles to shared decision making. *Health Aff (Millwood).* 2012;31(5):1030-1038.
13. Gaston CM, Mitchell G. Information giving and decision-making in patients with advanced cancer: a systematic review. *Soc Sci Med.* 2005;61(10):2252-2264.
14. Levinson W, Kao A, Kuby A, Thisted RA. Not all patients want to participate in decision making: a national study of public preferences. *J Gen Intern Med.* 2005;20(6):531-535.
15. Beaver K, Bogg J, Luker KA. Decision-making role preferences and information needs: a comparison of colorectal and breast cancer. *Health Expect.* 1999;2(4):266-276.
16. Montori VM, Rothman RL. Weakness in numbers. The challenge of numeracy in health care. *J Gen Intern Med.* 2005;20(11):1071-1072.
17. Whelan T, Gafni A, Charles C, Levine M. Lessons learned from the Decision Board: a unique and evolving decision aid. *Health Expect.* 2000;3(1):69-76.
18. Stacey D, Légaré F, Col NF, et al. Decision aids for people facing health treatment or screening decisions. *Cochrane Database of Syst Rev.* 2014;1:CD001431. doi: 10.1002/14651858.CD001431.pub4.
19. Charles C, Gafni A, Whelan T, O'Brien MA. Treatment decision aids: conceptual issues and future directions. *Health Expect.* 2005;8(2):114-125.
20. Elwyn G, O'Connor A, Stacey D, et al; International Patient Decision Aids Standards (IPDAS) Collaboration. Developing a quality criteria framework for patient decision aids: online international Delphi consensus process. *BMJ.* 2006;333(7565):417.
21. Joseph-Williams N, Newcombe R, Politi M, et al. Toward minimum standards for certifying patient decision aids: a modified delphi consensus process. *Med Dec Making.* 2013;34(6):699-710.

22. Elwyn G, Scholl I, Tietbohl C, et al. "Many miles to go ..." a systematic review of the implementation of patient decision support interventions into routine clinical practice. *BMC Med Inform Decis*. 2013;13(suppl 2):S14. doi: 10.1186/1472-6947-13-S2-S14.
23. Elwyn G, Frosch D, Volandes AE, Edwards A, Montori VM. Investing in deliberation: a definition and classification of decision support interventions for people facing difficult health decisions. *Med Decis Making*. 2010;30(6):701-711.
24. Weymiller AJ, Montori VM, Jones LA, et al. Helping patients with type 2 diabetes mellitus make treatment decisions: statin choice randomized trial. *Arch Intern Med*. 2007;167(10):1076-1082.
25. Mullan RJ, Montori VM, Shah ND, et al. The diabetes mellitus medication choice decision aid: a randomized trial. *Arch Intern Med*. 2009;169(17):1560-1568.
26. Montori VM, Shah ND, Pencille LJ, et al. Use of a decision aid to improve treatment decisions in osteoporosis: the osteoporosis choice randomized trial. *Am J Med*. 2011;124(6):549-556.
27. Hess EP, Knoedler MA, Shah ND, et al. The chest pain choice decision aid: a randomized trial. *Circ Cardiovasc Qual Outcomes*. 2012;5(3):251-259.
28. Branda ME, LeBlanc A, Shah ND, et al. Shared decision making for patients with type 2 diabetes: a randomized trial in primary care. *BMC Health Serv Res*. 2013;13:301. doi:10.1186/1472-6963-13-301.
29. Zuger A. Dissatisfaction with medical practice. *N Engl J Med*. 2004;350(1):69-75.
30. Yarnall KS, Pollak KI, Østbye T, Krause KM, Michener JL. Primary care: is there enough time for prevention? *Am J Public Health*. 2003;93(4):635-641.
31. Mechanic D, McAlpine DD, Rosenthal M. Are patients' office visits with physicians getting shorter? *N Engl J Med*. 2001;344(3):198-204.
32. Getz L, Sigurdsson JA, Hetlevik I. Is opportunistic disease prevention in the consultation ethically justifiable? *BMJ*. 2003;327(7413):498-500.
33. Légaré F, Ratté S, Gravel K, Graham ID. Barriers and facilitators to implementing shared decision-making in clinical practice: update of a systematic review of health professionals' perceptions. *Patient Educ Couns*. 2008;73(3):526-535.
34. Vandvik PO, Lincoff AM, Gore JM, et al; American College of Chest Physicians. Primary and secondary prevention of cardiovascular disease: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest*. 2012;141(2 suppl):e637S-e668S.
35. Rothberg MB, Sivalingam SK, Ashraf J, et al. Patients' and cardiologists' perceptions of the benefits of percutaneous coronary intervention for stable coronary disease. *Ann Intern Med*. 2010;153(5):307-313.
36. Mintzes B, Barer ML, Kravitz RL, et al. Influence of direct to consumer pharmaceutical advertising and patients' requests on prescribing decisions: two site cross sectional survey. *BMJ*. 2002;324(7332):278-279.

37. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *Lancet*. 2012;380(9836):37-43.
38. May C, Montori VM, Mair FS. We need minimally disruptive medicine. *BMJ*. 2009;339:b2803.
39. Woolf SH, Chan EC, Harris R, et al. Promoting informed choice: transforming health care to dispense knowledge for decision making. *Ann Intern Med*. 2005;143(4):293-300.
40. Ottawa Hospital Research Institute. Decision Aid Library Inventory (DALI). Patient Decision Aids website. <http://decisionaid.ohri.ca/cochinvent.php>. Updated June 25, 2012. Accessed August 4, 2014.

*This page intentionally left blank*

# GLOSSARY

Term	Definition
Absolute Difference	The absolute difference in rates of good or harmful outcomes between experimental groups (experimental group risk [EGR]) and control groups (control group risk [CGR]), calculated as the risk in the control group minus the risk in the experimental group ( $CGR - EGR$ ). For instance, if the rate of adverse events is 20% in the control group and 10% in the treatment group, the absolute difference is $20\% - 10\% = 10\%$ .
Absolute Risk (or Baseline Risk or Control Event Rate [CER])	The risk of an event (eg, if 10 of 100 patients have an event, the absolute risk is 10% expressed as a percentage and 0.10 expressed as a proportion).
Absolute Risk Increase (ARI)	The absolute difference in the risk of harmful outcomes between experimental groups (experimental group risk [EGR]) and control groups (control group risk [CGR]), calculated as the risk of harmful outcomes in the experimental group minus the rate of harmful outcomes in the control group ( $EGR - CGR$ ). Typically used to describe a harmful exposure or intervention (eg, if the rate of adverse outcomes is 20% in the treatment group and 10% in the control group, the absolute risk increase would be 10% expressed as a percentage and 0.10 expressed as a proportion). See also Absolute Risk Reduction; Number Needed to Harm.
Absolute Risk Reduction (ARR) (or Risk Difference [RD])	The absolute difference (risk difference) in risks of harmful outcomes between experimental groups (experimental group risk [EGR]) and control groups (control group risk [CGR]), calculated as the risk of harmful outcome in the control group minus the risk of harmful outcome in the experimental group ( $CGR - EGR$ ). Typically used to describe a beneficial exposure or intervention

*(Continued)*

Term	Definition
Absolute Risk Reduction (ARR) (or Risk Difference [RD]) ( <i>Continued</i> )	(eg, if 20% of patients in the control group have an adverse event, as do 10% among treated patients, the ARR or risk difference would be 10% expressed as a percentage and 0.10 expressed as a proportion).
Academic Detailing (or Educational Outreach Visits)	A strategy for changing clinician behavior. Use of a trained person who meets with health care professionals in their practice settings to provide information with the intent of changing their practice. The pharmaceutical industry frequently uses this strategy, to which the term “detailing” is applied. Academic detailing is such an interaction initiated by an academic group or institution rather than the pharmaceutical industry.
Additive	In genetic association studies, this describes any trait that increases proportionately in expression when comparing those with no copy, 1 copy, or 2 copies of that allele (ie, those with 1 copy of the allele show more of the trait than those without and, in turn, those with 2 copies show more of the trait than those with 1 copy).
Adherence (or Compliance)	The extent to which patients follow health care recommendations or the extent to which clinicians follow recommendations for use of diagnostic tests, monitoring equipment, interventional requirements, and other technical specifications that define optimal patient management.
Adjusted Analysis	An adjusted analysis takes into account differences in prognostic factors (or baseline characteristics) between groups that may influence the outcome. For instance, when comparing an experimental and control intervention, if the experimental group is on average older, and thus at higher risk of an adverse outcome than the control group, the analysis adjusted for age will have a larger treatment effect than the unadjusted analysis.

Term	Definition
Adjusted Indirect Comparison	A statistical technique that permits comparison between 2 interventions that have not been compared directly (head-to-head) but have both been compared to the same third comparator. This method preserves the principle of randomization.
Alerting (or Alerting Systems)	A strategy for changing clinician behavior. A type of computer decision support system that alerts the clinician to a circumstance that might require clinical action (eg, a system that highlights out-of-range laboratory values).
Algorithm	An explicit description of an ordered sequence of steps with branching logic that can be applied under specific clinical circumstances. The logic of an algorithm is if <i>a</i> , then do <i>x</i> ; if <i>b</i> , then do <i>y</i> ; etc.
Allele	In genetic association studies, this is one of several variants of a gene, usually referring to a specific site within the gene.
Allocation Concealment (or Concealment)	Randomization is concealed if the person who is making the decision about enrolling a patient is unaware of whether the next patient enrolled will be entered in the intervention or control group (using techniques such as central randomization or sequentially numbered opaque sealed envelopes). If randomization is not concealed, patients with differing prognoses may be differentially recruited to treatment or control groups. Of particular concern, patients with better prognoses may tend to be preferentially enrolled in the active treatment arm, resulting in exaggeration of the apparent benefit of the intervention (or even the false conclusion that the intervention is efficacious).
$\alpha$ Level (or Type I Error)	The probability of erroneously concluding that there is a difference between comparison groups when there is in fact no difference (also called a type I error). Typically, investigators decide on the chance of a false-positive result they are willing to accept when they plan the sample size for a study (eg, investigators often set $\alpha$ level at .05).



Term	Definition
Anchor Based	One way to establish the interpretability of measures of patient-reported outcomes is anchor based (the other is distribution based). Anchor-based methods require an independent standard, or anchor, that is itself interpretable and at least moderately correlated with the instrument being assessed. This anchor typically helps establish a minimum important difference of instruments that measure patient-reported outcomes.
Applicability	See Generalizability.
As-Treated Analysis	Includes patients according to the intervention they received rather than the intervention to which they were randomized. Thus, intervention group patients who received the control are counted in the control group, and control group patients who received the intervention are counted in the treatment group. This analysis is very likely to destroy the prognostic balance randomization achieved and provide misleading results.
Audit and Feedback	A strategy for changing clinician behavior. Any written or verbal summary of clinician performance (eg, based on medical record review or observation of clinical practice) during a specified period. The summary may also include recommendations to improve practice.
Background Questions	These clinical questions are about physiology, pathology, epidemiology, and general management and are often asked by clinicians in training. The answers to background questions are often best found in textbooks or narrative review articles.
Base Case	In an economic evaluation, the base case is the best estimates of each of the key variables that bear on the costs and effects of the alternative management strategies.

Term	Definition
Baseline Characteristics	Factors that describe study participants at the beginning of the study (eg, age, sex, disease severity). In comparison studies, it is important that these characteristics be initially similar between groups; if not balanced or if the imbalance is not statistically adjusted, these characteristics can cause confounding and can bias study results.
Baseline Risk (or Baseline Event Rate or Control Event Rate [CER])	The proportion or percentage of study participants in the control group in whom an adverse outcome is observed.
Bayesian Analysis	A statistical method that uses prior knowledge combined with data. See also Bayesian Diagnostic Reasoning.
Bayesian Diagnostic Reasoning	The essence of Bayesian reasoning is that one starts with a prior probability or probability distribution and incorporates new information to arrive at a posterior probability or probability distribution. The approach to diagnosis presented in the <i>Users' Guides to the Medical Literature</i> assumes that diagnosticians are intuitive Bayesian thinkers and move from pretest to posttest probabilities as information accumulates.
Before-After Design (or One-Group Pretest-Posttest Design)	A study in which the investigators compare the status of a group of study participants before and after the implementation of an intervention. In a controlled before-after study, investigators identify a control population with characteristics and performance similar to those of the study population. Data are collected and outcomes measured in both the study and control populations before and after the introduction of an intervention to the study population. Observed differences between groups in the postintervention period or in change scores (from baseline in each group) are assumed attributed to the intervention.

(Continued)

Term	Definition
Before-After Design (or One-Group Pretest-Posttest Design) (Continued)	In an uncontrolled before-after study, outcomes are measured before and after the introduction of an intervention in the same study setting. Observed differences in the outcomes are assumed to be attributable to the intervention.
$\beta$ Error (or Type II Error)	Otherwise known as type II error, $\beta$ error is the probability that a study will fail to rule out a null hypothesis when in fact that null hypothesis (typically that the treatment effect is 0; for instance, the relative risk is 1.0) is true. In other words, it is the probability of missing a true treatment effect. In sample-size calculations, $\beta$ is typically set at .2 or .1.
Bias (or Systematic Error)	<p>Systematic deviation from the underlying truth because of a feature of the design or conduct of a research study (eg, overestimation of a treatment effect because of failure to randomize). Sometimes, authors label specific types of bias in a variety of contexts.</p> <ol style="list-style-type: none"> <li>1. Channeling Effect or Channeling Bias: Tendency of clinicians to prescribe treatment according to a patient's prognosis. As a result of this behavior in observational studies, treated patients are more or less likely to be high-risk patients than untreated patients, leading to a biased estimate of treatment effect.</li> <li>2. Data Completeness Bias: Using a computer decision support system (CDSS) to log episodes in the intervention group and using a manual system in the non-CDSS control group can create variation in the completeness of data.</li> <li>3. Detection Bias (or Surveillance Bias): Tendency to look more carefully for an outcome in one of the comparison groups.</li> </ol>

(Continued)

Term	Definition
Bias (or Systematic Error) <i>(Continued)</i>	<ol style="list-style-type: none"> <li data-bbox="366 240 899 488">4. <b>Differential Verification Bias:</b> When test results influence the choice of the reference standard (eg, test-positive patients undergo an invasive test to establish the diagnosis, whereas test-negative patients undergo long-term follow-up without application of the invasive test), the assessment of test properties may be biased.</li> <li data-bbox="366 508 881 722">5. <b>Expectation Bias:</b> In data collection, an interviewer has information that influences his or her expectation of finding the exposure or outcome. In clinical practice, a clinician's assessment may be influenced by previous knowledge of the presence or absence of a disorder.</li> <li data-bbox="366 742 888 956">6. <b>Incorporation Bias:</b> Occurs when investigators use a reference standard that incorporates a diagnostic test that is the subject of investigation. The result is a bias toward making the test appear more powerful in differentiating target-positive from target-negative patients than it actually is.</li> <li data-bbox="366 976 891 1094">7. <b>Interviewer Bias:</b> Greater probing by an interviewer of some participants than others, contingent on particular features of the participants.</li> <li data-bbox="366 1115 905 1265">8. <b>Lead-Time Bias:</b> Occurs when outcomes such as survival, as measured from the time of diagnosis, may be increased not because patients live longer, but because screening lengthens the time that they know they have disease.</li> <li data-bbox="366 1286 909 1497">9. <b>Length-Time Bias:</b> Occurs when patients whose disease is discovered by screening may also appear to do better or live longer than people whose disease presents clinically with symptoms because screening tends to detect disease that is destined to progress slowly and that therefore has a good prognosis.</li> </ol>

*(Continued)*

Term	Definition
Bias (or Systematic Error) (Continued)	<p>10. Observer Bias: Occurs when an observer's observations differ systematically according to participant characteristics (eg, making systematically different observations in treatment and control groups).</p> <p>11. Partial Verification Bias: Occurs when only a selected sample of patients who underwent the index test is verified by the reference standard, and that sample is dependent on the results of the test. For example, patients with suspected coronary artery disease whose exercise test results are positive may be more likely to undergo coronary angiography (the reference standard) than those whose exercise test results are negative.</p> <p>12. Publication Bias: Occurs when the publication of research depends on the direction of the study results and whether they are statistically significant.</p> <p>13. Recall Bias: Occurs when patients who experience an adverse outcome have a different likelihood of recalling an exposure than patients who do not experience the adverse outcome, independent of the true extent of exposure.</p> <p>14. Referral Bias: Occurs when characteristics of patients differ between one setting (such as primary care) and another setting that includes only referred patients (such as secondary or tertiary care).</p> <p>15. Reporting Bias (or Selective Outcome Reporting Bias): The inclination of authors to differentially report research results according to the magnitude, direction, or statistical significance of the results.</p>

(Continued)

Term	Definition
Bias (or Systematic Error) <i>(Continued)</i>	<p>16. Social Desirability Bias: Occurs when participants answer according to social norms or socially desirable behavior rather than what is actually the case (for instance, underreporting alcohol consumption).</p> <p>17. Spectrum Bias: Ideally, diagnostic test properties will be assessed in a population in which the spectrum of disease in the target-positive patients includes all those in whom clinicians might be uncertain about the diagnosis, and the target-negative patients include all those with conditions easily confused with the target condition. Spectrum bias may occur when the accuracy of a diagnostic test is assessed in a population that differs from this ideal. Examples of spectrum bias would include a situation in which a substantial proportion of the target-positive population has advanced disease and target-negative participants are healthy or asymptomatic. Such situations typically occur in diagnostic case-control studies (eg, comparing those with advanced disease with healthy individuals). Such studies are liable to yield an overly sanguine estimate of the usefulness of the test.</p> <p>18. Surveillance Bias. See Detection Bias.</p> <p>19. Verification Bias. See Differential Verification Bias.</p> <p>20. Workup Bias. See Differential Verification Bias.</p>
Binary Outcome (or Dichotomous Outcome)	A categorical variable that can take 1 of 2 discrete values rather than an incremental value on a continuum (eg, pregnant or not pregnant, dead or alive).
Bivariable Regression Analysis	Regression when there is only 1 independent variable under evaluation with respect to a dependent variable. See also Multivariate Regression Analysis (or Multivariable Regression Analysis).

Term	Definition
Blind (or Blinded or Masked)	Patients, clinicians, data collectors, outcome adjudicators, or data analysts unaware of which patients have been assigned to the experimental or control group. In the case of diagnostic tests, those interpreting the test results are unaware of the result of the reference standard or vice versa.
Bonferroni correction	A statistical adjustment to the threshold $P$ value to adjust for multiple comparisons. The usual threshold for statistical significance ( $\alpha$ ) is 0.05. To perform a Bonferroni correction, one divides the critical $P$ value by the number of comparisons being made. For example, if 10 hypotheses are being tested, the new critical $P$ value would be $\alpha/10$ , usually 0.05/10 or 0.005. The Bonferroni correction represents a simple adjustment but is very conservative (ie, less likely than other methods to give a significant result).
Boolean Operators (or Logical Operators)	Words used when searching electronic databases. These operators are AND, OR, and NOT and are used to combine terms (AND/OR) or exclude terms (NOT) from the search strategy.
Bootstrap Technique	A statistical technique for estimating parameters, such as standard errors and confidence intervals, based on resampling from an observed data set with replacement from the original sample.
Burden	The term "burden" is used in 2 ways in the <i>Users' Guides to the Medical Literature</i> . One is burden of illness, which refers to the frequency of an illness in a population and its associated effect on quality of life, morbidity, mortality, and health care costs. Another is burden of treatment, which refers to the inconvenience of attending to the treatment's optimal use, of its monitoring, the limitations in lifestyle that it entails, and the possibility of interactions with other treatments.
Burden of Illness	See Burden.

Term	Definition
Burden of Treatment	See Burden.
Candidate Gene Study	A study that evaluates the association of specific genetic variants with outcomes or traits of interest, selecting the variants to be tested according to explicit considerations (known or postulated biology or function, previous studies, etc).
Case-Control Study	A study designed to determine the association between an exposure and outcome in which patients are sampled by outcome. Those with the outcome (cases) are compared with those without the outcome (controls) with respect to exposure to the suspected harmful agent.
Case Series	A report of a study of a collection of patients treated in a similar manner, without a control group. For example, a clinician might describe the characteristics of an outcome for 25 consecutive patients with diabetes who received education for prevention of foot ulcers.
Case Study	In qualitative research, an exploration of a case defined by some boundaries or contemporary phenomena, usually within a real-life context.
Categorical Variable	A categorical variable may be nominal or ordinal. Categorical variables can be defined according to attributes without any associated order (eg, medical admission, elective surgery, or emergency surgery); these are called nominal variables. A categorical variable can also be defined according to attributes that are ordered (eg, height, such as high, medium, or low); these are called ordinal variables.
Censoring	Censoring occurs when the value of a measurement or observation is only partially known. The problem of censored data, in which the observed value of some variables is partially known, is related to the problem of missing data. Many statistical methods can be used to estimate, impute, or otherwise model censored data.



Term	Definition
Chance-Corrected Agreement	The proportion of possible agreement achieved beyond that which one would expect by chance alone, often measured by the $\kappa$ statistic.
Chance-Independent Agreement	The proportion of possible agreement achieved that is independent of chance and unaffected by the distribution of ratings, as measured by the $\varphi$ statistic.
Channeling Effect or Channeling Bias	The tendency of clinicians to prescribe treatment according to a patient's prognosis. As a result of this behavior in observational studies, treated patients are more or less likely to be high-risk patients than untreated patients, leading to a biased estimate of treatment effect. See also Bias.
Checklist Effect	The improvement seen in medical decision making because of more complete and structured data collection (eg, clinicians fill out a detailed form, so their decisions improve).
$\chi^2$ Test	A nonparametric test of statistical significance used to compare the distribution of categorical outcomes in 2 or more groups, the null hypothesis of which is that the underlying distributions are identical.
Chromosome	Self-replicating structures in the nucleus of a cell that carry genetic information.
Class Effect (or Drug Class Effect)	When similar effects are produced by most or all members of a class of drugs (eg, $\beta$ -blockers or calcium antagonists).
Clinical Decision Rules (or Decision Rules, Clinical Prediction Rules, or Prediction Rules)	A guide for practice that is generated by initially examining, and ultimately combining, a number of variables to predict the likelihood of a current diagnosis or a future event. Sometimes, if the likelihood is sufficiently high or low, the rule generates a suggested course of action.
Clinical Decision Support System	A strategy for changing clinician behavior. An information system used to integrate clinical and patient information and provide support for decision making in patient care. See also Computer Decision Support System.

Term	Definition
Clinical Practice Guidelines (or Guidelines or Practice Guidelines)	A strategy for changing clinician behavior. Systematically developed statements or recommendations to assist clinician and patient decisions about appropriate health care for specific clinical circumstances.
Cluster Analysis	A statistical procedure in which the unit of analysis matches the unit of randomization, which is something other than the patient or participant (eg, school, clinic). See also Cluster Assignment (or Cluster Randomization).
Cluster Assignment (or Cluster Randomization)	The assignment of groups (eg, schools, clinics) rather than individuals to intervention and control groups. This approach is often used when assignment by individuals is likely to result in contamination (eg, if adolescents within a school are assigned to receive or not receive a new sex education program, it is likely that they will share the information they learn with one another; instead, if the unit of assignment is schools, entire schools are assigned to receive or not receive the new sex education program). Cluster assignment is typically randomized, but it is possible (though not advisable) to assign clusters to treatment or control by other methods.
Cochrane Collaboration	An international network working to help health care practitioners, policymakers, patients, patient advocates, and caregivers make well-informed decisions about health care, by preparing, updating, and promoting the accessibility of more than 5000 Cochrane Reviews, published online in the <i>Cochrane Database of Systematic Reviews</i> , as part of <i>The Cochrane Library</i> . The Cochrane Collaboration also prepares records of randomized clinical trials, in a database called CENTRAL, as part of <i>The Cochrane Library</i> .
Cochrane Q	A test for heterogeneity that assumes the null hypothesis that all of the apparent variability among individual study results is due to chance.

(Continued)

Term	Definition
Cochrane Q (Continued)	Cochrane Q generates a probability, presented as a <i>P</i> value, based on a $\chi^2$ distribution, that between-study differences in results equal to or greater than those observed are likely to occur simply by chance. See also <i>I</i> <sup>2</sup> Statistic.
Coefficient	See Correlation Coefficient.
Coherence	The agreement in treatment effect estimates between direct and indirect evidence, as in network meta-analyses.
Cohort	A group of persons with a common characteristic or set of characteristics. Typically, the group is followed for a specified period to determine the incidence of a disorder or complications of an established disorder (prognosis).
Cohort Study (or Longitudinal Study or Prospective Study)	This is an investigation in which a cohort of individuals who do not have evidence of an outcome of interest but who are exposed to the putative cause is compared with a concurrent cohort of individuals who are also free of the outcome but not exposed to the putative cause. Both cohorts are then followed forward in time to compare the incidence of the outcome of interest. When used to study the effectiveness of an intervention, it is an investigation in which a cohort of individuals who receive the intervention is compared with a concurrent cohort who does not receive the intervention, wherein both cohorts are followed forward to compare the incidence of the outcome of interest. Cohort studies can be conducted retrospectively in the sense that someone other than the investigator has followed patients, and the investigator obtains the database and then examines the association between exposure and outcome.
Cointerventions	Interventions other than the intervention under study that affect the outcome of interest and that may be differentially applied to intervention and control groups and thus potentially bias the result of a study.

Term	Definition
Comorbidity	Diseases or conditions that coexist in study participants in addition to the index condition that is the subject of the study.
Compliance (or Adherence)	See Adherence.
Composite End Point (or Composite Outcome)	When investigators measure the effect of treatment on an aggregate of end points of various levels of importance, this is a composite end point. Inferences from composite end points are strongest in the rare situations in which (1) the component end points are of similar patient importance, (2) the end points that are more important occur with at least similar frequency to those that are less important, and (3) strong biologic rationale supports results that, across component end points, reveal similar relative risks with sufficiently narrow confidence intervals.
Computer Decision Support System (CDSS)	A strategy for changing clinician behavior. Computer-based information systems are used to integrate clinical and patient information and provide support for decision making in patient care. In clinical decision support systems that are computer based, detailed individual patient data are entered into a computer program and are sorted and matched to programs or algorithms in a computerized database, resulting in the generation of patient-specific assessments or recommendations. Computer decision support systems can have the following purposes: alerting, reminding, critiquing, interpreting, predicting, diagnosing, and suggesting. See also Clinical Decision Support System.
Concealment (or Allocation Concealment)	See Allocation Concealment.
Concepts	The basic building blocks of theory.
Conceptual Framework	An organization of interrelated ideas or concepts that provides a system of relationships between those ideas or concepts.

Term	Definition
Conditional Probabilities	The probability of a particular state, given another state (ie, the probability of A, given B).
Confidence Interval (CI)	The range of values within which it is probable that the true value of a parameter (eg, a mean, a relative risk) lies.
Conflict of Interest	A conflict of interest exists when investigators, authors, institutions, reviewers, or editors have financial or nonfinancial relationships with other persons or organizations (such as study sponsors), or personal investments in research projects or the outcomes of projects that may inappropriately influence their interpretation or actions. Conflicts of interest can lead to biased design, conduct, analysis, and interpretation of study results as well as bias in review articles and opinion-based articles.
Confounder (or Confounding Variable or Confounding)	A factor that is associated with the outcome of interest and is differentially distributed in patients exposed and unexposed to the outcome of interest.
Consecutive Sample (or Sequential Sample)	A sample in which all potentially eligible patients treated throughout a period are enrolled.
Consequentialist (or Utilitarian)	A consequentialist or utilitarian view of distributive justice contends that, even in individual decision making, the clinician should take a broad social view, favoring actions that provide the greatest good to the greatest number. In this broader view, the effect on others of allocating resources to a particular patient's care would bear on the decision. This is an alternative to the deontologic view.
Construct Validity	In measurement theory, a construct is a theoretically derived notion of the domain(s) we wish to measure. An understanding of the construct will lead to expectations about how an instrument should behave if it is valid. Construct validity therefore involves comparisons between the

(Continued)

Term	Definition
Construct Validity ( <i>Continued</i> )	instrument being evaluated and other measures (eg, characteristics of patients or other scores) and the logical relationships that should exist between them.
Contamination	Occurs when participants in either the experimental or control group receive the intervention intended for the other arm of the study.
Content Validity	The extent to which a measurement instrument represents all facets of a given social construct.
Continuous Variable (or Interval Data)	A variable that can theoretically take any value and in practice can take a large number of values with small differences between them (eg, height). Continuous variables are also sometimes called interval data.
Control Event Rate (CER) (or Baseline Risk or Baseline Event Rate)	See Baseline Risk.
Control Group	A group that does not receive the experimental intervention. In many studies, the control group receives either usual care or a placebo.
Control Group Risk (CGR)	The risk of an event occurring in the control group of a study.
Controlled Time Series Design (or Controlled Interrupted Time Series)	Data are collected at several times both before and after the intervention in the intervention group and at the same times in a control group. Data collected before the intervention allow the underlying trend and cyclical (seasonal) effects to be estimated. Data collected after the intervention allow the intervention effect to be estimated while accounting for underlying secular trends. Use of a control group addresses the greatest threat to the validity of a time series design, which is the occurrence of another event at the same time as the intervention, both of which may be associated with the outcome.

Term	Definition
Convenience Sample	A sample of participants chosen primarily for their convenience to the researcher rather than for their salience to the research questions or the analysis. This is generally considered a scientifically inferior sampling approach to probability sampling in quantitative research or purposive sampling in qualitative research.
Correlation	The magnitude of the association between 2 variables. The strength of the association is described by the correlation coefficient. See also Correlation Coefficient.
Correlation Coefficient	A numeric expression (eg, $r^2$ or $R^2$ ) of the magnitude and direction of the association between 2 variables, which can take values from $-1.0$ (perfect negative relationship) to $0$ (no relationship) to $1.0$ (perfect positive relationship). If the analysis is bivariable, the correlation coefficient may be indicated as $r$ and the coefficient of determination is $r^2$ , and if the correlation coefficient is derived from multivariable (or multivariate) analysis, the correlation coefficient may be indicated as $R$ and the coefficient of determination is $R^2$ .
Cost Analysis	An economic analysis in which only costs of various alternatives are compared. This comparison informs only the resource-use half of the decision (the other half being the expected outcomes).
Cost-Benefit Analysis	An economic analysis in which both the costs and the consequences (including increases in the length and quality of life) are expressed in monetary terms.
Cost-Effectiveness Acceptability Curve	The cost-effectiveness acceptability is plotted on a graph that relates the maximum amount one is willing to pay for a particular treatment alternative (eg, how many dollars one is willing to pay to gain 1 life-year) on the horizontal axis to the probability that a treatment alternative is cost-effective compared with all other treatment alternatives on the vertical axis.

(Continued)

Term	Definition
Cost-Effectiveness Acceptability Curve (Continued)	The curves are generated from uncertainty around the point estimates of costs and effects in trial-based economic evaluations or uncertainty around values for variables used in decision analytic models. As one is willing to pay more for health outcomes, treatment alternatives that initially might be considered unattractive (eg, a high cost per life-year saved) will have a higher probability of becoming more cost-effective. Cost-effectiveness acceptability curves are a convenient method of presenting the effect of uncertainty on economic evaluation results on a single figure instead of through the use of numerous tables and figures of sensitivity analyses.
Cost-Effectiveness Analysis	An economic analysis in which the consequences are expressed in natural units (eg, cost per life saved or cost per bleeding event averted). Sometimes, cost-utility analysis is classified as a subcategory of cost-effectiveness analysis.
Cost-Effectiveness Efficiency Frontier	The cost and effectiveness results of each treatment alternative from an economic evaluation can be graphed on a figure known as the cost-effectiveness plane. The cost-effectiveness plane plots cost on the vertical axis (ie, positive infinity at the top and negative infinity at the bottom) and effects such as life-years on the horizontal axis (ie, negative infinity at the far left and positive infinity at the far right). One treatment alternative, such as usual care, is plotted at the origin (ie, 0, 0), and all other treatment alternatives are plotted relative to the treatment at the origin. Treatment alternatives are considered dominated if they have both higher costs and lower effectiveness relative to any other. Line segments can be drawn connecting the nondominated treatment alternatives, and the combination of line segments that join these nondominated treatment alternatives is referred to as the cost-effectiveness efficiency frontier. Constructed in this way, any treatment

(Continued)



Term	Definition
Cost-Effectiveness Efficiency Frontier (Continued)	alternative that lies above the cost-effectiveness efficiency frontier is considered to be inefficient (dominated) by a treatment alternative or combination of alternatives on the efficiency frontier.
Cost-Minimization Analysis	An economic analysis conducted in situations in which the consequences of the alternatives are identical and the only issue is their relative costs.
Cost-to-Charge Ratio	Where there is a systematic deviation between costs and charges, an economic analysis may adjust charges using a cost-to-charge ratio to approximate real costs.
Cost-Utility Analysis	A type of economic analysis in which the consequences are expressed in terms of life-years adjusted by peoples' preferences. Typically, one considers the incremental cost per incremental gain in quality-adjusted life-years (QALYs).
Cox Regression Model	A regression technique that allows adjustment for known differences in baseline characteristics or time-dependent characteristics between 2 groups applied to survival data.
Credibility (or Trustworthiness)	In qualitative research, a term used (in preference to quantitative terms such as "validity") to reflect the extent to which readers can trust researchers' empirical interpretations or descriptions as sound and insightful. Signs of credibility can be found not only in the procedural descriptions of methods but also through an assessment of the coherence and depth of the findings reported.
Credible Intervals	The Bayesian analogy to confidence intervals.
Criterion Standard (or Gold Standard or Reference Standard)	A method having established or widely accepted accuracy for determining a diagnosis that provides a standard to which a new screening or diagnostic test can be compared. The method need not be a single or simple procedure but could include patient follow-up to observe the evolution of a condition or the consensus of an adjudication committee about the patient's outcome.

Term	Definition
Critical Theory	A qualitative research tradition focused on understanding the nature of power relationships and related constructs, often with the intention of helping to remedy systemic injustices in society.
Critiquing (or Critiquing System)	A strategy for changing clinician behavior. A decision support approach in which the computer evaluates a clinician's decision and generates an appropriateness rating or an alternative suggestion.
Cronbach $\alpha$ Coefficient	Cronbach $\alpha$ is an index of reliability, homogeneity, or internal consistency of items on a measurement instrument. The Cronbach $\alpha$ increases with the magnitude of the interitem correlation and with the number of items.
Cross-Sectional Study	The observation of a defined population at a single point in time or during a specific interval. Exposure and outcome are determined simultaneously.
Data Completeness Bias	Using a computer decision support system (CDSS) to log episodes in the intervention group and using a manual system in the non-CDSS control group can create variation in the completeness of data. See also Bias.
Data Dredging	Searching a data set for differences among groups on particular outcomes, or in subgroups of patients, without explicit a priori hypotheses.
Decision Aid	A tool that presents patients with the benefits and harms of alternative courses of action in a manner that is quantitative, comprehensive, and understandable.
Decision Analysis	A systematic approach to decision making under conditions of uncertainty. It involves identifying all available alternatives and estimating the probabilities of potential outcomes associated with each alternative, valuing each outcome, and, on the basis of the probabilities and values, arriving at a quantitative estimate of the relative merit of each alternative.

Term	Definition
Decision Rules (or Clinical Decision Rules)	See Clinical Decision Rules.
Decision Tree	Most clinical decision analyses are built as decision trees; articles usually will include 1 or more diagrams showing the structure of the decision tree used for the analysis.
Degrees of Freedom	A technical term in a statistical analysis that has to do with the power of the analysis. The more degrees of freedom, the more powerful the analysis. The degrees of freedom typically refer to the number of observations in a sample minus the number of unknown parameters estimated for the model. It reflects a sort of adjusted sample size, with the adjustment based on the number of unknowns that need to be estimated in a model. For example, in a 2-sample <i>t</i> test, the degrees of freedom is $n_1 + n_2 - 1 - 1$ because there are $n_1 + n_2$ subjects altogether and 1 mean estimated in one group and 1 mean in another, giving $n_1 + n_2 - 2$ .
Deontologic	A deontologic approach to distributive justice holds that the clinician's only responsibility should be to best meet the needs of the individual under his or her care. This is an alternative to the consequentialist or utilitarian view.
Dependent Variable (or Outcome Variable or Target Variable)	The target variable of interest. The variable that is hypothesized to depend on or be caused by another variable, the independent variable.
Detection Bias (or Surveillance Bias)	The tendency to look more carefully for an outcome in one of the comparison groups. See also Bias.
Determinants of Outcome	The factors most strongly determining whether a target event will occur.
Dichotomous Outcome (or Binary Outcome)	A categorical variable that can take 1 of 2 discrete values rather than an incremental value on a continuum (eg, pregnant or not pregnant, dead or alive).

Term	Definition
Differential Diagnosis (or Active Alternatives)	The set of diagnoses that can plausibly explain a patient's presentation.
Differential Verification Bias (or Verification Bias or Workup Bias)	When test results influence the choice of the reference standard (eg, test-positive patients undergo an invasive test to establish the diagnosis, whereas test-negative patients undergo long-term follow-up without application of the invasive test), the assessment of test properties may be biased. See also Bias.
Dimensional Analysis	One of several possible approaches to analysis in grounded theory research, in which complex phenomena are characterized in terms of component parts (attributes, context, conditions, processes or actions, meanings).
Directness	A key element to consider when grading the quality of evidence for a health care recommendation. Evidence is direct to the extent that study participants, interventions, and outcome measures are similar to those of interest.
Direct Observation	See Field Observation.
Discriminant Analysis	A statistical technique similar to logistic regression analysis that identifies variables that are associated with the presence or absence of a particular categorical (nominal) outcome.
Disease-Specific Health-Related Quality of Life	See Health-Related Quality of Life.
Distribution Based	One way to establish the interpretability of measures of patient-reported outcomes is distribution based (the other is anchor based). Distribution-based methods interpret results in terms of the relation between the magnitude of observed effect and some measure of variability in instrument scores. The magnitude of effect may be the difference in patients' scores

(Continued)

Term	Definition
Distribution Based (Continued)	before and after treatment or the difference in end point scores. As a measure of variability, investigators may choose between-patient variability (eg, the SD of scores measured in patients at baseline) or within-patient variability (eg, the SD of change in scores that patients experienced during a study).
Document Analysis	In qualitative research, this is 1 of 3 basic data collection methods. It involves the interpretive review of written material.
Dominant	In genetic association studies, this describes any trait that is expressed in a heterozygote (ie, one copy of that allele is sufficient to manifest its effect).
Dominate	In economic evaluation, if the intervention of interest is both more effective and less costly than the control strategy, it is said to dominate the alternative.
Dose-Response Gradient (or Dose Dependence)	Exists when the risk of an outcome changes in the anticipated direction as the quantity or the duration of exposure to the putative harmful or beneficial agent increases.
Downstream Costs	Costs of resources consumed in the future and associated with clinical events in the future that are attributable to the intervention.
Drug Class Effects (or Class Effects)	See Class Effects.
Ecologic Study	Ecologic studies examine relationships between groups of individuals with exposure to a putative risk factor and an outcome. Exposures are measured at the population, community, or group level rather than at the individual level. Ecologic studies can provide information about an association; however, they are prone to bias: the ecologic fallacy. The ecologic fallacy holds that relationships observed for groups

(Continued)

Term	Definition
Ecologic Study (Continued)	necessarily hold for individuals (eg, if countries with more dietary fat have higher rates of breast cancer, then women who eat fatty foods must be more likely to get breast cancer). These inferences may be correct but are only weakly supported by the aggregate data.
Economic Analysis (or Economic Evaluation)	A set of formal, quantitative methods used to compare 2 or more treatments, programs, or strategies with respect to their resource use and their expected outcomes.
Educational Meetings (or Interactive Workshops)	A strategy for changing clinician behavior. Participation of professionals in workshops that include interaction and discussion.
Educational Outreach Visits (or Academic Detailing)	See Academic Detailing.
Effect Size	The difference in outcomes between the intervention and control groups divided by some measure of variability, typically the standard deviation.
Efficacy Analysis (Effectiveness Analysis)	This analysis includes the subset of patients in the trial who received the intervention of interest, regardless of initial randomization, and who do not have missing data for any reason. This approach is ill-named in that it does not tell one about either efficacy or effectiveness because it compromises the prognostic balance that randomization achieves and is therefore likely to provide a biased estimate of treatment effect.
Efficiency	Technical efficiency is the relationship between inputs (costs) and outputs (in health, quality-adjusted life-years [QALYs]). Interventions that provide more QALYs for the same or fewer resources are more efficient. Technical efficiency is assessed using cost minimization,

(Continued)

Term	Definition
Efficiency (Continued)	cost-effectiveness, and cost-utility analysis. Allocative efficiency recognizes that health is not the only goal that society wishes to pursue, so competing goals must be weighted and then related to costs. This is typically done through cost-benefit analysis.
Efficiency Frontier	When the cost and effectiveness results of an economic evaluation are graphed on a cost-effectiveness plane along with incremental cost-effectiveness ratios, the resultant line segments are referred to as the efficiency frontier. Any strategy that has a base-case cost-effectiveness that is above the efficiency frontier would be considered dominated.
End Point	An event or outcome that leads to completion or termination of follow-up of an individual in a study (eg, death or major morbidity).
Equivalence Study (or Equivalence Trial)	Trials that estimate treatment effects that exclude any patient-important superiority of interventions under evaluation are equivalence trials. Equivalence trials require a priori definition of the smallest difference in outcomes between these interventions that patients would consider large enough to justify a preference for the superior. The confidence interval for the estimated treatment effect at the end of the trial should exclude that difference for the authors to claim equivalence. Equivalence trials are helpful when investigators want to see whether a cheaper, safer, or simpler (or, increasingly often, better method to generate income for the sponsor) intervention is neither better nor worse (in terms of efficacy) than a current intervention.
Ethnography (or Ethnographic Study)	In qualitative research, an approach to inquiry that focuses on the culture or subculture of a group of people to try to understand the world view of those under study.

Term	Definition
Evidence	A broad definition of evidence is any empirical observation, whether systematically collected or not. The unsystematic observations of the individual clinician constitute one source of evidence. Physiologic experiments constitute another source. Clinical research evidence refers to systematic observation of clinical events and is the focus of the <i>Users' Guides to the Medical Literature</i> .
Evidence-Based Experts	Clinicians, who can, in a sophisticated manner, independently find, appraise, and judiciously apply the best evidence to patient care.
Evidence-Based Health Care (EBHC)	The conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. Evidence-based clinical practice requires integration of individual clinical expertise and patient preferences with the best available external clinical evidence from systematic research and consideration of available resources.
Evidence-Based Medicine (EBM)	Evidence-based medicine can be considered a subcategory of evidence-based health care, which also includes other branches of health care practice, such as evidence-based nursing or evidence-based physiotherapy. Subcategories of EBM include evidence-based surgery and evidence-based cardiology. See also Evidence-Based Health Care.
Evidence-Based Policy Making	Policy making is evidence based when practice policies (eg, use of resources by clinicians), service policies (eg, resource allocation, pattern of services), and governance policies (eg, organizational and financial structures) are based on research evidence of benefit or cost benefit.
Evidence-Based Practice (EBP)	Evidence-based practice is clinical practice in which patient management decisions are consistent with the principles of evidence-based health care. This means that decisions will be, first of all, consistent with the best evidence

(Continued)



Term	Definition
Evidence-Based Practice (EBP) (Continued)	about the benefits and downsides of the alternative management strategies. Second, decisions will be consistent with the values and preferences of the individual patient.
Evidence-Based Practitioners	Clinicians who can differentiate evidence-based summaries and recommendations from those that are not evidence-based and understand results sufficiently well to apply them judiciously in clinical care, ensuring decisions are consistent with patients' values and preferences.
Evidence Profile	An evidence profile is a tabular or list summary of a body of evidence addressing a structured clinical question of alternative management strategies. It includes, at minimum, the number of studies and patients, the study design(s), the reasons for increasing or decreasing confidence ratings in estimates, and measures of relative and absolute effect. The evidence profile is an expanded version of the summary-of-findings table.
Evidentialism	A theory of knowledge that holds that the justification or reason of a belief is determined by the quality of the believer's evidence for the belief.
Exclusion Criteria	The characteristics that render potential participants ineligible to participate in a study or that render studies ineligible for inclusion in a systematic review.
Expectation Bias	In data collection, an interviewer has information that influences his or her expectation of finding the exposure or outcome. In clinical practice, a clinician's assessment may be influenced by previous knowledge of the presence or absence of a disorder. See also Bias.
Experimental Therapy (or Experimental Treatment or Experimental Intervention)	A therapeutic alternative to standard or control therapy, which is often a new intervention or different dose of a standard drug.

Term	Definition
Exposure	A condition to which patients are exposed (either a potentially harmful intervention or a potentially beneficial one) that may affect their health.
Face Validity	The extent to which a measurement instrument appears to measure what it is intended to measure.
Fail-Safe N	The minimum number of undetected studies with negative results that would be needed to change the conclusions of a meta-analysis. A small fail-safe N suggests that the conclusion of the meta-analysis may be susceptible to publication bias.
False Negative	Those who have the target disorder, but the test incorrectly identifies them as not having it.
False Positive	Those who do not have the target disorder, but the test incorrectly identifies them as having it.
Federated Search Engine	A federated search engine searches several online information sources simultaneously and is especially useful when there is no single comprehensive, current, rigorous resource, as is currently the case for evidence-based health care. Examples of evidence-based federated search engines include ACCESSSS ( <a href="http://plus.mcmaster.ca/accesssss">http://plus.mcmaster.ca/accesssss</a> ) and Trip ( <a href="http://www.tripdatabase.com">http://www.tripdatabase.com</a> ).
Feedback Effect	The improvement seen in medical decision making because of performance evaluation and feedback.
Feeling Thermometer	A feeling thermometer is a visual analog scale presented as a thermometer, typically with markings from 0 to 100, with 0 representing death and 100 full health. Respondents use the thermometer to indicate their utility rating of their health state or of a hypothetical health state.

Term	Definition
Field Observation	In qualitative research, this is 1 of 3 basic data collection methods. It involves investigators witnessing and recording events as they occur. There are 3 approaches to field observation. With direct observation, investigators record detailed field notes from the milieu they are studying. In nonparticipant observation, the researcher participates relatively little in the interactions he or she is studying. In participant observation, the researcher assumes a role in the social setting beyond that of a researcher (eg, clinician, committee member).
Fixed-Effects Model	A model to generate a summary estimate of the magnitude of effect in a meta-analysis that restricts inferences to the set of studies included in the meta-analysis and assumes that a single true value underlies all of the primary study results. The assumption is that if all studies were infinitely large, they would yield identical estimates of effect; thus, observed estimates of effect differ from one another only because of random error. This model takes only within-study variation into account and not between-study variation.
Focus Group	See Interview.
Follow-up (or Complete Follow-up)	The extent to which investigators are aware of the outcome in every patient who participated in a study. If follow-up is complete, the outcome is known for all study participants.
Foreground Questions	These clinical questions are more commonly asked by seasoned clinicians. They are questions asked when browsing the literature (eg, what important new information should I know to optimally treat my patients?) or when problem solving (eg, defining specific questions raised in caring for patients and then consulting the literature to resolve these problems).

Term	Definition
Forest Plot	A forest plot is a graphic display that illustrates the magnitude of effect of an intervention vs a control in several studies. It provides a visual representation of the best estimate of effect and the range of plausible truth (confidence interval) for each study and for the pooled estimate combining all studies. A vertical line represents no effect. The area of each square or dot (typically representing individual studies) or diamond (typically representing the pooled estimates) is sometimes proportional to the study's weight in the meta-analysis.
Frequentist Analysis	A statistical approach that places the emphasis on available data (conventional approach to statistical analysis, contrast with Bayesian).
Funnel Plot	A graphic technique for assessing the possibility of publication bias in a systematic review. The effect measure is typically plotted on the horizontal axis and a measure of the random error associated with each study on the vertical axis. In the absence of publication bias, because of sampling variability, the graph should have the shape of a funnel. If there is bias against the publication of null results or results revealing an adverse effect of the intervention, one quadrant of the funnel plot will be partially or completely missing.
Generalizability (or Applicability)	The degree to which the results of a study can be generalized to settings or samples other than the ones studied.
Generic Health-Related Quality of Life	See Health-Related Quality of Life.
Genetic Association Study	A study that attempts to identify and characterize genomic variants underlying the susceptibility to multifactorial disease.
Genome	The entire collection of genetic information (or genes) that an organism possesses.

Term	Definition
Genome-wide Association Study (GWAS)	A study that evaluates the association of genetic variation with outcomes or traits of interest by using 100 000 to 1 000 000 or more markers across the genome.
Genotype	The genetic constitution of an individual, either overall or at a specific gene.
Geometry of a Network	A graphic representation of the distribution of treatments and their comparisons across a network.
Gold Standard (or Reference Standard or Criterion Standard)	See Criterion Standard.
GRADE (Grading of Recommendations Assessment, Development and Evaluation)	The Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach is a system for rating the quality of evidence and strength of recommendations that is explicit, comprehensive, and increasingly adopted by guideline organizations. The system classifies the confidence in estimates of effect into 1 of 4 levels (high, moderate, low, or very low). Recommendations are graded as strong or weak.
Grounded Theory	In qualitative research, an approach to collecting and analyzing data with the aim of developing a theory grounded in real-world observations.
Haplotype	Alleles that tend to occur together on the same chromosome because of single-nucleotide polymorphisms being in proximity and therefore inherited together.
Harm	Adverse consequences of exposure to an intervention.
Hawthorne Effect	The tendency for human performance to improve when participants are aware that their behavior is being observed.

Term	Definition
Hazard Ratio (HR)	The weighted relative risk of an outcome (eg, death) during the entire study period; often reported in the context of survival analysis.
Health Costs (or Health Care Costs)	Health care resources that are consumed. These reflect the inability to use the same resources for other worthwhile purposes (opportunity costs).
Health Outcomes	All possible changes in health status that may occur for a defined population or that may be associated with exposure to an intervention. These include changes in the length and quality of life, major morbid events, and mortality.
Health Profile	A type of data collection tool, intended for use in the entire population (including the healthy, the very sick, and patients with any sort of health problem), that attempts to measure all important aspects of health-related quality of life (HRQL).
Health-Related Quality of Life (HRQL)	<ol style="list-style-type: none"> <li data-bbox="353 855 919 987">1. Health-Related Quality of Life (HRQL): Measurements of how people are feeling or the value they place on their health state. Such measurements can be disease specific or generic.</li> <li data-bbox="353 992 919 1158">2. Disease-Specific Health-Related Quality of Life: Disease-specific HRQL measures evaluate the full range of patients' problems and experiences relevant to a specific condition or disease.</li> <li data-bbox="353 1163 919 1419">3. Generic Health-Related Quality of Life: Generic HRQL measures contain items that cover all relevant areas of HRQL. They are designed for administration to people with any kind of underlying health problem (or no problem at all). Generic HRQL measures allow comparisons across diseases or conditions.</li> </ol>
Health State	The health condition of an individual or group during a specified interval (commonly assessed at a particular point).

Term	Definition
Heterogeneity	Differences among individual studies included in a systematic review, typically referring to study results; the term can also be applied to other study characteristics.
Heterozygous	An individual is heterozygous at a gene location if he or she has 2 different alleles (one on the maternal chromosome and one on the paternal) at that location.
Hierarchic Regression	Hierarchic regression examines the relation between independent variables or predictor variables (eg, age, sex, disease severity) and a dependent variable (or outcome variable) (eg, death, exercise capacity). Hierarchic regression differs from standard regression in that one predictor is a subcategory of another predictor. The lower-level predictor is nested within the higher-level predictor. For instance, in a regression predicting likelihood of withdrawal of life support in intensive care units (ICUs) participating in an international study, city is nested within country and ICU is nested within city.
Hierarchy of Evidence	A system of classifying and organizing types of evidence, typically for questions of treatment and prevention. Clinicians should look for the evidence from the highest position in the hierarchy.
Historiography	A qualitative research method concerned with understanding both historical events and approaches to the writing of historical narratives.
Homogeneity	The inverse of heterogeneity.
Homozygous	An individual is homozygous at a gene location if he or she has 2 identical alleles at that location.
$I^2$ Statistic	The $I^2$ statistic is a test of heterogeneity. $I^2$ can be calculated from Cochrane Q according to the formula: $I^2 = 100\% \times (\text{Cochrane } Q - \text{degrees of freedom})$ . Any negative values of $I^2$ are considered equal to 0, so that the range of $I^2$ values is 0% to 100%, indicating no heterogeneity to high heterogeneity, respectively.

Term	Definition
Imprecision	In rating the quality of evidence, GRADE (Grading of Recommendations Assessment, Development and Evaluation) suggests that examination of 95% confidence intervals (CIs) provides the optimal primary approach to decisions regarding imprecision. Decreasing the rating in the quality of evidence (ie, confidence in estimates of effect) is required if clinical action would differ if the upper vs the lower boundary of the CI represented the truth. An exception to this rule occurs when an effect is large, and consideration of CIs alone suggests a robust effect, but the total sample size is not large and the number of events is small. Under these circumstances, one should consider rating the quality of evidence down for imprecision.
Incidence	The number of new cases of disease that occur during a specified period, expressed as a proportion of the number of people at risk during that time.
Inclusion Criteria	The characteristics that define the population eligible for a study or that define the studies that will be eligible for inclusion in a systematic review.
Incoherence	The disagreement in treatment effect estimates between direct and indirect evidence, as in network meta-analyses.
Inconsistency	In the GRADE (Grading of Recommendations Assessment, Development and Evaluation) system of recommendations, a body of evidence is not rated up in quality for consistency but may be rated down in quality if inconsistent. Criteria for evaluating consistency include similarity of point estimates, extent of overlap of confidence intervals, and statistical criteria, including tests of heterogeneity and $I^2$ . To explore heterogeneity, a small number of a priori subgroups may be examined related to the population, intervention, outcomes, and risk of bias.



Term	Definition
Incorporation Bias	Occurs when investigators use a reference standard that incorporates a diagnostic test that is the subject of investigation. The result is a bias toward making the test appear more powerful in differentiating target-positive from target-negative patients than it actually is. See also Bias.
Incremental Cost-Effectiveness Ratio	The price at which additional units of benefit can be obtained.
Independent Association	When a variable is associated with an outcome after adjusting for multiple other potential prognostic factors (often after regression analysis), the association is an independent association.
Independent Variable	The variable that is believed to cause, influence, or at least be associated with the dependent variable.
Indicator Condition	A clinical situation (eg, disease, symptom, injury, or health state) that occurs reasonably frequently and for which there is sound evidence that high-quality care is beneficial. Indicator conditions can be used to evaluate quality of care by comparing the care provided (as assessed through medical record review or observation) to that which is recommended.
Indirect Costs and Benefits	The effect of alternative patient management strategies on the productivity of the patient and others involved in the patient's care.
Indirect Evidence	Evidence bearing on the relative effect of treatments that have not been compared directly against each other but have a common comparator. Indirect evidence may be evaluated using accepted statistical approaches, including adjusted indirect comparisons and network meta-analyses.
Indirectness	In rating confidence in estimates of effect (quality of evidence), the GRADE (Grading of Recommendations Assessment, Development

(Continued)

Term	Definition
Indirectness (Continued)	and Evaluation) approach suggests examining directness. Directness in GRADE has 2 elements. The first is the extent to which the research evidence is about the patients and interventions of interest and measuring outcomes important to patients. Rating down the confidence in estimates is required if evidence is sufficiently indirect, which occurs in 4 ways: (1) if patients differ from those of interest, (2) if interventions differ from those of interest, (3) if outcomes differ from those of interest to patients (eg, surrogate outcomes), and (4) if interventions have not been tested in head-to-head comparisons and, as a result, indirect comparisons are required.
Individual Patient Data Meta-analysis	A meta-analysis in which individual patient data from each primary study are used to create pooled estimates. Such an approach can facilitate more accurate intention-to-treat analyses and informed subgroup analyses.
Informational Redundancy	In qualitative research, the point in the analysis at which new data fail to generate new themes and new information becomes redundant. This is considered an appropriate stopping point for data collection in most methods and an appropriate stopping point for analysis in some methods.
Informed Consent	A participant's expression (verbal or written) of willingness, after full disclosure of the risks, benefits, and other implications, to participate in a study.
Intention-to-Treat Analysis, Intention-to-Treat Principle	Authorities differ on the definition of an intention-to-treat analysis. All agree that it means that patients for whom data are available are analyzed in the groups to which they are randomized irrespective of what treatment they received. How one handles those patients for whom data are not available (loss to follow-up) in an intention-to-treat analysis is controversial.

(Continued)

Term	Definition
Intention-to-Treat Analysis, Intention-to-Treat Principle (Continued)	The authors of the <i>Users' Guides to the Medical Literature</i> believe that the term "intention-to-treat" should be restricted to patients with follow-up data. Thus, how one handles those patients lost to follow-up should be an issue separate from intention-to-treat.
Internal Validity	Whether a study provides valid results depends on whether it was designed and conducted well enough that the study findings accurately represent the direction and magnitude of the underlying true effect (ie, studies that have higher internal validity have a lower likelihood of bias/systematic error).
Interrater Reliability	The extent to which 2 or more raters are able to consistently differentiate subjects with higher and lower values on an underlying trait (typically measured with an intraclass correlation).
Interrupted Time Series Design (or Time Series Design)	See Time Series Design.
Interval Data (or Continuous Variable)	See Continuous Variable.
Intervention Effect (or Treatment Effect)	See Treatment Effect.
Interview	In qualitative research, this is 1 of 3 basic data collection methods. It involves an interviewer asking questions to engage participants in dialogue to allow interpretation of experiences and events in the participants' own terms. The 2 most common interviews are interviews of individuals or focus groups, which are group interviews in which a researcher facilitates discussion among multiple participants. Statements and interactions are then used as data. In quantitative research, an interview is a method of collecting data in which an interviewer obtains information from a participant through conversation.

Term	Definition
Interviewer Bias	Greater probing by an interviewer of some participants than others, contingent on particular features of the participants. See also Bias.
Intraclass Correlation Coefficient	This is a measure of reproducibility that compares variance between patients to the total variance, including both between-patient and within-patient variance.
Intrarater Reliability	The extent to which a rater is able to consistently differentiate participants with higher and lower values of an underlying trait on repeated ratings over time (typically measured with an intraclass correlation).
Inverse Rule of 3s	A rough rule of thumb that tells us the following: If an event occurs, on average, once every $x$ days, we need to observe $3x$ days to be 95% confident of observing at least 1 event.
Investigator Triangulation	See Triangulation.
Isoform	Variant in the amino acid sequence of a protein.
Jackknife Technique (or Jackknife Dispersion Test)	A statistical technique for estimating the variance and bias of an estimator. It is applied to a predictive model that is derived from a study sample to determine whether the model fits different subsamples from the model equally well.
Judgmental Sampling (or Purposive Sampling or Purposeful Sampling)	See Purposive Sampling.
Kaplan-Meier Curve (or Survival Curve)	A graphical plot of the Kaplan-Meier statistical estimate of survival in a survival analysis. See also Survival Curve and Survival Analysis.
$\kappa$ Statistic (or Weighted $\kappa$ or $\kappa$ Value)	A measure of the extent to which observers achieve agreement beyond the level expected to occur by chance alone. The value ranges from 0 to 100, with 0 indicating no agreement and typically values greater than 75 indicating excellent agreement.

Term	Definition
Law of Multiplicative Probabilities	The law of multiplicative probabilities for independent events (where one event in no way influences the other) tells us that the probability of 10 consecutive heads in 10 coin flips can be found by multiplying the probability of a single head ( $1/2$ ) 10 times over; that is, $1/2$ , $1/2$ , $1/2$ , and so on.
Leading Hypothesis (or Working Diagnosis)	See Working Diagnosis.
Lead Time Bias	Occurs when outcomes such as survival, as measured from the time of diagnosis, may be increased not because patients live longer but because screening lengthens the time that they know they have disease. See also Bias.
Length Time Bias	Occurs when patients whose disease is discovered by screening also may appear to do better or live longer than people whose disease presents clinically with symptoms because screening tends to detect disease that is destined to progress slowly and that therefore has a good prognosis. See also Bias.
Levels of Evidence	A hierarchy of research evidence to inform practice, usually ranging from strongest to weakest.
Likelihood Ratio (LR)	For a screening or diagnostic test (including clinical signs or symptoms), the likelihood ratio (LR) expresses the relative likelihood that a given test result would be expected in a patient with, as opposed to one without, a disorder of interest. An LR of 1 means that the posttest probability is identical to the pretest probability. As LRs increase above 1, the posttest probability progressively increases in relation to the pretest probability. As LRs decrease below 1, the posttest probability progressively decreases in relation to the pretest probability. An LR is calculated as the proportion of target positive with a particular test result (which, with a single cut point, would be either a positive or negative result) divided by the proportion of target negative with the same test result.

Term	Definition
Likert Scales	Scales, typically with 3 to 9 possible values, that include extremes of attitudes or feelings (such as from totally disagree to totally agree) that respondents mark to indicate their rating.
Linear Regression	The term used for a regression analysis when the dependent variable or target variable is a continuous variable and the relationship between the dependent variable and independent variable is thought to be linear.
Linkage	The tendency of genes or other DNA sequences at specific loci to be inherited together as a consequence of their physical proximity on a single chromosome.
Linkage Disequilibrium	A measure of association between alleles at different loci.
Local Consensus Process	A strategy for changing clinician behavior. Inclusion of participating clinicians in discussions to create agreement with a suggested approach to change clinician practice.
Local Opinion Leaders (or Opinion Leaders)	A strategy for changing clinician behavior. These persons are clinician peers who are recognized by their colleagues as model caregivers or who are viewed as having particular content expertise.
Locus	The site(s) on a chromosome at which the gene for a particular trait is located or on a gene at which a particular single-nucleotide polymorphism is located.
Logical Operators (or Boolean Operators)	See Boolean Operators.
Logistic Regression	A regression analysis in which the dependent variable is binary.
Longitudinal Study (or Cohort Study or Prospective Study)	See Cohort Study.

Term	Definition
Lost to Follow-up	Patients whose status on the outcome or end point of interest is unknown.
Markov Model (or Multistate Transition Model)	Markov models are tools used in decision analyses. Named after a 19th-century Russian mathematician, Markov models are the basis of software programs that model what might happen to a cohort of patients during a series of cycles (eg, periods of 1 year). The model allows for the possibility that patients might move from one health state to another. For instance, one patient may have a mild stroke in one 3-month cycle, continue with minimal functional limitation for a number of cycles, have a gastrointestinal bleeding episode in a subsequent cycle, and finally experience a major stroke. Ideally, data from randomized trials will determine the probability of moving from one state to another during any cycle under competing management options.
Masked (or Blind or Blinded)	See Blind.
Matching	A deliberate process to make the intervention group and comparison group comparable with respect to factors (or confounders) that are extraneous to the purpose of the investigation but that might interfere with the interpretation of the study's findings. For example, in case-control studies, individual cases may be matched with controls on the basis of comparable age, sex, or other clinical features.
Median Survival	The length of time that half the study population survives.
Medical Subject Headings (MeSH)	The National Library of Medicine's controlled vocabulary used for indexing articles for MEDLINE/PubMed. Medical Subject Headings (MeSH) terms provide a consistent way to retrieve information that may use different terms for the same concepts.

Term	Definition
Member Checking	In qualitative research, this involves sharing draft findings with the participants to get feedback on whether the findings make sense to them, whether researchers interpreted their viewpoints faithfully, or whether they perceive errors of fact. Note that any discrepancies would not necessarily indicate that the research is biased or in error but rather that the next stage of empirical analysis should interpret and account for the discrepancies.
Messenger RNA	An RNA-containing single-strand copy of a gene that migrates out of the cell nucleus to the ribosome, where it is translated into a protein.
Meta-analysis	A statistical technique for quantitatively combining the results of multiple studies that measure the same outcome into a single pooled or summary estimate.
Meta-regression Analysis	A regression in which the dependent variable is the magnitude of treatment effect in individual studies and the independent variable is study characteristics. Meta-regression is used to determine whether study characteristics can explain differences in magnitude of treatment effect across studies. Meta-regression techniques can be used to explore whether patient characteristics (eg, younger or older patients) or design characteristics (eg, studies of low or high quality) are related to the size of the treatment effect.
Meta-synthesis	A procedure for combining qualitative research on a specific topic in which researchers compare and analyze the texts of individual studies and develop new interpretations.
Minimal Important Difference	The smallest difference in a patient-important outcome that patients perceive as beneficial and that would mandate, in the absence of troublesome adverse effects and excessive cost, a change in the patient's health care management.



Term	Definition
Minimally Disruptive Medicine	Medicine practiced to minimize the burden of treatment or intervention on the patient's life.
Mixed-Methods Study	A study that combines data collection approaches, sometimes both qualitative and quantitative, into the study methods and is commonly used in the study of service delivery and organization. Some mixed-methods studies combine study designs (eg, investigators may embed qualitative or quantitative process evaluations alongside quantitative evaluative designs to increase understanding of factors that influence a phenomenon). Some mixed-methods studies include a single overarching research design but use mixed-methods for data collection (eg, surveys, interviews, observation, and analysis of documentary material).
Model	The term "model" is often used to describe statistical regression analyses that involve more than 1 independent variable and 1 dependent variable. This is a multivariable or multiple regression (or multivariate) analysis.
Multifaceted Interventions	The use of multiple strategies to change clinician behavior. Multiple strategies may include a combination that includes 2 or more of the following: audit and feedback, reminders, local consensus processes, patient-mediated interventions, or computer decision support systems.
Multistate Transition Model	See Markov Model.
Multivariate Regression Analysis (or Multivariable Regression Analysis)	A type of regression that provides a mathematical model that attempts to explain or predict the dependent variable (or outcome variable or target variable) by simultaneously considering 2 or more independent variables (or predictor variables). Multivariable refers to multiple predictors (independent variables) for a single outcome (dependent variable). Multivariate refers to 1 or more independent variables for multiple outcomes. See also Bivariable Regression.

<b>Term</b>	<b>Definition</b>
Mutation	A rare variant in a gene, occurring in less than 1% of a population. See Polymorphism.
Narrative Review	A review article (such as a typical book chapter) that is not conducted using methods to minimize bias (in contrast to a systematic review).
Natural History	As distinct from prognosis, natural history refers to the possible consequences and outcomes of a disease or condition and the frequency with which they can be expected to occur when the disease condition is untreated.
Negative Predictive Value (NPV)	See Predictive Value.
Negative Study (or Negative Trial)	Studies in which the authors have concluded that the comparison groups do not differ statistically in the variables of interest. The research results fail to support the researchers' hypotheses.
Network Meta-analysis (or Multiple Treatment Comparison Meta-analysis)	This systematic review allows the comparison of multiple interventions, including head-to-head evaluations at the same time as indirect comparisons, in a connected network of comparisons.
Neural Network	The application of nonlinear statistics to pattern-recognition problems. Neural networks can be used to develop clinical prediction rules. The technique identifies those predictors most strongly associated with the outcome of interest that belong in a clinical prediction rule and those that can be omitted from the rule without loss of predictive power.
N-of-1 Randomized Clinical Trial (or N-of-1 RCT)	An experiment designed to determine the effect of an intervention or exposure on a single study participant. In one n-of-1 design, the patient undergoes pairs of treatment periods organized so that 1 period involves the use of the experimental treatment and 1 period involves the use of an alternate treatment or placebo. The patient

*(Continued)*

Term	Definition
N-of-1 Randomized Clinical Trial (or N-of-1 RCT) ( <i>Continued</i> )	and clinician are blinded if possible, and outcomes are monitored. Treatment periods are replicated until the clinician and patient are convinced that the treatments are definitely different or definitely not different.
Nomogram	A graphic scale facilitating calculation of a probability. The most used nomogram in evidence-based medicine is one developed by Fagan to move from a pretest probability, through a likelihood ratio, to a posttest probability.
Nonadherent	Patients are nonadherent if they are not exposed to the full course of a study intervention (eg, most commonly, they do not take the prescribed dose or duration of a drug or they do not participate fully in the study program).
Noninferiority Trial	Noninferiority trials address whether the effect of an experimental intervention is not worse than a standard intervention by more than a specified margin. This contrasts with equivalence trials, which aim to determine whether an intervention is similar to another intervention. Noninferiority of the experimental intervention with respect to the standard treatment may be of interest if the new intervention has some other advantage, such as greater availability, reduced cost, less invasiveness, fewer harms, or decreased burden—or a potential for increased income for the sponsor.
Nonparticipant Observation	See Field Observation.
Null Hypothesis	In the hypothesis-testing framework, this is the starting hypothesis that the statistical test is designed to consider and possibly reject, which contends that there is no association among the variables under study.
Null Result	A nonsignificant result; no statistically significant difference between groups.

Term	Definition
Number Needed to Harm (NNH)	The number of patients who, if they received the experimental intervention, would lead to 1 additional patient being harmed during a specific period. It is the inverse of the absolute risk increase (ARI), expressed as a percentage ( $100/\text{ARI}$ ).
Number Needed to Screen (NNS)	The number of patients who would need to be screened to prevent 1 adverse event.
Number Needed to Treat (NNT)	The number of patients who need to be treated during a specific period to achieve 1 additional good outcome. When NNT is discussed, it is important to specify the intervention, its duration, and the desirable outcome. If an NNT calculation results in a decimal, round up as per Cochrane guidance ( <a href="http://www.cochrane-net.org/openlearning/html/mod11-6.htm">http://www.cochrane-net.org/openlearning/html/mod11-6.htm</a> ). It is the inverse of the absolute risk reduction (ARR), expressed as a percentage ( $100/\text{ARR}$ ).
Number of People Needed to Invite to Screening (NNI)	The number of people who need to be invited to screen to prevent 1 adverse event (calculated from the absolute risk difference in intention-to-treat analyses of randomized trials of screening). The NNI is larger than the number needed to screen because it is dependent on the uptake of screening; however, it may underestimate the effect of screening among individuals who participate fully in a program.
Observational Study (or Observational Study Design)	An observational study can be used to describe many designs that are not randomized trials (eg, cohort studies or case-control studies that have a goal of establishing causation, studies of prognosis, studies of diagnostic tests, and qualitative studies). The term is most often used in the context of cohort studies and case-control studies in which patient or caregiver preference, or happenstance, determines whether a person is exposed to an intervention or putative harmful agent or behavior (in contrast to the exposure being under the control of the investigator, as in a randomized trial).

Term	Definition
Observer Bias	Occurs when an observer's observations differ systematically according to participant characteristics (eg, making systematically different observations in treatment and control groups). See also Bias.
Odds	The ratio of events to nonevents; the ratio of the number of study participants experiencing the outcome of interest to the number of study participants not experiencing the outcome of interest.
Odds Ratio (OR) (or Relative Odds)	A ratio of the odds of an event in an exposed group to the odds of the same event in a group that is not exposed.
Odds Reduction	The odds reduction expresses, for odds, what relative risk reduction expresses for risks. Just as the relative risk reduction is $1 - \text{relative risk}$ , the odds reduction is $1 - \text{relative odds}$ (the relative odds and odds ratio being synonymous). Thus, if a treatment results in an odds ratio of 0.6 for a particular outcome, the treatment reduces the odds for that outcome by 0.4.
One-Group Pretest-Posttest Design (or Before-After Design)	See Before-After Design.
Open-Ended Interviews/ Questions	Questions that offer no specific structure for the respondents' answers and allow the respondents to answer in their own words. In qualitative research, this is sometimes also referred to as "unstructured" interviews. Interviewers invite participants to narrate their stories or perspectives on a very general topic in their own terms, with as little prompting or steering from the interviewer as possible. Open-ended questions are used.
Opinion Leaders (or Local Opinion Leaders)	See Local Opinion Leaders.

Term	Definition
Opportunity Costs	The value of (health or other) benefits forgone in alternative uses when a resource is used.
Optimal Information Size (OIS)	When using a GRADE (Grading of Recommendations Assessment, Development and Evaluation) approach to interpreting precision, examining the 95% confidence intervals (CIs) provides the optimal primary approach. We are skeptical about early studies with large effects and apparently satisfactory CIs. The optimal information size (OIS) is a way of dealing with such situations. The OIS is the number of patients required for an adequately powered individual trial assuming a modest treatment effect. If the CIs appear satisfactory but the sample size is less than the OIS, we lose confidence in estimates because of imprecision.
Outcome Variable (or Dependent Variable or Target Variable)	The target variable of interest. The variable that is hypothesized to depend on or be caused by another variable (the independent variable).
Overdetection	The detection of inconsequential disease—that is, disease that meets pathologic criteria for disease but that would not cause symptoms or become life-threatening if left undetected and untreated.
Partial Verification Bias	Occurs when only a selected sample of patients who underwent the index test is verified by the reference standard, and that sample is dependent on the results of the test. For example, patients with suspected coronary artery disease whose exercise test results are positive may be more likely to undergo coronary angiography (the reference standard) than those whose exercise test results are negative. See also Bias.
Participant Observation	See Field Observation.

Term	Definition
Patient-Important Outcomes	Outcomes that patients value directly. This is in contrast to surrogate, substitute, or physiologic outcomes that clinicians may consider important. One way of thinking about a patient-important outcome is that, were it to be the only thing that changed, patients would be willing to undergo an intervention with associated risk, cost, or inconvenience. This would be true of treatments that ameliorated symptoms or prevented morbidity or mortality. It would not be true of treatments that lowered blood pressure, improved cardiac output, improved bone density, or the like, without improving the quality or increasing the length of life.
Patient-Mediated Interventions	A strategy for changing clinician behavior. Any intervention aimed at changing the performance of health care professionals through interactions with, or information provided by or to, patients.
Patient Preferences	The relative value that patients place on various health states. Preferences are determined by values, beliefs, and attitudes that patients bring to bear in considering what they will gain—or lose—as a result of a management decision. Explicit enumeration and balancing of benefits and risks that are central to evidence-based clinical practice bring the underlying value judgments involved in making management decisions into bold relief.
Patient-Reported Outcomes	Any report of the status of a patient's health condition that comes directly from the patient without interpretation of the patient's response by a clinician or anyone else. Patient-reported outcomes can be measured in absolute terms (eg, severity of a sign, symptom, or state of a disease) or as a change from a previous measure.
Pedigree	A diagram that depicts heritable traits across 2 or more generations of a family.

Term	Definition
Pearson Correlation Coefficient	A statistical test of correlation between 2 groups of normally distributed data. The Pearson correlation provides a measure of association rather than measure of agreement. See also Correlation Coefficient.
Per-Protocol Analysis (Efficacy Analysis or Effectiveness Analysis)	Includes the subset of patients who complete the entire clinical trial according to the protocol. This approach compromises the prognostic balance that randomization achieves and is therefore likely to provide a biased estimate of treatment effect.
Pharmacogenomics	The analysis of how genetic makeup affects an individual's response to drugs. Pharmacogenomics deals with the influence of genetic variation on drug response in patients by correlating gene expression or single-nucleotide polymorphisms with a drug's efficacy or toxicity. The goal is to optimize drug therapy according to a patient's genotype to ensure maximum efficacy with minimal adverse effects.
Phase 1 Studies	Studies, often conducted in healthy volunteers, that investigate a drug's physiologic effect and evaluate whether it manifests unacceptable early toxic effects.
Phase 2 Studies	Initial studies on patients that provide preliminary evidence of possible drug effectiveness.
Phase 3 Studies	Randomized clinical trials designed to test the magnitude of benefit and harm of a drug.
Phase 4 Studies (or Postmarketing Surveillance Studies)	Studies conducted after the effectiveness of a drug has been established and the drug marketed, typically to establish the frequency of uncommon or unanticipated toxic effects.
Phenomenology	In qualitative research, an approach to inquiry that emphasizes the complexity of human experience and the need to understand the experience holistically as it is actually lived.



Term	Definition
Phenotype	The observable characteristics of a cell or organism, usually being the result of the product coded by a gene (genotype).
$\varphi$ (or $\varphi$ Statistic)	A measure of chance-independent agreement.
PICO (Patient, Intervention, Comparison, Outcome)	A method for answering clinical questions.
Placebo	A biologically inert substance (typically a pill or capsule) that is as similar as possible to the active intervention. Placebos are sometimes given to participants in the control arm of a drug trial to help ensure that the study is blinded.
Placebo Effect	The effect of an intervention independent of its biologic effect.
Point Estimate	The single value that best represents the value of the population parameter.
Polymorphism	The existence of 2 or more variants of a gene, occurring in a population, with at least 1% frequency of the less common variant. See also Mutation.
Pooled Estimate	A statistical summary measure representing the best estimate of a parameter that applies to all of the studies that contribute to addressing a similar question (such as a pooled relative risk and 95% confidence intervals from a set of randomized trials).
Positive Predictive Value (PPV)	See Predictive Value.
Positive Study (or Positive Trial)	A study with results that reveal a difference that investigators interpret as beyond the play of chance.
Posttest Odds	The odds of the target condition being present after the results of a diagnostic test are available.

<b>Term</b>	<b>Definition</b>
Posttest Probability	The probability of the target condition being present after the results of a diagnostic test are available.
Power	The ability of a study to reject a null hypothesis when it is false (and should be rejected). Power is linked to the adequacy of the sample size: if a sample size is too small, the study will have insufficient power to detect differences between groups.
Practice Guidelines (or Clinical Practice Guidelines or Guidelines)	See Clinical Practice Guidelines.
Prediction Rules (or Clinical Prediction Rules)	See Clinical Prediction Rules.
Predictive Value	There are 2 categories of predictive value. Positive predictive value is the proportion of people with a positive test result who have the disease; negative predictive value is the proportion of people with a negative test result and who are free of disease.
Preferences	See Values and Preferences.
Pretest Odds	The odds of the target condition being present before the results of a diagnostic test are available.
Pretest Probability	The probability of the target condition being present before the results of a diagnostic test are available.
Prevalence	Proportion of persons affected with a particular disease at a specified time. Prevalence rates obtained from high-quality studies can inform pretest probabilities.
Prevent (Prevention)	A preventive maneuver is an action that decreases the risk of a future event or the threatened onset of disease. Primary prevention

*(Continued)*

Term	Definition
Prevent (Prevention) (Continued)	is designed to stop a condition from developing. Secondary prevention is designed to stop or slow progression of a disease or disorder when patients have a disease and are at risk for developing something related to their current disease. Often, secondary prevention is indistinguishable from treatment. An example of primary prevention is vaccination for pertussis. An example of secondary prevention is administration of an antiosteoporosis intervention to women with low bone density and evidence of a vertebral fracture to prevent subsequent fractures. An example of tertiary prevention is a rehabilitation program for patients experiencing the adverse effects associated with a myocardial infarction.
Primary Studies	Studies that collect original data. Primary studies are differentiated from synopses that summarize the results of individual primary studies, and they are different from systematic reviews that summarize the results of a number of primary studies.
Principal Components Analysis	A series of microarray experiments that produces observations of differential expression for thousands of genes across multiple conditions. Principal components analysis is a statistical technique for determining the key variables in a multidimensional data set that explain the differences in the observations and can be used to simplify the analysis and visualization of multidimensional data sets.
Probabilistic Sensitivity Analysis	Related to economic analysis, this is an approach for dealing with uncertainty in economic models whereby distributions are defined for model variables and simulation techniques used to make random draws of the distributions to estimate the variability in estimated costs and outcomes.

Term	Definition
Probability	Quantitative estimate of the likelihood of a condition existing (as in diagnosis) or of subsequent events (such as in an intervention study).
Prognosis	The possible consequences and outcomes of a disease and the frequency with which they can be expected to occur.
Prognostic Factors	Patient or participant characteristics that confer increased or decreased risk of a positive or adverse outcome.
Prognostic Study	A study that enrolls patients at a point in time and follows them forward to determine the frequency and timing of subsequent events.
Prospective Study (or Cohort Study or Longitudinal Study)	See Cohort Study.
Publication Bias	Occurs when the publication of research depends on the direction of the study results and whether they are statistically significant. See also Bias.
Purposive Sampling (or Purposeful Sampling or Judgmental Sampling)	In qualitative research, a type of nonprobability sampling to select participants based on key characteristics relevant to the research question and on analytic questions as they arise during analysis. Specific sampling criteria may evolve during a project. Depending on the topic, examples include maximum variation sampling to document range or diversity; extreme case sampling, in which one selects cases that are opposite in some way; typical or representative case sampling to describe what is common in terms of the phenomenon of interest; critical sampling to make a point dramatically; and criterion sampling, in which all cases that meet some predetermined criteria of importance are studied.

Term	Definition
<i>P</i> Value (or <i>P</i> )	The probability that results as extreme as or more extreme than those observed would occur if the null hypothesis were true and the experiment were repeated over and over. $P < .05$ means that there is a less than 1 in 20 probability that, on repeated performance of the experiment, results as extreme as or more extreme than those observed would occur if the null hypothesis were true.
Pyramid of EBM Resources	This term refers to the way evidence-based medicine resources can be viewed in 3 broad categories: summaries and guidelines, preappraised research, and nonpreappraised research.
Qualitative Research	Qualitative research focuses on social and interpreted, rather than quantifiable, phenomena and aims to discover, interpret, and describe rather than to test and evaluate. Qualitative research makes inductive, descriptive inferences to theory concerning social experiences or settings, whereas quantitative research makes causal or correlational inferences to populations. Qualitative research is not a single method but a family of analytic approaches that rely on the description and interpretation of qualitative data. Specific methods include, for example, grounded theory, ethnography, phenomenology, case study, critical theory, and historiography.
Quality-Adjusted Life-Year (QALY)	A unit of measure for survival that accounts for the effects of suboptimal health status and the resulting limitations in quality of life. For example, if a patient lives for 10 years and his or her quality of life is decreased by 50% because of chronic lung disease, survival would be equivalent to 5 quality-adjusted life-years (QALYs).
Quality Improvement	An approach to defining, measuring, improving, and controlling practices to maintain or improve the appropriateness of health care services.
Quality of Care	The extent to which health care meets technical and humanistic standards of optimal care.

Term	Definition
Quantitative Research	The investigation of phenomena that lend themselves to test well-specified hypotheses through precise measurement and quantification of predetermined variables that yield numbers suitable for statistical analysis.
Random	Governed by a formal chance process in which the occurrence of previous events is of no value in predicting future events. For example, the probability of assigning a participant to 1 of 2 specified groups is 50%.
Random Allocation (or Randomization)	See Randomization.
Random-Effects Model	A model used to give a summary estimate of the magnitude of effect in a meta-analysis that assumes that the studies included are a random sample of a population of studies that address the question posed in the meta-analysis. Each study estimates a different underlying true effect, and the distribution of these effects is assumed to be normal around a mean value. Because a random-effects model takes into account both within-study and between-study variability, the confidence interval around the point estimate is, when there is appreciable variability in results across studies, wider than it could be if a fixed-effects model were used.
Random Error (or Chance)	We can never know with certainty the true value of an intervention effect because of random error. It is inherent in all measurement. The observations that are made in a study are only a sample of all possible observations that could be made from the population of relevant patients. Thus, the average or mean value of any sample of observations is subject to some variation from the true value for that entire population. When the level of random error associated with a measurement is high, the measurement is less precise, and we are less certain about the value of that measurement.

Term	Definition
Randomization (or Random Allocation)	The allocation of participants to groups by chance, usually done with the aid of a table of random numbers. Not to be confused with systematic allocation or quasi-randomization (eg, on even and odd days of the month) or other allocation methods used at the discretion of the investigator.
Randomized Clinical Trial (RCT) or Randomized Trial	An experiment in which individuals are randomly allocated to receive or not receive an experimental diagnostic, preventive, therapeutic, or palliative procedure and then followed up to determine the effect of the intervention.
Random Sample	A sample derived by selecting sampling units (eg, individual patients) such that each unit has an independent and fixed (generally equal) chance of selection. Whether a given unit is selected is determined by chance (eg, by a table of randomly ordered numbers).
Recall Bias	Occurs when patients who experience an adverse outcome have a different likelihood of recalling an exposure than patients who do not experience the adverse outcome, independent of the true extent of exposure. See also Bias.
Receiver Operating Characteristic (ROC) Curve	A figure depicting the power of a diagnostic test. The receiver operating characteristic (ROC) curve presents the test's true-positive rate (ie, sensitivity) on the horizontal axis and the false-positive rate (ie, $1 - \text{specificity}$ ) on the vertical axis for different cut points dividing a positive from a negative test result. An ROC curve for a perfect test has an area under the curve of 1.0, whereas a test that performs no better than chance has an area under the curve of only 0.5.
Recessive	Describes any trait that is expressed in a homozygote but not a heterozygote (ie, 2 copies of that allele are necessary to manifest its effect).

Term	Definition
Recursive Partitioning Analysis	A technique for determining the optimal way of using a set of predictor variables to estimate the likelihood of an individual's experiencing a particular outcome. The technique repeatedly divides the population (eg, old vs young, among young and old) according to status on variables that discriminate between those who will have the outcome of interest and those who will not.
Reference Standard (or Criterion Standard or Gold Standard)	See Criterion Standard.
Referral Bias	Occurs when characteristics of patients differ between one setting (such as primary care) and another setting that includes only referred patients (such as secondary or tertiary care). See also Bias.
Reflexivity	In qualitative research using field observation, whichever of the 3 approaches used, the observer will always have some effect on what is being observed, small or large. This interaction of the observer with what is observed is called reflexivity. Whether it plays a positive or negative role in accessing social truths, the researcher must acknowledge and investigate reflexivity and account for it in data interpretation.
Regression (or Regression Analysis)	A technique that uses predictor or independent variables to build a statistical model that predicts an individual patient's status with respect to a dependent variable or target variable.
Relative Diagnostic Odds Ratio	The diagnostic odds ratio is a single value that provides one way of representing the power of the diagnostic test. It is applicable when we have a single cut point for a test and classify test results as positive and negative. The diagnostic odds ratio is calculated as the product of the true-positive and true-negative results

(Continued)



Term	Definition
Relative Diagnostic Odds Ratio ( <i>Continued</i> )	divided by the product of the false-positive and false-negative results. The relative diagnostic odds ratio is the ratio of one diagnostic odds ratio to another.
Relative Odds	See Odds Ratio. Just as relative risk and risk ratio are synonymous, relative odds and odds ratio are synonymous.
Relative Risk (RR) (or Risk Ratio)	The ratio of the risk of an event among an exposed population to the risk among the unexposed.
Relative Risk Increase (RRI)	The proportional increase in risk of harmful outcomes between experimental and control participants. It is calculated by dividing the risk of a harmful outcome in the experimental group (experimental group risk [EGR]) minus the risk of a harmful outcome in the control group (control group risk [CGR]) by the risk of a harmful outcome in the control group $([EGR - CGR]/CGR)$ . Typically used with a harmful exposure.
Relative Risk Reduction (RRR)	The proportional reduction in risk of harmful outcomes between experimental and control participants. It is calculated by dividing the risk of a harmful outcome in the control group (control group risk [CGR]) minus the risk of a harmful outcome in the experimental group (experimental group risk [EGR]) by the risk of a harmful outcome in the control group $([CGR - EGR]/CGR)$ . Used with a beneficial exposure or intervention. See also Relative Risk; Risk; Treatment Effect.
Reliability	A technical statistical term that refers to a measurement instrument's ability to differentiate among subjects, patients, or participants in some underlying trait. Reliability increases as the variability between subjects increases and decreases as the variability within subjects (over time or over raters) increases. Reliability is typically expressed as an intraclass correlation coefficient with between-subject variability in the numerator and total variability (between-subject and within-subject) in the denominator.

Term	Definition
Reminding (or Reminders or Reminder Systems)	A strategy for changing clinician behavior. Manual or computerized reminders to prompt behavior change.
Reporting Bias (or Selective Outcome Reporting Bias)	The inclination of authors to differentially report research results according to the magnitude, direction, or statistical significance of the results. See also Bias.
Residual Confounding	Unknown, unmeasured, or suboptimally measured prognostic factors that remain unbalanced between groups after full covariable adjustment by statistical techniques. The remaining imbalance will lead to a biased assessment of the effect of any putatively causal exposure.
Responsiveness	The sensitivity or ability of an instrument to detect change over time.
Review	A general term for an article that systematically evaluates and summarizes the results of more than 1 primary study, as in a systematic review, or an article that summarizes a topic without an evidence-based approach, as in a narrative review. See also Systematic Review and Narrative Review.
Ribosome	The protein synthesis machinery of a cell where messenger RNA translation occurs.
Risk	A measure of the association between exposure and outcome (including incidence, adverse effects, or toxicity).
Risk Difference	The absolute difference in risk of a harmful outcome between experimental and control participants. It is calculated by subtracting the risk of a harmful outcome in the control group (control group risk [CGR]) minus the risk of a harmful outcome in the experimental group (experimental group risk [EGR]) (CGR – EGR).

Term	Definition
Risk Factors	Risk factors are patient characteristics associated with the development of a disease in the first place. Prognostic factors are patient characteristics that confer increased or decreased risk of a positive or adverse outcome from a given disease.
Risk of Bias	The extent to which study results are subject to systematic error.
Risk Ratio (or Relative Risk)	See Relative Risk.
Screening	Services designed to detect people at high risk of experiencing a condition associated with a modifiable adverse outcome, offered to persons who have neither symptoms of nor risk factors for a target condition.
Secondary Evidence-Based Journal	A secondary journal does not publish original research but rather includes synopses of published research studies that meet prespecified criteria of both clinical relevance and methodologic quality.
Secular Trends	Changes in the probability of events with time, independent of known predictors of outcome.
Semistructured Interview	In qualitative research, interviews that are structured in the sense of covering a specific list of issues relevant to the analysis but unstructured in the sense that both the way questions are asked and the way they are answered will vary from one interview to the next. Interviewers systematically touch on specific topics but pose questions in natural, conversational language and invite open-ended answers from participants.
Sensitivity	The proportion of people with a positive test result among those with the target condition. See also Specificity.
Sensitivity Analysis	Any test of the stability of the conclusions of a health care evaluation over a range of probability estimates, value judgments, and

*(Continued)*

Term	Definition
Sensitivity Analysis ( <i>Continued</i> )	assumptions about the structure of the decisions to be made. This may involve the repeated evaluation of a decision model in which one or more of the parameters of interest are varied.
Sentinel Effect	The tendency for human performance to improve when participants are aware that their behavior is being evaluated, in contrast to the Hawthorne effect, which refers to behavior change as a result of being observed but not evaluated.
Sequential Sample (or Consecutive Sample)	See Consecutive Sample.
Sign	Any abnormality indicative of disease, discoverable by the clinician at an examination of the patient. It is an objective aspect of a disease.
Signal-to-Noise Ratio	Signal refers to the target of the measurement; noise, to random error that obscures the signal. When one is trying to discriminate among people at a single point in time (who is better off, who is worse off), the signal comes from differences in scores among patients. The noise comes from variability or differences in score within patients over time. The greater the noise, the more difficult it is to detect the signal. When one is trying to evaluate change over time, the signal comes from the difference in scores in patients whose status has improved or deteriorated. The noise comes from the variability in scores in patients whose status has not changed.
Sign Test	A nonparametric test for comparing 2 paired groups according to the relative ranking of values between the pairs.
Silo Effect	One of the main reasons for considering narrower viewpoints in conducting an economic analysis is to assess the effect of change on the main budget holders because budgets may need to be adjusted before a new intervention can be adopted (the silo effect).

Term	Definition
Single-Nucleotide Polymorphism (SNP)	A single base-pair change in the DNA sequence at a particular point compared with the common or wild-type sequence.
Social Desirability Bias	Occurs when participants answer according to social norms or socially desirable behavior rather than what is actually the case (eg, under-reporting alcohol consumption). See also Bias.
Specificity	The proportion of people who are truly free of a designated disorder who are so identified by the test. The test may consist of, or include, clinical observations. See also Sensitivity.
Spectrum Bias	Ideally, diagnostic test properties will be assessed in a population in which the spectrum of disease in the target-positive patients includes all of those in whom clinicians might be uncertain about the diagnosis, and the target-negative patients include all of those with conditions easily confused with the target condition. Spectrum bias may occur when the accuracy of a diagnostic test is assessed in a population that differs from this ideal. Examples of spectrum bias would include a situation in which a substantial proportion of the target-positive population have advanced disease and target-negative participants are healthy or asymptomatic. Such situations typically occur in diagnostic case-control studies (for instance, comparing those with advanced disease to healthy individuals). Such studies are liable to yield an overly sanguine estimate of the usefulness of the test. See also Bias.
Stakeholder Analysis	A strategy that seeks to increase understanding of stakeholder behavior, plans, relationships, and interests and to generate information about stakeholders' levels of influence, support, and resources.
Standard Error	The standard deviation of an estimate of a population parameter. The standard error of the mean is the standard deviation of the estimate of the population mean value.

Term	Definition
Standard Gamble	A direct preference or utility measure that effectively asks respondents to rate their quality of life on a scale from 0 to 1.0, where 0 is death and 1.0 is full health. Respondents choose between a specified time $x$ in their current health state and a gamble in which they have probability $P$ (anywhere from 0 to .99) of full health for time $x$ and a probability $1 - P$ of immediate death.
Standardized Mean Difference (SMD)	A statistic used in meta-analysis when the studies all assess the same outcome but measure that outcome using different measurement instruments (eg, different instruments to measure anxiety or pain). Reported as $d$ . See also Effect Size.
Statistical Process Control	A statistical method used for quality improvement based on understanding expected variation in process or outcomes. It involves measuring, plotting, and analyzing data over time to detect stable, improving, or declining performance, the last of which prompts controlling or corrective action.
Statistical Significance	A term that indicates that the results obtained in an analysis of study data are unlikely to have occurred by chance and the null hypothesis is rejected. When statistically significant, the probability of the observed results, given the null hypothesis, falls below a specified level of probability (most often $P < .05$ ). One-sided significance testing is conducted when only effects in one direction are considered. Note: $P$ values do not provide an estimate of the magnitude of an effect or the precision of the estimate of magnitude. The results of specific statistical tests and measures of variance (eg, odds ratios and 95% confidence intervals, medians and interquartile ranges, means and standard deviations) should be provided.

Term	Definition
Stepped Wedge Design	<p>The sequential rollout of a quality improvement (QI) intervention to study units (clinicians, organizations) during a number of periods so that by the end of the study all participants have received the intervention. The order in which participants receive the intervention may be randomized (similar rigor to cluster randomized designs). Data are collected and outcomes measured at each point at which a new group of participants (“step”) receives the QI intervention. Observed differences in outcomes between the control section of the wedge with those in the intervention section are attributed to the intervention.</p>
Stopped Early Trials (or Truncated Trials)	<p>Truncated randomized clinical trials (RCTs) are trials stopped early due to apparent harm because the investigators have concluded that they will not be able to demonstrate a treatment effect (futility) or because of apparent benefit. Believing the treatment from RCTs stopped early for benefit will be misleading if the decision to stop the trial resulted from catching the apparent benefit of treatment at a random high.</p>
Stopping Rules	<p>These are methodologic and statistical guides that inform decisions to stop trials early. They can incorporate issues such as the planned sample size, planned and conducted interim analyses, presence and type of data monitoring including independent research oversight, statistical boundaries, and statistical adjustments for interim analyses and stopping.</p>
Structured Abstract	<p>A brief summary of the key elements of an article following prespecified headings. For example, the <i>ACP Journal Club</i> therapy abstracts include major headings of question, methods, setting, patients, intervention, main results, and conclusion. More highly structured abstracts include subheadings. For example, <i>ACP Journal Club</i> therapy abstracts methods sections include design, allocation, blinding, and follow-up period.</p>

Term	Definition
Subgroup Analysis	The separate analysis of data for subgroups of patients, such as those at different stages of their illness, those with different comorbid conditions, or those of different ages.
Substitute Outcomes or End Points (or Surrogate Outcomes or End Points)	See Surrogate End Points.
Summary-of-Findings Table	In a practice guideline developed according to the GRADE (Grading of Recommendations Assessment, Development and Evaluation) method, the summary-of-findings table provides the confidence ratings for all important outcomes and the associated estimates of relative and absolute effects. Summary-of-findings tables can facilitate shared decision making.
Superiority Trial	Superiority trials are designed to determine whether an experimental intervention is better than a control (typically a standard intervention or existing standard of care). Interpreting the results of superiority trials requires an a priori definition of the smallest difference in outcomes between the interventions that patients would consider large enough in favor of the experimental intervention to justify a preference for it given possible harms, burden, or cost.
Surrogate Outcomes or End Points (or Substitute Outcomes or End Points)	Outcomes that are not in themselves important to patients but are associated with outcomes that are important to patients (eg, bone density for fracture, cholesterol for myocardial infarction, and blood pressure for stroke). These outcomes would not influence patient behavior if they were the only outcomes that would change with an intervention.
Surveillance Bias	See Detection Bias.



<b>Term</b>	<b>Definition</b>
Survey	An observational study that focuses on obtaining information about activities, beliefs, preferences, knowledge, or attitudes from respondents through interviewer-administered or self-administered methods.
Survival Analysis	A statistical procedure used to compare the proportion of patients in each group who experience an outcome or end point at various intervals throughout the study (eg, death).
Survival Curve (or Kaplan-Meier Curve)	A curve that starts at 100% of the study population and shows the percentage of the population still surviving (or free of disease or some other outcome) at successive times for as long as information is available. See also Kaplan-Meier Curve.
Symptom	Any phenomenon or departure from the normal in function, appearance, or sensation reported by the patient and suggestive or indicative of disease.
Syndrome	A collection of signs or symptoms or physiologic abnormalities.
Synonymous single-nucleotide polymorphism	A single-nucleotide polymorphism (SNP) that does not lead to a change in the amino acid sequence compared with the common or wild-type sequence; in a nonsynonymous SNP, there is a change in the amino acid sequence as a result of the SNP.
Synopsis	A brief summary that encapsulates the key methodologic details and results of a single study or systematic review.
Systematic Error (or Bias)	See Bias.
Systematic Review	The identification, selection, appraisal, and summary of primary studies that address a focused clinical question using methods to reduce the likelihood of bias.

Term	Definition
Systems	Systems include practice guidelines, clinical pathways, or evidence-based textbook summaries that integrate evidence-based information about specific clinical problems and provide regular updates to guide the care of individual patients.
Target Condition (or Target Disease)	In diagnostic test studies, the condition the investigators or clinicians are particularly interested in identifying (such as tuberculosis, lung cancer, or iron deficiency anemia).
Target Negative	In diagnostic test studies, patients who do not have the target condition.
Target Outcome (or Target End Points or Target Events)	In intervention studies, the condition the investigators or clinicians are particularly interested in identifying and in which it is anticipated the intervention will decrease (such as myocardial infarction, stroke, or death) or increase (such as ulcer healing).
Target Positive	In diagnostic test studies, patients who have the target condition.
Target Variable (or Dependent Variable or Outcome Variable)	See Dependent Variable.
Test Threshold	The probability below which the clinician decides a diagnosis warrants no further consideration.
Themes	A generic term for the elements of qualitative research findings. Researchers usually express themes in terms of labels and definitions for the phenomena they describe or interpret from patterns in their data.
Theoretical Saturation	In qualitative research, this is the point in the analysis at which themes are well organized into a coherent theory or conceptual framework; new data fit easily without requiring revision to the theory. This is considered an appropriate stopping point for data analysis, especially in grounded theory methods.

Term	Definition
Theory	Theory consists of concepts and their relationships.
Theory Triangulation	See Triangulation.
Threshold Number Needed to Treat (or Threshold Number Needed to Harm)	The maximum number needed to treat or number needed to harm accepted as justifying the benefits and harms of therapy. See also Number Needed to Treat and Number Needed to Harm.
Time Series Design (or Interrupted Time Series Design)	In this study design, data are collected at several points both before and after the intervention. Data collected before the intervention allow the underlying trend and cyclical (seasonal) effects to be estimated. Data collected after the intervention allow the intervention effect to be estimated while accounting for underlying secular trends. The intervention may be interrupted then reintroduced multiple times. The time series design monitors the occurrence of outcomes or end points during a number of cycles and determines whether the pattern changes coincident with the intervention.
Transferability	The extent to which knowledge based on research findings can reasonably be applied to situations that differ from the original research setting. This requires judgment and expertise and necessarily draws on information from other sources in addition to information provided by the research study.
Treatment Effect (or Intervention Effect)	The results of comparative clinical studies can be expressed using various intervention effect measures. Examples are absolute risk reduction (ARR), relative risk reduction (RRR), odds ratio (OR), number needed to treat (NNT), and effect size. The appropriateness of using these to express an intervention effect and whether probabilities, means, or medians are used to calculate them depend on the type of outcome

(Continued)

Term	Definition
Treatment Effect (or Intervention Effect) (Continued)	variable used to measure health outcomes. For example, ARR, RRR, and NNT are used for dichotomous variables, and effect sizes are normally used for continuous variables.
Treatment Target	The manifestation of illness (a symptom, sign, or physiologic abnormality) toward which a treatment is directed.
Treatment Threshold (or Therapeutic Threshold)	Probability above which a clinician would consider a diagnosis confirmed and would stop testing and initiate treatment.
Trial of Therapy	In a trial of therapy, the physician offers the patient an intervention, reviews the effect of the intervention on that patient at some subsequent time, and, depending on the effect, recommends either continuation or discontinuation of the intervention.
Triangulation	In qualitative research, triangulation is an analytic approach in which key findings are corroborated using multiple sources of information. There are different types of triangulation. Investigator triangulation requires more than 1 investigator to collect and analyze the raw data, such that the findings emerge through consensus among a team of investigators. Theory triangulation is a process whereby emergent findings are corroborated with existing social science theories. Note that any discrepancies would not necessarily indicate that the research is biased or in error but rather that the next stage of empirical analysis should interpret and account for the discrepancies.
Trim-and-Fill Method	When publication bias is suspected in a systematic review, investigators may attempt to estimate the true intervention effect by removing, or trimming, small positive-result studies that do not have a negative-result study counterpart and then calculating a supposed true effect from the resulting symmetric funnel plot.

(Continued)

Term	Definition
Trim-and-Fill Method (Continued)	The investigators then replace the positive-result studies they have removed and add hypothetical studies that mirror these positive-result studies to create a symmetric funnel plot that retains the new pooled effect estimate. This method allows the calculation of an adjusted confidence interval and an estimate of the number of missing trials.
True Negative	Those whom the test correctly identifies as not having the target disorder.
True Positive	Those whom the test correctly identifies as having the target disorder.
Truncated Trials (Stopped Early Trials)	See Stopped Early Trials.
Trustworthiness (or Credibility)	See Credibility.
<i>t</i> Test	A parametric statistical test that examines the difference between the means of 2 groups of values.
Type I Error	An error created by rejecting the null hypothesis when it is true (ie, investigators conclude that an association exists among variables when it does not). See also $\alpha$ Level and Type II Error.
Type II Error	An error created by accepting the null hypothesis when it is false (ie, investigators conclude that no association exists among variables when, in fact, an association does exist). See also $\beta$ Error and Type I Error.
Unblinded (or Unmasked)	Patients, clinicians, those monitoring outcomes, judicial assessors of outcomes, data analysts, and manuscript authors are aware of whether patients have been assigned to the experimental or control group.
Unit of Allocation	The unit or focus used for assignment to comparison groups (eg, individuals or clusters such as schools, health care teams, hospital wards, outpatient practices).

Term	Definition
Unit of Analysis	The unit or focus of the analysis; although it is most often the individual study participant, in a study that uses cluster allocation, the unit of analysis is the cluster (eg, school, clinic).
Unit of Analysis Error	When investigators use any sort of cluster randomization (randomize by physician instead of patient, practice instead of physician or patient, or village instead of participant) and analyze as if they have randomized according to patient or participant, they have made a unit of analysis error. The appropriate analysis acknowledges the cluster randomization and takes into account the extent to which outcomes differ among clusters independent of treatment effect.
Univariate Regression (or Univariable Regression or Simple Regression)	This term is for simple descriptive analyses. It is often erroneously used for bivariable regression. See also Bivariable Regression.
Unmasked (or Unblinded)	See Unblinded.
Upfront Costs	Costs incurred to “produce” the treatment, such as the physician’s time, nurse’s time, and materials.
Utilitarian (or Consequentialist)	See Consequentialist.
Utility	Utility in the context of health economic modeling refers to the value of a health state, typically expressed from 0 (death) to 1.0 (full health).
Validity (or Credibility)	In terms of health status measurement, validity is the extent to which an instrument measures what it is intended to measure. In critical appraisal terms, validity reflects the extent to which the limitations in study design leave a study vulnerable to systematic error or spurious inferences. See also Credibility.

Term	Definition
Values and Preferences	When used generically, as in “values and preferences,” we refer to the collection of goals, expectations, predispositions, and beliefs that individuals have for certain decisions and their potential outcomes. The incorporation of patient values and preferences in decision making is central to evidence-based medicine. These terms also carry specific meaning in other settings. Measurement tools that require a choice under conditions of uncertainty to indirectly measure preference for an outcome in health economics (such as the standard gamble) quantify preferences. Measurement tools that evaluate the outcome on a scale with defined favorable and unfavorable ends (eg, visual analog scales, feeling thermometers) quantify values.
Variance	The technical term for the statistical estimate of the variability in results.
Variant Allele	The allele at a particular single-nucleotide polymorphism that is the least frequent in a population.
Verification Bias	See Differential Verification Bias.
Visual Analog Scale	A scaling procedure that consists of a straight line anchored on each end with words or phrases that represent the extremes of some phenomenon (eg, “worst pain I have ever had” to “absolutely no pain”). Respondents are asked to make a mark on the line at the point that corresponds to their experience of the phenomenon.
Washout Period	In a crossover or n-of-1 trial, the period required for the treatment to cease to act once it has been discontinued.
Weighted Mean Difference	The weighted mean difference is the difference between initial and final values of a continuous measure in a group of patients in a study. The weighted mean difference is also a way of presenting the magnitude of effect in a meta-analysis in which all studies have used the same

*(Continued)*

Term	Definition
Weighted Mean Difference (Continued)	continuous variable (such as exercise capacity or a specific quality-of-life instrument). It presents the best estimate of the difference between 2 treatments using the units of the particular outcome used in all of the studies. It is calculated as the sum of the differences in the individual studies, weighted by the individual variances for each study.
Wild-Type Allele	The allele at a particular single-nucleotide polymorphism that is most frequent in a population, also called a common allele.
Willingness to Pay	In some economic analyses, it may be desirable to compare costs and outcomes using the same metric (ie, costs). In this case, an attempt is made to ask people how much they would pay to achieve an improvement in health or to avoid a negative health event/outcome.
Working Diagnosis (or Leading Hypothesis)	The clinician's single best explanation for the patient's clinical problem(s).
Workup Bias	See Differential Verification Bias.



*This page intentionally left blank*

# INDEX

---

## A

---

- Absolute difference**
  - confidence intervals, 171
  - precision of results, 313–314
- Absolute effect estimates, meta-analyses, 300**
- Absolute risk (AR)**
  - incremental risk, 204–205
  - negative trials, 172
- Absolute risk (AR), dichotomous outcomes, 152**
- Absolute risk reduction (ARR)**
  - confidence intervals and, 173–177
  - evidence-based practice and, 161–162
  - incremental risk, 204–205
  - precision of results, 313–314
  - relative risk and, 154–155
  - systematic reviews and clinical application, 285–287
  - treatment effects, 152
  - treatment effect studies, 110–111
- Accessibility of EBM resources, assessment of, 57–59**
- ACCESSSS, evidence updates, 67**
- ACP Journal Club, 28**
  - preappraised research, 65–67
- Adjusted analysis, prognostic studies, 258**
- Adjusted indirect comparison, network meta-analyses, 343**
- Adverse outcome**
  - harm, observational studies, 181–205
  - network meta-analyses, 331–334, 350–351
  - risk analysis, 152
  - systematic reviews and clinical application, 285–287
- Allocation concealment**
  - network meta-analyses, 338
  - risk of bias, 303–304
- Alternative treatments**
  - patient management recommendations, 365–366
  - shared physician-patient decision making, 411
- American Heart Association, patient management recommendations, 370–371**
- AR. See Absolute risk (AR)**
- ARR. See Absolute risk reduction (ARR)**
- Availability of EBM resources, assessment of, 57–59**

## B

---

- Background questions**
  - current best evidence searches, 45–46
  - defined, 29
- Baseline risk**
  - defined, 152
  - relative risk reduction and, 154–155
  - systematic reviews and clinical application, 285–287
- Bayesian analysis**
  - network meta-analyses, 333–334, 337–338
  - posttest probabilities, diagnostic testing, 216–220

- Bayley scales, prognosis studies**, 253, 264–266
- Before–after studies**, 199
- Benefits analysis. See also Cost-benefit analysis**  
 in evidence-based medicine, 10–19  
 harm, observational studies, 205  
 negative trials, 172  
 network meta-analyses, 336  
 treatment effects, 119–123, 142–145
- Between-study differences**  
 network meta-analyses, 336–337  
 in systematic reviews, 284–285
- Bias**  
 confidence intervals and, 168–169  
 diagnostic testing, 226–233  
 harm, observational studies, and minimization of, 185–187  
 loss to follow-up and, 90, 106–108  
 misleading results, 88–90  
 pretest probabilities, 215  
 random error vs, 3  
 systematic error, 303–304  
 treatment effect, 101–103
- Blinded studies**  
 between-study differences, 337  
 clinical scenario, 166–167  
 diagnostic testing, 230  
 noninferiority trial validity, 135–140  
 prognostic balance and, 104–105  
 risk of bias reduction and, 93, 303–304  
 systematic reviews, 284
- BMJ EvidenceUpdates**, 58, 65, 67

- Boolean operator, search term selection and**, 77–78
- Broad search strategies**, 78–79
- Burden of treatment**  
 confidence intervals for estimation of, 176–177  
 decision tree example, 362–363  
 evidence-based medicine and, 10–19  
 multiple chronic conditions, 410–411  
 noninferiority trial, 129  
 systematic reviews and clinical application, 285–287

---

## C

- Case-control studies**  
 diagnostic testing, 228–229  
 exposures and outcomes in, 185–187  
 harm study as example of, 182–205, 192–197  
 questions matched to, 32
- Case reports, risk of bias in**, 198–199
- Case series, risk of bias in**, 198–199
- CGR. See Control group risk (CGR)**
- Chance-corrected agreement, systematic reviews**, 287
- Chance nodes, decision tree**, 360–362
- Channeling bias, in cohort studies**, 189
- $\chi^2$  distribution, test for heterogeneity**, 308–309
- CI. See Confidence intervals (CIs)**
- Class effects, therapy studies**, 116

**Clinical decision making**

- evidence and, 18–19
- harm, observational studies, 205
- meta-analyses credibility and, 278–281
- patient applicability of results, 314–316
- pooled studies, 272–273
- systematic reviews and, 285–287

**Clinical decision rules**

- diagnosis, 215
- prognostic studies, 258

**Clinical decision support systems, EBM evidence resources, 50–54****Clinical practice guidelines**

- clinical scenario, 359
- confidence rating of evidence, 301–322
- diagnostic test results applicability, 244–245
- evidence-based medicine and, 19–22
- online summaries, 62
- patient management recommendations, 359–360
- processing of evidence in, 49–50
- prognosis studies, 264–266
- reproducibility of diagnostic testing results and, 243–244
- strong recommendations, 371–376
- values and preferences and, 4, 18–19
- values and preferences of patient, 391

**Clinical prediction rules, diagnostic testing and, 246–247****Clinical Queries search**

- filter, nonpreappraised research, 68–72

**Clinical trials. See Randomized clinical trials (RCTs)****Clinician-as-perfect-agent approach, shared physician-patient decision making, 393****Closed loops, meta-analyses, 331, 340****Cluster analysis, diagnosis based on, 213****Cochrane Collaboration, weak recommendations, 376–377****Cochrane Controlled Trials Registry**

- nonpreappraised research, 71–72
- systematic review searches, 282–283

**Cochrane Decision Aid Registry, 411–412****Cochrane Library, preappraised research synopses, 67****Cochrane Q test, heterogeneity, 308–309****Cohort studies**

- diagnostic testing, 240–243
- follow-up in, 191–192
- harm study as example of, 182, 187–192
- outcome detection methods in, 190–192
- questions matched to, 32–33

**Cointerventions**

- prognostic balance and, 105
- risk of bias reduction and, 93

**Coin toss analogy, random error, 86–87****Comorbidity**

- number needed to treat and, 156–157
- prognosis and, 34

- Comorbidity, target outcomes and**, 101–103
- Competing risks, survival analysis**, 160–161
- Composite end points**
  - patient-important outcomes in therapy research, 119
  - questions involving, 42
  - reporting bias, 318–321
- Comprehensibility of evidence, patient management recommendations**, 371–377
- Concealed randomization, therapy research**, 99–100
- Concealment bias, systematic reviews**, 284
- Confidence intervals (CIs)**
  - best evidence summaries, 11–14
  - characteristics of, 169
  - in cohort studies, 189
  - meta-analyses, 166, 296–300
  - negative trials, 172
  - network meta-analyses, 333–334, 337–338
  - noninferiority trials, 131–134
  - patient decision aids, 404
  - pooled studies, 272–273
  - precision of results, 312–314
  - prognosis likelihood estimates, 262–264
  - properties of, 167–170
  - results analysis using, 170–171
  - risk estimation, 202
  - study size and, 172–177
  - systematic reviews, 287–288
  - therapy research, 97
  - treatment effect studies, 112–115, 158
  - visual assessment of results variability, 306–308
- Confidence ratings**
  - estimates, 301–322
  - GRADE confidence rating, 301–303
  - grades of recommendation, 370–371
  - increase in, 322
  - network meta-analyses, 337–338
  - precision of results, 313–314
  - treatment ranking, network meta-analyses, 343–344
- Conflicts of interest, decision analyses**, 377–378
- Confounding**
  - in case-control studies, 193–197
  - in cohort studies, 188–192
- Consecutive sampling, diagnostic testing**, 232–233
- Continuous variables**
  - meta-analyses, 294–300
  - systematic reviews and clinical application, 286–287
- Control event rate**
  - adverse outcomes, 152
  - noninferiority trial, 136–140
  - precision of results, 312–314
- Control group risk (CGR)**
  - adverse outcomes, 152
  - in cohort studies, 188–192
  - treatment effect studies, 110–111
- Control groups**
  - bias and, 89–90
  - harm, observational studies, 184
  - patient management recommendations, 365–366
  - prognosis research, 260–261
  - prognostic balance with patients in therapy trials, 103–104
  - questions matched to, 31–32
  - risk of bias in assessment of, 100–103

**Cost-benefit analysis**

- evidence-based medicine and, 10–19
- experimental treatment, 142–145
- systematic reviews and clinical application, 285–287
- treatment effects studies, 119–123

**Coverage in EBM resources, assessment of, 57****Credibility, 276–277**

- clinical issues, 277–281
- large between-study variability, 310–312
- network meta-analyses, 351–352
- results analysis, 293

**Credible intervals, network meta-analyses, 333–334, 337–338****Criterion standard**

- clinical questions
  - concerning, 30
- diagnostic testing, 229–230
- reporting bias, 318–321
- risk of bias reduction and, 90–93

**Cross-sectional studies, risk of bias in, 197****Current best evidence**

- EBM resources based on, 55–56
- patient management
  - recommendations, 367–368
- pyramid of EBM resources
  - and, 59–76
- searching resources for, 45–59

**D****DARE (Database of Abstracts of Reviews of Effects), preappraised research synopses, 65–67****Database searches**

- nonpreappraised research, 68–72
- preappraised research, 63–67

**Decision analyses, 4. See also Shared physician-patient decision making**

- clinician-as-perfect-agent approach, 393
- conflict of interest
  - minimization, 377–378
- patient management
  - recommendations, 359–362
- processing of evidence and, 49–50
- strong recommendations, 371–376, 378–383
- summaries and guidelines, 62
- weak recommendations, 376–377, 383–384

**Decision talk model, shared physician-patient decision making, 397–398****Decision tree**

- example, 362–363
- patient management
  - recommendations, 360–362

**Determinants of outcome. See Prognostic factors****Diagnosis. See also Diagnostic testing; Differential diagnosis**

- clinical scenario, 212
- clinician selection of
  - possibilities, 214
- cluster analysis, 213
- complementary approaches, 212–213
- posttest probability, 215–216
- pretest probability, 214–215
- process of, 211–220
- threshold/posttest probability
  - and clinical action, 216–220

**Diagnostic and Statistical  
Manual of Mental  
Disorders, 233**

**Diagnostic testing**

- clinical questions
  - concerning, 30
- clinical scenario, 225, 248–249
- continuous dichotomization, 239–243
- evidence sources, 226
- foreground questions
  - concerning, 31
- likelihood ratios, 233–239
- management strategies based on, 245–247
- matching questions to, 33–34
- patient care applications, 243–248
- reproducibility in clinical settings, 243–244
- results analysis, 233–243
- results applicability in clinical practice, 244–245
- risk of bias, 226–233

**Dichotomous outcomes**

- absolute risk reduction, 152
- analysis of, 150–162
- confidence intervals
  - and, 158
- diagnostic testing, 239–243
- meta-analyses, 294–300, 295–300
- number needed to harm, 158
- number needed to treat and, 156–157
- odds ratio, 153–154
- relative risk, 152–153
- relative risk reduction, 153
- risk analysis, 152
- survival analysis, 158–161
- treatment effect studies, 110–111
- 2 x 2 table analysis, 150–151

**Differential diagnosis**

- cluster analysis, 213
- foreground questions
  - concerning, 31
- questions matched to, 32–33, 38–39

**Differential verification**

- bias, diagnostic testing, 231**

**Difficult decisions, shared physician-patient decision making, 407–408**

**Direct comparisons, network meta-analyses, 341–343**

**Direct evidence, network meta-analyses, 331–334**

**Dissemination bias, results analysis, 317**

---

## E

**EBM. See Evidence-based medicine (EBM)**

**Economic analysis**

- decision analyses, 62
- patient-important outcomes in therapy research, 119
- patient management
  - recommendations, 362

**Effect estimates**

- credibility of, 276–277
- meta-analyses, 295–300
- strong recommendations, 373–376
- systematic reviews, 287–288, 295–300

**Effect size**

- GRADE confidence rating, 302–303
- network meta-analyses, 337

**EGR. See Experimental group risk (EGR)**

**Eligibility criteria**

- network meta-analyses, 334–335
- systematic reviews, 274–276, 278–281

**E-mail alert systems**

- availability and access of research and, 58–59
- for evidence updates, 26–28
- guidelines for using, 67

**EMBASE, systematic review searches, 282–283****End points, risk of bias reduction and, 93****Equivalence trials, noninferiority trial vs, 129–130****Estimates**

- confidence in, 14–18, 300–322
- GRADE confidence rating, 301–303
- precision, treatment effect studies, 112–115

**Event rate**

- loss to follow-up and, 107
- noninferiority trial, 136–140

**Evidence-based medicine (EBM)**

- clinical skills and, 19–22
- future challenges for, 21–22
- guiding principles, 2–7, 9–22
- humanism and, 19–22
- medical literature search strategies for, 27–28
- online resources for, 45
- overview of resources, 53–54
- pyramid of resources, 50–54
- resource organization and processing for, 47–54
- resource selection criteria, 55–59
- values and preferences of patient, 391

**Evidence-Based Medicine website, preappraised research, 65–67****Evidence-Based Oncology, preappraised research, 65–67****Evidence-based practice**

- clinical skills, 19–22
- guiding principles, 2–7
- measures of association and, 161–162
- network meta-analyses, 350–351
- quality of evidence, 300–322
- shared physician-patient decision making, 394–402

**Evidence processing and organization**

- diagnosis, 212
- diagnostic testing, 226
- in EBM resources, 47–54
- in harm, observational studies, 181–182
- meta-analyses, 271–272, 293
- pyramid of EBM evidence, 50–54
- risk of bias, 303–304
- systematic reviews, 271–272, 293
- therapy research, 97–98
- values and preferences of patient, 391

**Evidence profile, 323****EvidenceUpdates**

- e-mail alerts for, 26–28, 67
- online summaries, 60–62

**EvidenceUpdates e-mail alert system, 27–28****Exclusion criteria, therapy studies, 116****Experimental group risk (EGR), treatment effect studies, 110–111**



**Experimental intervention,  
systematic reviews, 283**

**Experimental treatment**  
harm and cost vs benefits of,  
142–145  
prognostic balance and, 93  
questions matched to, 31  
relative risk, 152–153

**Expertise, procedural  
interventions, 116–118**

**Exposure**  
benefits analysis and risk  
of, 205  
in case-control studies,  
192–197  
in cohort studies, 188–192  
in harm, observational studies,  
181–182, 201–202  
outcome association with,  
201–202  
patient exposure data,  
203–204  
risk of bias, 303–304  
risk of bias concerning,  
185–187  
systematic reviews, 272

## F

**Fagan's nomogram, diagnostic  
testing, 241–243**

**False-positive results, risk of,  
166–167**

**FDA. See Food and Drug  
Administration (FDA)**

**Federated search engines**  
EBM evidence resources,  
50–54  
pyramid of EBM resources  
searches, 72–74

**Filtering search strategies,  
nonpreappraised research,  
68–72**

## Follow-up

bias in losses to, 90  
in cohort studies,  
191–192  
completion of, 106–108  
diagnostic testing,  
229–230  
eligibility criteria in systematic  
reviews, 280–281  
harm, observational studies,  
184–187, 203  
noninferiority trial validity,  
136–140  
number needed to treat and,  
156–157  
prognostic risk of bias and,  
259, 266

## Food and Drug

**Administration (FDA)**  
noninferiority trial guidelines,  
131–134  
selective outcome reporting  
bias, 317–318  
systematic review searches,  
282–283

## Foreground questions

categories, 31  
current best evidence  
searches, 45–46  
defined, 29

## Forest plots

best evidence summaries,  
11–14  
confidence intervals and,  
173–177  
meta-analyses, 295–300  
network meta-analyses,  
346–349

**Framingham risk score,  
systematic reviews and  
clinical application,  
285–287**

**Framingham study, prognosis  
and patient similarity  
in, 258**

**Frequentist analysis, network meta-analyses**, 333–334  
**Funnel plots, reporting bias**, 319–321

## G

---

**Generalizable results**  
 network meta-analyses, 334–335  
 therapy studies, 115–116  
**Geometry of network, meta-analyses**, 331, 340  
**Glasgow Coma Scale, prognosis and**, 257–258  
**Global health resources, availability and access issues**, 58–59  
**Gold standard**  
 diagnostic testing, 33–34, 229–230  
 reporting bias, 318–321  
 risk of bias reduction and, 90–93  
 systematic review eligibility and, 279–281  
**Google Scholar**, 75–76  
**Google searches, guidelines for using**, 74–76  
**GRADE (Grading of Recommendations Assessment, Development and Evaluation) Working Group**, 6–7  
 confidence in estimates, 301–303  
 confidence rating increases, 322  
 estimation confidence, 293  
 evidence profile, 323  
 network meta-analyses, 338, 353  
 online summaries, 60–62

patient applicability of results, 314–316  
 patient management recommendations, 370–371  
 weak recommendations, 376–377

**Guidelines. See Clinical practice guidelines; Summaries and guidelines**

## H

---

**Harm studies**, 3  
 case-control studies, 192–197  
 case series/case reports, 198–199  
 clinical scenario, 181, 206–207  
 cohort studies, 187–192  
 cross-sectional studies, 197  
 evidence search, 181–182  
 experimental treatment, 142–145  
 foreground questions concerning, 31  
 network meta-analyses, 336  
 patient care applications, 202–205  
 results analysis, 201–202  
 risk of bias, 182–187, 197–200  
 treatment effects studies, 119–123  
**Hazard ratio**  
 meta-analyses, 294–300  
 survival analysis, 160–161  
 systematic reviews and clinical application, 286–287  
 treatment effect studies, 111  
**Health-related quality of life**  
 meta-analyses measurements, 294–300  
 systematic reviews and clinical application, 287

**Heterogeneity**

- between-study differences in systematic reviews, 285
- large between-study results variability, 310–312
- magnitude of, statistical tests, 309–310
- meta-analyses, 276
- results analysis, 305
- treatment effects, 330
- visual assessment of variability, 305–308
- yes-or-no statistical tests, 308–309

**Hierarchy of evidence, in primary studies, 47–49****Homogeneity, network meta-analyses, 332–334****Humanism, evidence-based medicine and, 19–22****Hypothesis testing**

- confidence intervals and, 171
- sample size, 166

**I** **$I^2$  statistic**

- magnitude of heterogeneity, 309–310
- network meta-analyses, 346–349

**Important decisions, shared physician-patient decision making, 406–407****Impossibility diagnosis, 214****Inclusion criteria**

- systematic reviews, 280–281
- therapy studies, 116

**Incoherence, test for, network meta-analyses, 341–343****Inconsistency, network meta-analyses, 338****Incremental risk, harm, observational studies, 204–205****Indirect comparisons, network meta-analyses, 341–343****Indirect evidence**

- GRADE confidence rating, 302–303
- network meta-analyses, 331–334, 338
- patient applicability of results, 314–316

**Individual patient data, meta-analyses, 283****Informed decision-making approach, 393–394****Intention-to-treat principle**

- meta-analyses, 283
- noninferiority trial, 138–140
- patient randomization and trial bias, 109–110
- risk of bias, 304

**International Statistical Classification of Diseases, 10th Revision (ICD-10) criteria, diagnostic testing, 233****Interviewer bias, in case-control studies, 195–197****J****JAMA, The Rational Clinical Examination reviews in, 238–239****K****K statistic, systematic reviews, 287**

**L**

- 
- Leading hypotheses, diagnosis and, 214**
  - Lifestyle factors, in case-control studies, 193–197**
  - Likelihood ratios/likelihood estimates**
    - diagnostic testing, 233–239
    - dichotomous continuous test scores, 239–243
    - management strategies based on diagnostic testing, 245–247
    - meta-analyses, 294–300
    - posttest probability, 216, 236
    - prognosis studies, 262–264
  - Loss to follow-up**
    - bias in, 90
    - network meta-analyses, 338
    - prognostic balance and, 106–108
    - prognostic risk of bias and, 259
    - risk of bias, 303–304
    - survival analysis, 161

**M**

- 
- Management strategies, diagnostic test results and, 245–247**
  - Markov models, 363**
  - McMaster *PLUS***
    - evidence updates on, 67
    - preappraised research, 63–67
  - Medical literature**
    - browsing strategies for, 27–28
    - evidence processing and organization in, 47–54
    - guidelines for using, 26–29
    - limitations of searches in, 46–47
    - new evidence in, 26–28
    - optimal patient care and, 4–7
    - problem solving and, 28–29
    - related article searches, 79–80
    - selection criteria for EBM resources, 55–59
  - Medical Subject Headings (MeSH) thesaurus, search term selection and, 78**
  - MEDLINE database**
    - nonpreappraised research, 68
    - systematic review searches, 282–283
  - MeSH. See Medical Subject Headings (MeSH) thesaurus**
  - Meta-analyses**
    - absolute effect estimates, 300
    - amount of evidence, 339–340
    - benefits, 274
    - best evidence summaries, 11–14
    - clinical scenario, 271, 288–289, 293
    - confidence intervals and, 172–177
    - consistency of results, cross-study analysis, 304–305
    - credibility of effect estimates in, 276–277
    - credibility of process, 277–281
    - definitions, 272–277
    - effect estimates in, 287–288
    - eligibility criteria in, 279–281
    - evidence sources, 271–272
    - harm, observational studies, 184–187
    - individual patient data, 283
    - large between-study variability, 310–312
    - network meta-analyses, 329–353
    - pooled estimates, 276
    - precision of results, 312–314

**Meta-analyses** (*continued*)

- process synopsis, 274–276
- quality of evidence, 300–322
- reporting bias, 316–321
- results interpretation, 293–323
- risk of bias, 283–284
- sample size, 166
- study relevance in, 281–283
- summary estimates, 293–300
- visual assessment of
  - variability, 305–308
- yes-or-no statistical test of heterogeneity, 308–309

**Meta-regression, between-study differences, 336–337****Minimal important difference, systematic reviews and clinical application, 287****Minimally disruptive medicine, multiple chronic conditions, 411****Misinformed participants, shared physician-patient decision making, 408–410****Misleading results**

- bias, 88–90
- random error, 86–88
- risk of bias reduction, 90–93
- shared physician-patient decision making, 408–410

**Multilevel likelihood ratios, diagnostic testing, 242–243****Multiple chronic conditions, shared physician-patient decision making for patients with, 410–411****Multistate transition models, 363**


---

## N

**Narrative reviews, 273–274****Narrow search strategies, 78–79****Negative trials**

- clinical scenario for, 167
- confidence intervals, 172
- exclusion of treatment effects in, 175–177
- reporting bias, 317

**Network meta-analyses**

- between-study differences, 336–337
- clinical scenario, 329, 352
- confidence ratings, 337–338
- direct/indirect comparisons, 341–343
- elements of, 330–334
- eligibility criteria, 334–335
- evidence sources, 329
- geometry of, 331
- patient care applications, 350–352
- results, 339–349
- risk of bias, 334–339
- selection and reporting bias, 335–336
- sensitivity and bias in results, 344–349
- similarity of results, 340–341
- treatment ranking, 343–344

**NNH. See Number needed to harm (NNH)****NNT. See Number needed to treat (NNT)****Nomograms, diagnostic testing, 236–239****Nonadherent patients**

- informed decision-making approach, 393–394
- noninferiority trial, 138–140
- randomized trials and, 109–110

**Noninferiority trial**

- benefit-cost analysis, 142–145
- clinical scenario example, 127–128, 145
- confidence intervals and, 177
- overview, 128–134

- patient care application of
    - results, 142–144
    - results analysis, 140–142
    - unwarranted conclusion of noninferiority, 136–140
    - validity of results, 134–140
  - Nonpreappraised research**, 54
    - guidelines for using, 67–72
  - “Not much worse” paradigm, noninferiority trials**, 128–130
  - Null hypothesis**
    - sample size, 166
    - test for heterogeneity, 308–309
  - Number needed to harm (NNH), treatment effects studies**, 158
  - Number needed to treat (NNT)**
    - incremental risk, 204–205
    - systematic reviews and clinical application, 287
    - treatment benefit-harm analysis, 119–123
    - treatment effects studies, 156–157
- 
- # O
- Objective outcome criteria, prognostic studies**, 260
  - Observational studies**
    - cross-sectional studies, 197
    - GRADE confidence rating, 302–303
    - harm studies as, 179–205
    - prognostic factors in, 101–103
    - questions matched to, 32–33
    - risk of bias, 283–284
    - therapy research, 99–100
  - Odds ratio (OR)**
    - best evidence summaries, 11–14
    - confidence intervals and, 173–177
    - exposure-outcome association, 201–202
    - meta-analyses, 294–300
    - network meta-analyses, 331–334
    - systematic reviews and clinical application, 286–287
    - therapy research, 97
    - treatment effects studies, 153–154
  - Online resources**
    - for evidence-based medicine, 45
    - preappraised research, 63–67
    - summaries and guidelines, 60–62
  - Optimal patient care, medical literature and**, 4–7
  - Option talk model, shared physician-patient decision making**, 396–397
  - OR. See Odds ratio (OR)**
  - Outcomes. See also Criterion standard**
    - clinical questions concerning, 30
    - in cohort studies, 188–192
    - detection methods in cohort studies, 190–192
    - diagnostic testing and, 247–248
    - exposure association with, 201–202
    - harm, observational studies, 182–187, 201–202
    - meta-analyses measurements, 294–300
    - network meta-analyses, 336
    - noninferiority trials, 130–134
    - over time, in prognostic studies, 261–262
    - patient decision aids, 404
    - patient management recommendations, 366–367

**Outcomes. See also Criterion****standard** (*continued*)

prognostic factors and,

188–192, 260

reporting bias concerning,

317–321

variables in, 359–360

**Ovid database, nonpreappraised research, 68****P****Pairwise comparisons, network meta-analyses, 337–338****Partial verification bias,****diagnostic testing, 231****Paternalistic approach, shared physician-patient decision making, 392****Patient care**

applicability of results, 314–316

diagnostic testing, 243–248

harm, observational studies, and, 202–205

network meta-analysis

applications, 350–353

noninferiority trial results and, 142–145

prognosis studies, 264–266

therapy study results and, 115–123

**Patient decision aids, 403–406**

sources for, 411–412

**Patient-important outcomes**

absolute effect estimates, 300

decision analysis, 366–367

diagnostic testing and, 247–248

network meta-analyses, 350–351

straightforward vs difficult decisions involving, 407–408

systematic reviews, 272

therapy research, 98–100, 118–119

**Patient management****recommendations**

alternative treatments, 365–366

assessment, 364–378

clinical scenario, 359, 384

conflicts of interest

minimization, 377–378

current best evidence,

367–368

decision tree, 362–364

development of, 359–362

evidence comprehensibility, 371–377

grades of recommendation, 370–371

guidelines for using,

378–384

strong recommendations, 369–376, 378–383

treatment intervention, 365

weak recommendations, 376–377, 383–384

**Patient preference, misleading results in treatment studies, 101–103****Patient randomization**

group analysis of, 109–110

noninferiority trial, 138–140

prognostic balance with

control groups, 103–104

therapy research, 100–103

**Patient similarity, prognosis and, 257–258****Pattern recognition, diagnosis, 212–213****Per-protocol analysis,****noninferiority trial, 138–140****Physician preference, misleading results in treatment studies, 101–103**

**PICO framework**

- clinical questions, 30
- current best evidence searches and, 46
- network meta-analyses, eligibility criteria, 334–335
- pyramid of EBM resources searches, 72–74
- search term selection and, 76–78

**Placebo effect**

- noninferiority trials, 128–129
- prognostic balance and, 93

**Placebo trials**

- confidence intervals and, 168
- questions involving, 41–42

**POEMS (Patient-Oriented Evidence that Matters), preappraised research, 65–67****Point estimates**

- best evidence summaries, 13–14
- confidence intervals and, 169–170
- magnitude of heterogeneity, 309–310
- meta-analyses, 295–300
- network meta-analyses, 336
- noninferiority trials, 131–134
- reporting bias, 319–321
- survival analysis, 160–161
- treatment effect studies, 112–115
- visual assessment of variability, 305–308

**Pooled estimates**

- confidence intervals and, 173–177
- in meta-analyses, 276, 294–300
- systematic reviews and meta-analyses, 294

**Pooled studies, meta-analyses, 272–277****Positive trials**

- clinical scenario, 167
- reporting bias, 317

**Posttest probability**

- clinical practice guidelines and, 6
- diagnosis, 215–216
- likelihood ratios, 236–239
- threshold probabilities and, 216–220

**Preappraised research, 53–54**

- availability and access, 58–59
- guidelines for using, 63–67

**Precision of estimation**

- GRADE confidence rating, 302–303
- prognosis likelihood estimation, 262–264
- results analysis, 312–314
- risk analysis, 202
- treatment effect studies, 112–115

**Prediction rules, diagnosis, 215****Pretest probability**

- clinical practice guidelines and, 6
- diagnostic process, 214–215
- diagnostic testing, 225
- likelihood ratios, 236–239

**Prevention of complications, therapy research, 98–100****Primary studies**

- decision aids and, 403–404
- diagnostic testing, 216
- harm, observational studies, 182–184
- hierarchy of evidence, 47–49
- inclusion criteria, 280–281
- network meta-analyses, 350–351
- nonpreappraised research, 54
- processing of evidence in, 49–50
- risk of bias, 273–274, 280–281, 283–284
- summary estimates, 293–300



**Probability**

- diagnosis based on, 212–220
- random error and, 86–88
- survival analysis, 159–161
- test for heterogeneity, 308–309

**Probability tradeoff, values and preferences of patient, 400–402****Problem solving, medical literature resources for, 28–29****Procedural interventions, expertise in, 116–118****Processing of evidence, in primary studies, 49–50****Prognosis, 251–267**

- balance of intervention vs control groups, 100–103
- clinical practice guidelines and, 4
- clinical questions concerning, 30
- clinical scenario, 253, 267
- evidence-based medicine and, 10–19
- evidence resources for, 253–254
- follow-up and risk of bias in, 259
- foreground questions concerning, 31
- maintenance of balance in clinical trials, 104–105
- measurement, 254–255
- patient care and studies of, 264–266
- patient similarity and, 258–259
- results analysis of studies in, 261–264
- risk of bias, 256–261
- studies and questions involving, 34, 39

**Prognostic factors**

- balance at trial completion, 105–110
- in cohort studies, 188–192

- multiple chronic conditions, 410–411
- noninferiority trial validity and, 134–140
- patient-control group similarity, 103–104, 257–258
- target outcomes, 101–103

**Prospective cohort studies, 187–188****Prospective study registration, reporting bias, 321****Publication bias**

- GRADE confidence rating, 302–303
- meta-analyses, 318–321
- network meta-analyses, 335–336
- systematic reviews, 282–283, 318–321

**PubMed**

- nonpreappraised research strategies, 68–72
- related article searches, 79–80
- search strategies for, 45

**P values**

- large between-study results variability, 311–312
- network meta-analyses, 351–352
- test for heterogeneity, 308–309

**Pyramid of EBM resources,**

- 50–54
- availability and access criteria, 57–59
- coverage and specificity in, 57
- guidelines for using, 59–76
- multilevel searching of, 72–74

---

## Q

**Qualitative research**

- meta-analyses, 300–322
- systematic reviews, 272

**Questions**

- background questions, 29
- clarification strategies and examples, 30–41, 45–46
- definition, in systematic reviews, 274–276
- diabetes and target blood pressure example, 35–36
- foreground questions, 29
- framing strategies, 30
- improved, searchable questions, 36–37
- patient management recommendations, 364
- on resource coverage and specificity, 57
- search terms and, 76–81
- squamous cell carcinoma example, 39–41
- strategies for defining, 41–42
- study designs matched to, 31–34
- transient loss of consciousness example, 37–38

## R

---

**Random allocation, risk of bias reduction and, 93****Random error**

- bias vs, 3
- in case-control studies, 197
- misleading results, 86–88
- pretest probabilities, 215
- systematic reviews, 287
- visual assessment of results variability, 306–308

**Randomization, risk of bias, 303–304****Randomized clinical trials (RCTs)**

- best evidence summaries, 11–14
- clinical scenario, 166–167
- confidence intervals for, 170–171
- diagnostic testing and, 247–248
- GRADE confidence rating, 302–303
- harm, observational studies, 182–187
- meta-analyses, 299–300, 330
- noninferiority trials, 128–129
- patient decision aids, 404
- patient randomization in, 100–103
- questions matched to, 31–32
- random error in, 88
- risk of bias, 303–304
- search strategies for, 45
- strong recommendations, 376
- therapy research, 99–100
- 2 x 2 table analysis, 151

**Random sampling, diagnostic testing, 232–233****Rating system, for EBM resources, 56–59****Rational Clinical Examination reviews, 238–239****Recall bias, in case-control studies, 195–197****Receiver operating characteristic (ROC) curve, diagnostic testing, 240–243****Reference standard**

- diagnostic testing, 33–34, 229–232, 241–243
- systematic review eligibility and, 279–281

**Referral bias, prognosis and, 257****Regression analysis, prognostic studies, 258**

**Related article searches,**

79–80

**Relative odds reduction,**

**meta-analyses,** 294–300

**Relative risk (RR)**

in cohort studies, 189

confidence intervals and,  
168–171

meta-analyses, 294–300

patient decision aids, 404

precision of results, 312–314

risk difference vs, 154–155

systematic reviews and

clinical application,  
285–287

treatment effect, 88, 111,

152–153

**Relative risk reduction (RRR)**

confidence intervals and,  
168–171

estimate precision,

treatment effect studies,  
112–115

evidence-based practice and,  
161–162

meta-analyses, 294–300

number needed to treat and,  
156–157

relative risk in treatment effect

studies and, 88, 111

stopped early trials and,  
108–109

treatment benefit-harm

analysis, 120–213

treatment effects, 153

**Relevance analysis, systematic reviews,** 281–283

**Reporting bias**

network meta-analyses,  
335–336

results analysis, 316–321

strategies for managing,  
318–321

systematic review searches,  
282–283

**Reproducibility of results**

diagnostic testing and patient  
care, 243–244

systematic review selection  
and assessment, 287

**Residual confounding, in cohort studies,** 190

**Results analysis**

applicability in clinical practice,  
244–245

between-study differences  
in systematic reviews,  
284–285

clinical scenario, 166–167

confidence intervals for,  
170–171

cross-study consistency,  
304–305

diagnostic testing, 233–243

harm, observational studies,  
201–202

large between-study variability,  
310–312

meta-analyses, 293–323

network meta-analyses,  
339–349

noninferiority trial, 140–142

patient applicability, 314–316

precision of results, 312–314

prognostic studies, 261–264

reporting bias, 316–321

reproducibility in clinical  
settings, 243–244

systematic review, 293–323

treatment effect studies,  
110–115

visual assessment of  
variability, 305–308

**Retrospective cohort studies,**  
187–188

**Risk analysis (risk factors)**

case-control studies, 192–197

dichotomous outcomes, 152

in harm, observational studies,  
181–182

---

## S

- precision estimation of, 202
- prognosis risk of bias and, 257
- treatment benefit-harm analysis, 119–123
- Risk-benefit tradeoff, evidence-based medicine and, 10–19**
- Risk difference. See Absolute risk reduction**
- Risk of bias**
  - in case-control studies, 193–197
  - case series/case reports, 198–199
  - clinical practice guidelines and, 4–5
  - confidence intervals and, 158
  - cross-sectional studies, 197
  - diagnostic testing, 216, 226–233
  - evidence, 303–304
  - harm, observational studies, 182–187, 197–200
  - hierarchy of evidence and minimization of, 47–49
  - loss to follow-up and, 106–108
  - network meta-analyses, 334–339
  - noninferiority trial validity, 134–140
  - primary studies, 273–274, 280–281, 283–284
  - prognosis, 257–261
  - reduction strategies, 90–93
  - systematic reviews, 273–274, 280–281, 283–284
  - therapy research, 100–103
- Risk ratio, exposure-outcome association, 201–202**
- Robustness of results, network meta-analyses, 344–349**
- ROC. See Receiver operating characteristic (ROC) curve**
- RR. See Relative risk (RR)**
- RRR. See Relative risk reduction (RRR)**
- Sample size**
  - confidence intervals and, 169–170, 172–177
  - diagnostic testing risk of bias, 226–229
  - hypothesis testing and, 166
  - prognosis risk of bias and, 257
  - treatment effect, 110–111
- Screening tool**
  - diagnostic testing, 226
  - therapy research, 98–100
- Search terms**
  - broad vs narrow searches, 78–79
  - daily practice skills, 81–82
  - information specialists and, 81
  - selection and combination strategies, 76–78
  - strategies for using, 76–81
- Secondary evidence-based journals, 28**
  - preappraised research, 65–67
- Selection bias, network meta-analyses, 335–336**
- Selection criteria for EBM resources, 55–59**
  - in systematic reviews, 274–276
- Selective outcome reporting bias**
  - network meta-analyses, 335–336
  - results analysis, 317–318
  - systematic reviews, 283
- Self-selection by patients, in cohort studies, 188–192**
- Sensitivity analysis**
  - decision analysis, 369–371
  - diagnostic testing, 239–243
  - loss to follow-up and, 107
  - robustness of results, 344–349
  - strong recommendations, 376

**Sequential diagnostic testing, clinical application of,** 246–247

**Shared physician-patient decision making**  
 alternative treatments, 411  
 approaches to, 391–403  
 bidirectional exchange, 394–402  
 challenges to patient decisions, 403–406  
 clinician-as-perfect-agent approach, 393  
 guidelines for choosing, 402–403  
 important vs unimportant decisions, 406–407  
 informed decision-making approach, 393–394  
 misinformed participants, 408–410  
 multiple chronic conditions, patients with, 410–411  
 noninferiority trial analysis, 144–145  
 paternalistic approach, 392  
 patient decision aids, 403–406, 411–412  
 straightforward vs difficult decisions, 407–408  
 time factors, 406  
 values and preferences of patient, 391  
 weak recommendations, 376–377

**Signs, questions concerning,** 33

**SIS. See Six-Item Screener (SIS) instrument**

**Six-Item Screener (SIS) instrument,** 226, 232–233  
 likelihood ratios, 234–239  
 sensitivity and specificity of diagnostic testing, 240–243

## **Specificity**

assessment in EBM  
 resources, 57  
 diagnostic testing, 239–243

**Spectrum bias, diagnostic testing,** 228

**Standard error, reporting bias,** 319–321

**Standard treatment effects, noninferiority trial,** 136–140

**Star network, meta-analyses,** 340

**Statistical significance, tests of**  
 magnitude of heterogeneity, 309–310  
 meta-analyses, 299–300  
 network meta-analyses, 341–343  
 reporting bias, 321  
 risk estimation, 202  
 treatment effect studies, estimate precision, 112–115

## **Stopped early trials**

bias in, 90  
 risk of bias, 304  
 systematic reviews, 284  
 treatment effect assessment, 108–109  
 treatment effect studies, 108–109

**Straightforward decisions, shared physician-patient decision making,** 407–408

**Stratum-specific likelihood ratios, diagnostic testing,** 242–243

**Strong recommendations**  
 application of, 378–383  
 patient management recommendations, 369–376

**Study design, matching questions to,** 31–34

**Subgroup analyses**

- between-study differences, 336–337
- cross-study consistency of results, 305
- diagnostic test results
  - applicability and, 244–245
- meta-analyses and, 283
- network meta-analyses, 351–352
- therapy studies, 117–118

**Substitute/surrogate outcomes**

- decision analysis, 366–367
- patient applicability of results, 315–316
- patient-important outcomes in therapy research, 118–119
- questions involving, 42

**Summaries and guidelines**

- in EBM resources, 53, 60–62
- evidence profile, 322–323

**Summary estimates, meta-analyses, 293–300****Summary-of-findings table, 323, 376–377****SumSearch, pyramid of EBM resources searches, 72–74****Superiority trials, defined, 128–129****Surveillance bias, in cohort studies, 191****Survival analysis**

- clinical scenario, 167
- meta-analyses, 294–300
- treatment effect studies, 111, 158–161

**Survival curve**

- prognostic studies, 261–262
- treatment effects studies, 159–161

**Symptoms**

- questions concerning, 33
- strong recommendations and reduction of, 382–383

**Synopses**

- dissemination bias, 317
- preappraised research, 65–67

**Systematic error. See also Bias**

- bias, 303–304
- confidence intervals and, 168–169
- systematic reviews, 287

**Systematic reviews, 4**

- absolute effect estimates, 300
- benefits, 274
- best evidence summaries, 13–14
- between-study results
  - differences, 284–285
- clinical applications based on, 285–287
- clinical practice guidelines, 359–360
- clinical scenario, 271, 288–289, 293
- confidence intervals and, 173–177
- credibility of process, 277–281
- decision aids and, 403–404
- definitions, 272–277
- diagnostic testing, 216
- effect estimate credibility, 276–277, 287–288
- evidence-based summary, 322–323
- evidence sources, 271–272
- filtering search strategies for, 70–72
- harm, observational studies, 182–184
- likelihood ratios, diagnostic testing, 238–239
- meta-analyses and, 272–273
- preappraised research, 53–54
- preappraised research synopses, 65–67
- procedures in, 271–289
- processing of evidence in, 49–50

**Systematic reviews** (*continued*)

- process synopsis, 274–276
- quality of evidence, 300–322
- relevance criteria, 281–283
- reporting bias, 316–321
- reproducibility of study
  - selection and assessments, 287
- results analysis, 293–323
- risk of bias, 283–284
- search strategies for, 45
- selection of EBM resources
  - using, 56–59
- summary estimates, 293–300
- therapy research, 97

---

## T

**Talk model, shared physician-patient decision making, 394–402****Target condition**

- diagnostic testing risk of bias, 227–229
- foreground questions
  - concerning, 31
- management strategies based on diagnostic testing, 245–247

**Target events, bias and, 89–90****Target-negative patients, diagnostic testing, 228–229****Target outcomes**

- bias and, 89–90
- prognosis and, 34, 101–103, 260

**Target-positive patients, diagnostic testing, 228–229****Team talk model, shared physician-patient decision making, 394–396****Tertiary care centers, prognosis risk of bias and, 257****Test threshold**

- diagnostic testing, 225
- management strategies based on diagnostic testing, 245–247
- posttest probabilities, 217–220

**Theoretical distribution of results, 87–88****Therapeutic threshold, diagnostic testing, 225****Therapy studies. See also****Treatment intervention**

- clinical scenario example, 97–98, 122–123
- diagnostic testing and, 247–248
- evidence in, 97–98
- foreground questions in, 31
- improved, searchable
  - questions involving, 36–37
- patient care applications, 115–123
- randomized trials, 95–123
- risk of bias in research on, 100–103
- systematic reviews and meta-analyses, 294–300

**Threshold interpretations**

- confidence intervals and, 176–177
- noninferiority trials, 132–134
- posttest probabilities, 216–220

**Time factors, shared physician-patient decision making, 406–407****Time-to-event methods, survival analysis, 294–300****Treatment effect**

- bias in, 101–103
- confidence intervals and, 169–170
- meta-analyses, 330
- precision of estimates in studies of, 112–115
- random error and, 86

- reporting bias, 318–321
  - size of, 110–111
  - systematic review eligibility and, 278–281
  - uncertainty concerning magnitude of, 88
  - Treatment intervention**
    - bias and, 89–90
    - network meta-analyses, 330–334, 350–351
    - patient applicability of results, 315–316
    - patient management recommendations, 365
    - procedural intervention expertise, 116–118
    - risk of bias in assessment of, 100–103
  - Treatment network, meta-analyses**, 339–340, 346
  - Treatment ranking, network meta-analyses**, 343–344
  - Treatment threshold**
    - management strategies based on diagnostic testing, 245–247
    - posttest probabilities, 216–220
  - Trim-and-fill method, reporting bias**, 321
  - 2 x 2 table**
    - continuous diagnostic testing, 239–243
    - dichotomous outcomes analysis, 150–151
    - survival analysis, 158–161
- 
- U
- Unblinded assessment, diagnostic testing**, 230
  - Unimportant decisions, shared physician-patient decision making**, 406–407
  - Unrepresentative sampling, prognosis risk of bias and, 257**
  - Users' Guides to the Medical Literature***
    - advanced topics and, 7–8
    - credibility of systematic reviews, 276–277
    - current best evidence searches and, 46
    - diagnostic testing risk of bias, 226–227, 232–233, 243
    - harm, observational studies, 183, 200
    - large between-study results variability, 311–312
    - network meta-analyses, 333–334, 339, 345–349
    - noninferiority trial guidelines, 133–140
    - patient applicability of results, 316
    - patient decision aids, 404
    - preappraised research synopses, 67
    - precision of results, 312–314
    - prognosis articles, 255, 260–261, 264
    - reporting bias, 321
    - risk of bias, 304
    - shared physician-patient decision making, 410
    - strong recommendations, 381–383
    - structure of, 2–7
    - therapy research, 98–100
    - treatment recommendations, 364
  - US Preventive Services Task Force (USPSTF), patient management recommendations**, 370–371
  - Utility analyses**
    - patient management recommendations, 368–369
    - strong recommendations, 371–376



---

## V

---

**Validity**

- noninferiority trial results, 134–140
- risk of bias and, 5

**Values and preferences of patient**

- clinical practice guidelines and, 4, 18–19
- clinician sharing of, 400–402
- confidence intervals and, 177
- decision analysis, 360–362
- decision making and, 391–412
- evidence-based medicine and, 10–19
- noninferiority trial analysis, 144–145
- patient management recommendations, 368–369
- processing of evidence and, 49–50
- risk of bias and, 200
- treatment benefit-harm analysis, 122–123

**Variability of results**

- large between-study variability, 310–312
- visual assessment, 305–308

**Variance in studies, reporting bias,** 319–321

**Verification bias, diagnostic testing,** 230–232

---

## W

---

**Weak recommendations**

- application of, 383–384
- shared decision making, 376–377

**Weighted mean difference, meta-analyses,** 294–300

**Weighting process, systematic reviews and meta-analyses,** 294–300

**Wikipedia, EBM research and,** 75–76

**Working diagnoses,** 214

**Workup bias, diagnostic testing,** 230–232

---

## Y

---

**Yes-or-no statistical tests, heterogeneity,** 308–309