Methods Guide for Medical Test Reviews

Prepared for:
Center for Outcomes and Effectiveness
Agency for Healthcare Research and Quality (AHRQ)
U.S. Department of Health and Human Services
Rockville, MD 20850
www.ahrq.gov

Editors:
Stephanie M. Chang, M.D., M.P.H., AHRQ
David B. Matchar, M.D., Duke University

Journal of General Internal Medicine Guest Editors:
Gerald W. Smetana, M.D., Beth Israel Deaconess Medical Center and Harvard Medical School
Craig A. Umscheid, M.D., M.S., University of Pennsylvania

Editorial Assistants:
Rebecca Gray, D. Phil., Duke University
Marion M. Torchia, Ph.D., AHRQ

AHRQ Publication No. 12-EHC017
June 2012

This document was written with support from the Effective Health Care Program at the Agency for Healthcare Research and Quality (AHRQ). Its findings and conclusions are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this document should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

None of the authors has any affiliations or financial involvements that conflict with the information presented in this chapter.

Suggested citation: Methods Guide for Medical Test Reviews. AHRQ Publication No. 12-EC017. Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published as a special supplement to the Journal of General Internal Medicine, July 2012.

Preface

Effectiveness and Comparative Effectiveness Reviews, systematic reviews of existing research on the effectiveness, comparative effectiveness, and comparative harms of different medical tests, are intended to provide relevant evidence to inform real-world health care decisions for patients, providers, and policymakers. In an effort to improve the transparency, consistency, and scientific rigor of the work of the Effective Health Care (EHC) Program, the Agency for Healthcare Research and Quality (AHRQ), the Scientific Resource Center, and the Evidence-based Practice Centers (EPCs), have developed this *Methods Guide for Medical Test Reviews* (also referred to as the *Medical Test Methods Guide).* We intend it to serve as a resource for the EPCs as well as for other investigators interested in conducting systematic reviews on medical tests. We hope it will be a practical guide both for those who prepare the systematic reviews and those who use them in clinical practice, research development, and in making policy decisions].

The *Methods Guide for Medical Test Reviews* complements the *EPC Methods Guide for Comparative Effectiveness Reviews* (also referred to as the *General Methods Guide),* which focuses on methods to assess the effectiveness of treatments and other health care interventions. The guidance in this *Medical Test Methods Guide* applies the principles for the assessment of treatments to the issues and challenges entailed in assessing medical tests. It highlights particular areas where the inherently different qualities of medical tests necessitate a different or variation of the approach to systematic review compared to a review on treatments. It provides step-by-step guidance for those conducting a systematic review.

The series of chapters comprising this Medical Test Methods Guide is the result of a collaborative effort by EPC authors and external peer reviewers. In addition, leaders from AHRQ and the EPC program have collaborated with the editors of the Journal of General Internal Medicine to co-publish the Guide, which will be simultaneously posted on AHRQ's Effective Health Care Web site at www.effectivehealthcare.ahrq.gov/reports/final.cfm and as a special supplement to the journal (Volume 27, Supplement 1).

This *Medical Test Methods Guide* is a living document, and will be updated as further empirical evidence develops and our understanding of better methods improves. Comments and Suggestions on the *Methods Guides* and the Effective Health Care Program can be made at www.effectivehealthcare.gov.

Carolyn M. Clancy, M.D.
Director, Agency for Healthcare Research
and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie M. Chang M.D., M.P.H.
Director, EPC Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

AHRQ's Evidence-Based Practice Centers

In 1997, the Agency for Health Care Policy and Research (AHCPR), now known as the Agency for Healthcare Research and Quality (AHRQ), launched its initiative to promote evidence-based practice in everyday care through establishment of the Evidence-based Practice Center (EPC) Program. The EPCs develop evidence reports and technology assessments on topics relevant to clinical and other health care organization and delivery issues;—specifically those that are common, expensive, and/or significant for the Medicare and Medicaid populations. With this program, AHRQ has become a "science partner" with private and public organizations in their efforts to improve the quality, effectiveness, and appropriateness of health care by synthesizing the evidence and facilitating the translation of evidence-based research findings. Topics are nominated by non-federal partners such as professional societies, health plans, insurers, employers, and patient groups.

The EPCs are located at:

- Blue Cross and Blue Shield Association, Technology Evaluation Center, Chicago, IL
- Duke University, Durham, NC
- ECRI Institute, Plymouth Meeting, PA
- Johns Hopkins University, Baltimore, MD
- McMaster University, Hamilton, Ontario, CA
- Oregon Health & Science University, Portland, OR
- RAND Corporation, Santa Monica, CA
- RTI International—University of North Carolina, Chapel Hill, NC
- Tufts—New England Medical Center, Boston, MA
- University of Alberta, Edmonton, Alberta, CA
- University of Connecticut, Hartford, CT
- University of Minnesota—Minneapolis VA, Minneapolis, MN
- University of Ottawa, Ottawa, CA
- Vanderbilt University, Nashville, TN

For contacts and additional information about the current participating EPCs, go to: http://www.ahrq.gov/clinic/epc/epcenters.htm.

The Effective Health Care Program

The Effective Health Care Program was initiated in 2005 to provide valid evidence about the comparative effectiveness of different medical interventions. The object is to help consumers, health care providers, and others in making informed choices among treatment alternatives. Through its Comparative Effectiveness Reviews the Program supports systematic appraisals of existing scientific evidence regarding treatments for high-priority health conditions. It also promotes and generates new scientific evidence by identifying gaps in existing scientific evidence and supporting new research. The program puts special emphasis on translating findings for different stakeholders, including consumers.

List of Authors

Cynthia M. Balion, Ph.D.
Hamilton General Hospital and McMaster Evidence-based Practice Center
Hamilton, Ontario, CA

Eric B. Bass, M.D., M.P.H.
Johns Hopkins University School of Medicine and Johns Hopkins University Bloomberg School of Public Health
Baltimore, MD

Stephanie M. Chang, M.D., M.P.H.
Agency for Healthcare Research and Quality
Rockville, MD

Craig I. Coleman, Pharm.D.
University of Connecticut/Hartford Hospital Evidence-based Practice Center and University of Connecticut School of Pharmacy
Storrs, CT

Rongwei Fu, Ph.D.
Oregon Evidence-based Practice Center
Portland, OR

Lauren Griffith, Ph.D.
McMaster Evidence-based Practice Center
Hamilton, Ontario, CA

Katherine E. Hartmann, M.D., Ph.D.
Institute for Medicine and Public Health
Vanderbilt University
Nashville, TN

Daniel E. Jonas, M.D., M.P.H.
University of North Carolina
Chapel Hill, NC

Shalini Kulasingam, Ph.D.
University of Minnesota School of Public Health
Minneapolis, MN

William F. Lawrence, M.D., M.S.
Agency for Healthcare Research and Quality
Rockville, MD

David B. Matchar, M.D., FACP, FAHA
Duke-NUS Graduate Medical School Singapore;
Duke University Medical Center
Durham, NC

Thomas S. Rector, Ph.D.
Minneapolis Veterans Affairs Medical Center
Minneapolis, MN

Rose Relevo, M.L.I.S, M.S.
Oregon Health & Science University
Portland, OR

Crystal M. Riley, M.A.
Nanyang Technological University
Singapore

David Samson, M.S.
Blue Cross and Blue Shield Association Technology Evaluation Center
Chicago, IL

P. Lina Santaguida, B. Sc. P.T., Ph.D., M.Sc.
McMaster Evidence-based Practice Center
Hamilton, Ontario, CA

Karen M. Schoelles M.D., S.M., FACP
ECRI Institute Health Technology Assessment Group
Plymouth Meeting, PA

Jodi B. Segal, M.D., M.P.H.
Johns Hopkins University School of Medicine
Baltimore, MD

Sonal Singh, M.D., M.P.H.
Johns Hopkins University School of Medicine and Johns Hopkins University Bloomberg School of Public Health
Baltimore, MD

Brent C. Taylor, Ph.D., M.P.H.
Department of Veterans Affairs Health Care System
University of Minnesota
Minneapolis, MN

Thomas A. Trikalinos, M.D.
Tufts Evidence-based Practice Center
Boston, MA

Ben Vandermeer, M.Sc.
University of Alberta Evidence-based Practice Center
Edmonton, Alberta, CA

Tania M. Wilkins, M.S.
University of North Carolina
Chapel Hill, NC

Timothy J. Wilt, M.D., M.P.H.
Department of Veterans Affairs Health Care System
University of Minnesota
Minneapolis, MN

List of Peer Reviewers

Rex Astles
Centers for Disease Control and Prevention

David Atkins
Department of Veterans Affairs

Claudia Bonnell
BlueCross BlueShield Association Technology Evaluation Center

Joseph Boone
Battelle

Patrick Bossuyt
University of Amsterdam

Rose Campbell
Oregon Health & Sciences University

Rob Christenson
University of Maryland School of Medicine

Nananda Col
Tufts Medical School

Benjamin Djulbegovic
University of South Florida

Eileen Erinoff
ECRI Institute

Mark Helfand
Oregon Health & Sciences University

Joseph Jacobs
Abbott Molecular

Louis Jacques
Centers for Medicare & Medicaid Services

John Krolak
Centers for Disease Control and Prevention

Shalini Kulasingam
University of Minnesota

Joseph Lau
Tufts University

Kathleen Lohr
RTI International

Sally Lord
University of Sydney

Kirsten McCaffrey
University of Sydney

Diana Petitti
Arizona State University

Greg Raab
Raab & Associates, Inc.

Max Robinowitz
U.S. Food and Drug Administration

Jeff Roche
Center for Medicare & Medicaid Services

Donald Rucker
Siemens

Indy Rutks
Veterans Affairs Medical Center

Holger Schunemann
McMaster University

Jodi Segal
Johns Hopkins University

Lisa Tjosvold
University of Alberta

Frederic Wolf
University of Washington

—And members of the public

Contents

Editorial. Methods Guide for Authors of Systematic Reviews of Medical Tests: A Collaboration Between the Agency for Healthcare Research and Quality (AHRQ) and the Journal of General Internal Medicine E-1

Chapter 1. Introduction to the Methods Guide for Medical Test Reviews 1-1
Abstract 1-1
Introduction 1-1
Development of the *Medical Test Methods Guide* 1-2
Unique Challenges of Medical Tests 1-3
Recurrent Themes in the Test Evaluation Literature 1-4
Analytic Frameworks 1-4
A Note on Terminology 1-5
PICOTS Typology 1-6
Organization of This *Medical Test Methods Guide* 1-7
Summary 1-8
References 1-9

Chapter 2. Developing the Topic and Structuring Systematic Reviews of Medical Tests: Utility of PICOTS, Analytic Frameworks, Decision Trees, and Other Frameworks 2-1
Abstract 2-1
Introduction 2-1
Common Challenges 2-2
Principles for Addressing the Challenges 2-2
Principle 1: Engage Stakeholders Using the PICOTS Typology 2-2
Principle 2: Develop an Analytical Framework 2-3
Principle 3: Consider Using Decision Trees 2-4
Principle 4: Sometimes it is Sufficient To Focus Exclusively on Accuracy Studies 2-5
Principle 5: Other Frameworks may be Helpful 2-6
Illustrations 2-6
Summary 2-12
References 2-12

Chapter 3. Choosing the Important Outcomes for a Systematic Review of a Medical Test 3-1
Abstract 3-1
Introduction 3-1
Common Challenges 3-2
Principles for Addressing the Challenges and Recommended Approaches for Incorporating All Decision-Relevant Outcomes 3-2
Principle 1: Catalog Outcomes Methodically 3-3
Principle 2: Solicit Input From Stakeholders 3-5
Illustrations of the Principles 3-6
Summary 3-11
References 3-11

Chapter 4. Effective Search Strategies for Systematic Reviews of Medical Tests 4-1
Abstract 4-1
Introduction 4-1
Common Challenges 4-2
Principles for Addressing the Challenges 4-2
Principle 1: Do not Rely on Search Filters Alone 4-3
Principle 2: Do not Rely on Controlled Vocabulary (Subject Headings) Alone 4-3
Principle 3: Search in Multiple Locations 4-3
Illustration: Contrasting Search Strategies 4-6
Summary 4-6
References 4-7
Chapter 5. Assessing Risk of Bias as a Domain of Quality in Medical Test Studies 5-1
Abstract 5-1
Introduction 5-1
Evidence for Biases Affecting Medical Test Studies 5-2
Common Challenges 5-4
Principles for Addressing the Challenges 5-4
Principle 1: Use Validated Criteria To Address Relevant Sources of Bias 5-4
Principle 2: Standardize the Application of Criteria 5-5
Principle 3: Decide When Inadequate Reporting Constitutes a Fatal Flaw 5-6
Illustration 5-8
Summary 5-8
Key Points 5-8
References 5-8
Chapter 6. Assessing Applicability of Medical Test Studies in Systematic Reviews 6-1
Abstract 6-1
Introduction 6-1
Common Challenges 6-2
Unclear Key Questions 6-2
Studies Not Specific to the Key Questions 6-2
Rapidly Evolving Tests 6-3
Principles for Addressing the Challenges 6-3
Principle 1: Identify Important Contextual Factors 6-6
Principle 2: Be Prepared To Deal With Additional Factors Affecting Applicability 6-8
Principle 3: Justify Decisions To "Split" or Restrict the Scope of a Review 6-8
Principle 4: Maintain a Transparent Process 6-9
An Illustration 6-9
Summary 6-10
Key Points 6-11
References 6-12
Chapter 7. Grading a Body of Evidence on Diagnostic Tests 7-1
Abstract 7-1
Introduction 7-1
Common Challenges 7-3
Principles for Addressing the Challenges 7-4

Principle 1: Methods for Grading Intervention Studies can be Adapted for Studies Evaluating Studies on Diagnostic Tests With Clinical Outcomes 7-4
Principle 2: Consider Carefully What Test Characteristic Measures are the Most Appropriate Intermediate Outcomes for Assessing the Impact of a Test on Clinical Outcomes and for Assessing the Test's Precision in the Clinical Context Represented by the Key Question 7-5
Principle 3: The Principle Domains of GRADE can be Adapted To Assess a Body of Evidence on Diagnostic Test Accuracy 7-6
Principle 4: Additional GRADE Domains can be Adapted To Assess a Body of Evidence With Respect to Diagnostic Test Accuracy 7-9
Principle 5: Multiple Domains Should be Incorporated Into an Overall Assessment in a Transparent Way 7-9
Illustration 7-10
Summary 7-12
Key Points 7-12
References 7-13
Chapter 8. Meta-Analysis of Test Performance When There Is a "Gold Standard" 8-1
Abstract 8-1
Introduction 8-1
Nonindependence of Sensitivity and Specificity Across Studies and Why It Matters for Meta-Analysis 8-3
Principles for Addressing the Challenges 8-7
Recommended Approaches 8-7
Which Metrics To Meta-Analyze 8-7
Desired Characteristics of Meta-Analysis Methods 8-9
Preferred Methods for Obtaining a "Summary Point" (Summary Sensitivity and Specificity): Two Families of Hierarchical Models 8-10
Preferred Methods for Obtaining a "Summary Line" 8-11
A Special Case: Joint Analysis of Sensitivity and Specificity When Studies Report Multiple Thresholds 8-11
A Workable Algorithm 8-12
Step 1: Start by Considering Sensitivity and Specificity Independently 8-13
Step 2: Perform Multivariate Meta-Analysis (When Each Study Reports a Single Threshold) 8-13
Step 3. Explore Between-Study Heterogeneity 8-14
Illustrations 8-14
First Example: D-dimers for Diagnosis of Venous Thromboembolism 8-14
Second Example: Serial Creatine Kinase-MB Measurements for Diagnosing Acute Cardiac Ischemia 8-15
Overall Recommendations 8-16
References 8-17
Chapter 9. Options for Summarizing Medical Test Performance in the Absence of a "Gold Standard" 9-1
Abstract 9-1
Introduction 9-1
Imperfect Reference Standards 9-2

What is Meant by "Imperfect Reference Standard," and Why is it Important for Meta-Analysis and Synthesis in General? 9-2
Options for Systematic Reviewers 9-5
Option 1. Assess the Index Test's Ability To Predict Patient-Relevant Outcomes Instead of Test Accuracy 9-6
Option 2. Assess the Concordance of Difference Tests Instead of Test Accuracy 9-6
Option 3. Qualify the Interpretation of Naïve Estimates of the Index Test's Performance 9-7
Option 4. Adjust or Correct the Naïve Estimates of Sensitivity and Specificity 9-8
Illustration 9-8
Identifying (Defining) the Reference Standard 9-9
Deciding How To Summarize the Findings of Individual Studies and How To Present Findings 9-9
Qualitative Analyses of Naïve Sensitivity and Specificity Estimates 9-9
Qualitative Assessment of the Concordance Between Measurement Methods 9-9
Summary 9-12
References 9-13
Chapter 10. Deciding Whether To Complement a Systematic Review of Medical Tests With Decision Modeling 10-1
Abstract 10-1
Introduction 10-1
A Workable Algorithm 10-2
Step 1. Define how the Test Will be Used 10-2
Step 2. Use a Framework To Identify Consequences of Testing as Well as Management Strategies for Each Test Result 10-3
Step 3. Assess Whether Modeling may be Useful 10-3
Step 4. Evaluate Prior Modeling Studies 10-3
Step 5. Consider Whether Modeling is Practically Feasible in the Given Time Frame 10-6
Illustration 10-6
Step 1: Define how PET will be Used 10-7
Step 2: Create a Simplified Analytic Framework and Outline how Patient Management Will be Affected by Test Results 10-7
Step 3: Assess Whether Modeling Could be Useful in the PET and AD Evidence Report 10-9
Step 4: Assess Whether Prior Modeling Studies Could be Utilized 10-9
Step 5. Consider Whether Modeling is Practically Feasible in the Time Frame Given 10-10
Overall Suggestions 10-10
References 10-10
Chapter 11. Challenges in and Principles for Conducting Systematic Reviews of Genetic Tests Used as Predictive Indicators 11-1
Abstract 11-1
Introduction 11-1
Common Challenges 11-2
Penetrance 11-2

Time Lag 11-3
Variable Expressivity 11-3
Pleiotropy 11-3
Other Common Challenges 11-3
Principles for Addressing the Challenges 11-4
Principle 1: Use an Organizing Framework Appropriate for Genetic Tests 11-4
Principle 2: Develop Analytic Frameworks That Reflect the Predictive Nature of Genetic Tests and Incorporate Appropriate Outcomes 11-4
Principle 3: Search Databases Appropriate for Genetic Tests 11-7
Principle 4: Consult With Experts To Determine Which Technical Issues are Important To Address in Assessing Genetic Tests 11-8
Principle 5: Distinguish Between Functional Assays and DNA-Based Assays To Determine Important Technical Issues 11-9
Principle 6: Evaluate Case-Control Studies Carefully for Potential Selection Bias 11-9
Principle 7: Determine the Added Value of the Genetic Test Over Existing Risk Assessment Approaches 11-10
Principle 8: Understand Statistical Issues of Particular Relevance to Genetic Tests 11-10
Illustrations 11-11
Conclusions 11-13
References 11-14
Chapter 12. Systematic Review of Prognostic Tests 12-1
Abstract 12-1
Introduction 12-1
Step 1: Developing the Review Topic and Framework 12-2
Step 2: Searching for Studies 12-3
Step 3: Selecting Studies and Assessing Quality 12-4
Step 4: Extracting Statistics to Evaluate Test Performance 12-6
Discrimination Statistics 12-6
Reclassification Tables 12-6
Predictive Values 12-8
Step 5: Meta-Analysis of Estimates of Outcome Probabilities 12-8
Conclusion 12-9
Key Points 12-9
References 12-10

Tables

Table 1–1. Different Objectives of Medical Test Evaluation Studies 1-4
Table 1–2. The PICOTS Typology as Applied to Interventions and Medical Tests 1-7
Table 2–1. Examples of Initially Ambiguous Claims That Were Clarified Through the Process of Topic Development 2-8
Table 3–1. Outcomes That Might be Particularly Consequential Depending on Type of Medical Test 3-3
Table 4–1. Diagnosis Clinical Query for PubMed 4-2
Table 4–2. Specalized Databases 4-4
Table 4-3. Citation Tracking Databases 4-5

Table 5-1. Commonly Reported Sources of Systematic Bias in Studies of Medical Test Performance 5-3
Table 5–2. QUADAS-2 Questions for Assessing Risk of Bias in Diagnostic Accuracy Studies 5-5
Table 5–3. Categorizing Individual Studies Into General Quality Classes 5-6
Table 5–4. Interpretation of Partial Verification Bias: the Example of Family History 5-7
Table 6–1. Using the PICOTS Framework To Assess and Describe Applicability of Medical Tests 6-4
Table 7–1. Example of the Impact of Precision of Sensitivity on Negative Predictive Value 7-4
Table 7–2. Required and Additional Domains and Their Definitions 7-6
Table 7–3. Illustration of the Approach To Grading a Body of Evidence on Diagnostic Tests - Identifying Norovirus in a Health Care Setting 7-11
Table 8–1. Commonly Used Methods for Meta-Analysis of Medical Test Performance 8-9
Table 8–2. Meta-Regression-Based Comparison of Diagnostic Performance 8-16
Table 9–1. Parameterization When the Reference Standard is Assumed "Perfect" ("Gold Standard") 9-3
Table 9–2. Situations Where one can Question the Validity of the Reference Standard 9-4
Table 9–3. Parameterization When the Reference Test is Assumed To be Imperfect, and the Index and Reference Test Results are Assumed Independent Within the Strata of the Condition of Interest 9-4
Table 10–1: Proposed Algorithm To Decide if Modeling Should be a Part of the Systematic Review 10-2
Table 10–2. Cross-Tabulation of PET Results and Actual Clinical Status Among Patients With Initial Clinical Examination Suggestive of Alzheimer's 10-7
Table 11–1. Principles for Addressing Common Challenges When Evaluating Genetic Tests Used as Predictive Indicators 11-4
Table 11–2. ACCE Model Questions for Reviews of Genetic Tests 11-5
Table 11–3. Questions for Assessing Preanalytic, Analytic, and Postanalytic Factors for Evaluating Predictive Genetic Tests 11-8
Table 12–1. General PICOTS Typology for Review of Prognostic Tests 12-3
Table 12–2. Outline of Questions for Judging the Quality of Individual Studies of Prognostic Tests 12-5
Table 12–3. Example of a Reclassification Table Based on Predicted Outcome Probabilities 12-7

Figures

Figure 1–1. Causal Chain Diagram 1-3
Figure 1–2. A Mapping Across Three Major Organizing Frameworks for Evaluating Clinical Tests 1-6
Figure 2–1. Application of USPSTF Analytic Framework To Test Evaluation 2-4
Figure 2–2. Example of an Analytical Framework Within an Overarching Conceptual Framework in the Evaluation of Breast Biopsy Techniques 2-7
Figure 2–3. Replacement Test Example: Full-Field Digital Mammography Versus Screen-Film Mammography 2-9

Figure 2–4. Add-on Test Example: HER2 Protein Expression Assay Followed by HER2 Gene Amplification Assay To Select Patients for HER2-Targeted Therapy 2-10
Figure 2–5. Triage Test Example: Positron Emission Tomography (PET) To Decide Whether To Perform Breast Biopsy Among Patients With a Palpable Mass or Abnormal Mammogram 2-11
Figure 3–1. Balance of Outcomes Against Resources 3-2
Figure 3–2. Mapping Outcomes to the Testing Process and to the Test Results 3-5
Figure 3–3. Screening Example: Bacterial Vaginosis 3-7
Figure 7–1. Steps in Grading a Body of Evidence on Diagnostic Test Accuracy Outcomes 7-10
Figure 8–1. Typical Data on the Performance of a Medical Test (D-Dimers for Venous Thromboembolism) 8-4
Figure 8–2. Obtaining Summary (Overall) Metrics for Medical Test Performance 8-8
Figure 8–3. Graphical Presentation of Studies Reporting Data at Multiple Thresholds 8-12
Figure 8–4. HSROC for the ELISA-Based D-Dimer Tests 8-15
Figure 8–5. Sensitivity 1–Specificity Plot for Studies of Serial CK-MB Measurements 8-16
Figure 9–1. Correspondence of Test Results and True Proportions in the 2 X 2 Table 9-3
Figure 9–2. Naïve Estimates Versus True Values for the Performance of the Index Test 9-5 With an Imperfect Reference Standard 9-5
Figure 9–3. "Naïve" Estimates of the Ability of Portable Monitors Versus Laboratory-Based Polysomnography To Detect AHI>15 Events/Hour 9-10
Figure 9–4. Illustrative Example of a Difference Versus Average Analysis of Measurements With Facility-Based Polysomnography and Portable Monitors 9-11
Figure 9–5. Schematic Representation of the Mean Bias and Limits of Agreement Across Several Studies 9-12
Figure 10–1. Simplified Analytic Framework 10-8
Figure 10–2. Management Options for Mild Cognitive Impairment 10-9
Figure 11–1. Generic Analytic Framework for Evaluating Predictive Genetic Tests 11-6
Figure 11–2. Generic Analytic Framework for Evaluating Predictive Genetic Tests When the Impact on Family Members is Important 11-7
Figure 11–3. Analytic Framework for Evidence Gathering on CYP450 Genotype Testing for SSRI Treatment of Depression 11-12

Appendix: Test Performance Metrics

Editorial

Methods Guide for Authors of Systematic Reviews of Medical Tests: A Collaboration Between the Agency for Healthcare Research and Quality (AHRQ) and the Journal of General Internal Medicine

Gerald W. Smetana, M.D., Beth Israel Deaconess Medical Center, and Harvard Medical School, Boston, MA
Craig A. Umscheid, M.D., University of Pennsylvania Perelman School of Medicine, Philadelphia, PA
Stephanie Chang, M.D., Agency for Healthcare Research and Quality, Rockville, MD
David B. Matchar, M.D., National University of Singapore Graduate Medical School, Singapore, and Duke University Medical Center, Durham, NC

Over the past two decades, systematic reviews have risen in number, quality, and impact. The sheer volume of work is remarkable. For example, the annual number of meta-analyses (a subset of systematic reviews) indexed by MEDLINE has grown from 273 in 1990 to 4,526 in 2010. Well conceived and well written systemic reviews serve many functions for stakeholders. First, they help clinicians apply evidence from the medical literature to patient care by critically appraising and summarizing what is often, for a given topic, a large amount of published clinical investigation. Systematic reviews are particularly useful when substantial practice variation exists, actual practice differs from published standards of care, clinical guidelines differ in their recommendations, and a large body of recent literature provides new insights that may modify recommendations from those of published guidelines.

Second, systematic reviews can provide the basis for establishing and revising clinical guidelines as well as many quality-assessment metrics applied to physicians, group practices, and hospitals. Third, they can inform future research agendas by defining important unresolved questions. Lastly, they draw attention to differences in findings across studies addressing similar research questions, and propose a basis for the conflicting results. For all of these reasons, their impact can be substantial. For example, in one study of 170 journals in the fields of general internal medicine, family practice, nursing, and mental health, the average impact factor for systematic reviews was 26.5.[1] In contrast, the mean impact factor for the top 40 general medical journals is 7.4.

Guidelines have evolved to assist authors of systematic reviews in medicine. Published in 1999, the QUORUM (Quality of Reporting of Meta-Analyses) guideline for reporting systematic reviews[2] aimed to standardize and improve published reports of systematic reviews. Subsequent evolution of review methods, including increasingly rigorous assessments of the risk of bias and more frequent inclusion of observational data, prompted the development of an updated reporting tool, PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), which was published in 2009.[3,4] PRISMA aims to standardize the reporting of

systematic reviews; it offers less guidance to authors on the conduct and performance of such reviews. Guidelines also exist to assist authors in the conduct of reviews. Since 1994, the Cochrane Collaboration has published, and regularly updated, a detailed handbook for authors of systematic reviews.[5] This methods guide focuses primarily on reviews of randomized controlled trials of interventions. While developed for authors of Cochrane reviews, the handbook is freely available and has been a helpful resource for other authors of systematic reviews.

One mission of the Agency for Healthcare Research and Quality (AHRQ) is to solicit and publish systematic reviews (evidence reports and technology assessments) on topics to improve the clinical practice and delivery of healthcare services. In 1997, AHRQ formed the Evidence-based Practice Center (EPC) Program to commission and oversee these reviews.[6] One of us (SC) directs the EPC program under the umbrella of the Center for Outcomes and Evidence, and one (DM) directed the Duke EPC. EPCs conduct reviews for use by a variety of groups, including national guideline groups such as the U.S. Preventive Services Task Force (USPSTF),[7] which uses reviews to inform screening and prevention guidelines, and payers, such as the Centers for Medicare and Medicaid Services. To improve the quality and consistency of EPC reports, the Agency has published methods guidance, developed by EPC authors (*Methods Guide for Effectiveness and Comparative Effectiveness Reviews*; hereinafter called the *General Methods Guide*).[8] This guidance, along with those of other groups such as the Cochrane Collaboration, the USPSTF,[9] and the Institute of Medicine[10], form the basis for a standards in the conduct of systematic reviews.

The editors of the AHRQ *General Methods Guide* realized, however, that systematic reviews of medical tests pose unique challenges that are not adequately addressed in guidelines for authors of reviews of interventions or comparative efficacy. For example, the principal "outcome" of a study of a medical test is commonly a proxy or intermediate outcome. An illustration of this is ultrasound evaluation of the carotid arteries. The most common outcome in an evaluation of this test is the accuracy of the test in identifying clinically significant stenosis of the artery. Clinicians, while interested in this proxy outcome, would find more value in the ability of the test to predict clinically significant outcomes (such as 10-year risk of stroke or cardiovascular death) or the effect of subsequent carotid endarterectomy or stenting on stroke or death rates. A review of the operating characteristics of carotid ultrasound would optimally assess both the proxy result (as compared to a reference standard, in this case, invasive angiography) and the downstream result of testing on clinically significant outcomes.

Clinicians obtain medical tests for a number of non-overlapping reasons. These include screening, diagnosis, prognosis, and prediction of treatment response. In recognition of the unique challenges in conducting reviews of diagnostic tests, the Cochrane Collaboration has formed a working group specifically tasked with providing guidance in this arena. A draft version of their handbook, which is at present incomplete, has begun to address these challenges.[11]

AHRQ has also recognized the limitations of the *General Methods Guide* when applied to studies of medical tests. In 2007, AHRQ convened an expert working meeting on the methodologic challenges in performing systematic reviews of medical tests. Four white papers were commissioned and presented on May 28-29, 2008.[12] The discussions from this meeting formed the basis for the Medical Test EPC workgroups, led by DM, then director of the Center for Clinical Health Policy Research and of the Duke EPC. Three EPC workgroups identified and addressed practical challenges in each step of conducting systematic reviews of medical tests (understanding the context, performing the review, and synthesizing the evidence). From these

workgroups, EPC authors wrote nine draft papers providing guidance on steps for systematically reviewing medical test evidence that were either not covered in the existing *General Methods Guide* or that illustrated how to apply the *General Methods Guide* to medical test evaluation. An additional two workgroups addressed issues unique to genetic and prognostic tests. Each paper underwent extensive peer review by EPC investigators, external peer review, and public comment.

The Society of General Internal Medicine (SGIM) and the editorial leadership of the *Journal of General Internal Medicine* recognize that academic general internists share with AHRQ the desire to improve the quality of systematic reviews of medical tests through dissemination of methods guides to potential authors. AHRQ approached the *Journal's* editorial leadership and proposed a collaborative effort to review and publish this guide. The AHRQ Scientific Resource Center managed the peer and public review process through the usual Effective Health Care Program mechanisms. Two deputy editors from the *Journal* (GS and CU) reviewed the peer and public review comments, and author responses. All four of us reconciled any remaining issues and submitted a consensus letter to the corresponding author of each chapter with additional requests for revisions. In particular, we sought to expand the scope of the articles beyond EPC authors to provide relevant guidance to *all* authors of systematic reviews of medical tests. Likewise, we guided manuscript development so that the resulting chapters would be of value to readers of systematic reviews of medical tests who seek to determine the strengths and weaknesses of the review and its impact on clinical practice. We asked authors to identify potential differences between their chapters and the recommendations from the upcoming Cochrane handbook for systematic reviews of diagnostic test accuracy, and to comment on the basis for any disparities. The final versions of each chapter manuscript were submitted simultaneously to the *Journal* for typesetting and to AHRQ for public posting. AHRQ is also developing online training modules for authors based on the content of these manuscripts.[13]

This *Methods Guide for Medical Test Reviews*, published simultaneously on the AHRQ Web site and as a special supplement to the *Journal of General Interal Medicine*, represents the final product of these efforts. It covers twelve core aspects of the optimal conduct of systematic reviews of medical tests and serves as guidance for authors. However, each paper, or chapter, stands on its own. It is our sincere hope that EPC and non-ECP authors of systematic reviews, as well as other researchers and clinician readers, will find this collated *Methods Guide for Medical Test Reviews* to be helpful for the generation and appraisal of reviews, as well as the application of reviews to decision making about the use of specific medical tests in clinical practice.

References

1. Montori VM, Wilczynski NL, Morgan D, Haynes RB. Systematic reviews: a cross-sectional study of location and citation counts. BMC Med 2003;1:2.
2. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. Lancet 1999;354:1896-900.
3. 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:W65-94.
4. PRISMA: Transparent reporting of systematic reviews and meta-analyses. http://www.prisma-statement.org/index.htm. Accessed March 2, 2012.
5. Higgins J, Green S, (editors). Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 (updated March 2011). The Cochrane Collaboration.
6. Evidence-based Practice Centers: synthesizing scientific evidence to improve quality and effectiveness in health care. Available at http://www.ahrq.gov/clinic/epc/ Accessed March 2, 2012.
7. Agency for Healthcare Research and Quality: U.S. Preventive Services Task Force (USPSTF). http://www.ahrq.gov/clinic/uspstfix.htm.) Accessed March 9, 2012.
8. Agency for Healthcare Research and Quality. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville, MD: Agency for Healthcare Research and Quality. http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318 Accessed March 2, 2012.
9. Agency for Healthcare Research and Quality. U.S. Preventive Services Task Force Procedure Manual. AHRQ Publication No. 08-05118-EF. July 2008. http://www.uspreventiveservicestaskforce.org/uspstf08/methods/procmanual.pdf. Accessed May 24, 2012.
10. Institute of Medicine of the National Academies. Finding What Works in Health Care: Standards for Systematic Reviews; March 2011. http://www.iom.edu/Reports/2011/Finding-What-Works-in-Health-Care-Standards-for-Systematic-Reviews.aspx. Accessed March 9, 2012.
11. Cochrane Diagnostic Test Accuracy Working Group. Handbook for DTA reviews. http://srdta.cochrane.org/handbook-dta-reviews. Accessed March 9, 2012.
12. Agency for Healthcare Research and Quality. Effective Health Care Program. Medical tests. White paper series; Nov. 16, 2009. http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=350. Accessed March 9, 2012.
13. Agency for Healthcare Research and Quality. CME/CE Activities. 2010. http://www.effectivehealthcare.ahrq.gov/index.cfm/tools-and-resources/cmece-activities/#ahrq Accessed March 9, 2012.

Conflict of interest: None of the authors has affiliations or financial involvements that conflict with the information presented in this chapter.

Corresponding author: Gerald W. Smetana, M.D., Division of General Medicine and Primary Care, Beth Israel Deaconess Medical Center, Shapiro 621D, 330 Brookline Avenue, Boston, MA 02215. Phone: 617-667-9600. Fax : 617-667-9620. Email: gsmetana@bidmc.harvard.edu

Suggested citation: Smetana GW, Umscheid CA, Chang S, Matchar DB. Methods guide for authors of systematic reviews of medical tests: a collaboration between the Agency for Healthcare Research and Quality (AHRQ) and the Journal of General Internal Medicine. AHRQ Publication No. 12-EHC098-EF. Editorial printed in Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, Maryland: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.

Chapter 1

Introduction to the Methods Guide for Medical Test Reviews

David B. Matchar, M.D., FACP, FAHA; Duke-NUS Graduate Medical School Singapore; Duke University Medical Center, Durham, NC

Abstract

Evaluation of medical tests presents challenges distinct from those involved in the evaluation of therapies; in particular, the very great importance of context and the dearth of comprehensive RCTs aimed at comparing the clinical outcomes of different tests and test strategies. Available guidance provides some suggestions: (1) Use the PICOTS typology for clarifying the context relevant to the review, and (2) Use an organizing framework for classifying the types of medical test evaluation studies and their relationship to potential key questions. However, there is a diversity of recommendations for reviewers of medical tests and a proliferation of concepts, terms, and methods. As a contribution to the field, this *Methods Guide for Medical Test Reviews* seeks to provide practical guidance to achieving the goal of clarity, consistency, tractability, and usefulness.

Introduction

With the growing number, complexity, and cost of medical tests, which tests can reliably be expected to improve health outcomes, and under what circumstances? As reflected in the increasing number of requests for systematic reviews of medical tests under the Agency for Healthcare Research and Quality (AHRQ) Evidence-based Practice Center (EPC) Program, patients, clinicians, and policymakers have a profound need for guidance on this question.

Systematic reviews developed under the EPC Program (sometimes labeled "evidence reports" or "technology assessments") are expected to be technically excellent and practically useful. The challenge for EPC investigators is to complete such reviews with limited time and resources—a daunting prospect, particularly in the face of the near-exponential growth in the number of published studies related to medical tests (A MEDLINE® search using the keyword "test.mp" demonstrates a doubling of the number of citations approximately every 10 years since 1960). How can EPC investigators respond to this challenge with reviews that are timely, accessible, and practical, and that provide insight into where there have been (or should be) advances in the field of systematic review of medical tests?

This *Methods Guide for Medical Test Reviews* (referred to hereafter as the *Medical Test Methods Guide*), produced by researchers in AHRQ's EPC Program, is intended to be a practical guide for those who prepare and use systematic reviews of medical tests; as such, it complements AHRQ's *Methods Guide for Effectiveness and Comparative Effectiveness* Reviews[1] (hereafter referred to as the *General Methods Guide*).[1] Not only has the *Medical Test Methods Guide* been motivated by the increasing need for comprehensive reviews of medical tests; it has also been

created in recognition of features of medical tests and the evaluation literature that present unique problems for systematic reviewers. In particular, medical tests are used in—and are highly dependent on—a complex context. This context includes, among other factors, preexisting conditions, results of other tests, skill and knowledge of providers, availability of therapeutic resources, and so on. In this complex environment, researchers have tended to focus on narrow questions, such as the ability of a test to conform to technical specifications, to accurately classify patients into diagnostic or prognostic categories, or to influence thought or actions by clinicians and patients. Rarely are medical tests evaluated in randomized controlled trials with representative patient populations and comprehensive measures of patient-relevant outcomes. As a result, the reviewer must put together the evidence in puzzle-like fashion.

In addition to encouraging a high standard for excellence, usefulness, and efficiency in systematic reviews, this *Medical Test Methods Guide* is designed to promote consistency in how specific issues are addressed across the various systematic reviews produced by investigators. Even though consistency in approach may not always guarantee that a particular task in review development is done in an ideal way, it is certainly the case that inconsistency in approach increases the effort and energy needed to read, digest, and apply the results of systematic reviews of medical tests.

Development of the *Medical Test Methods Guide*

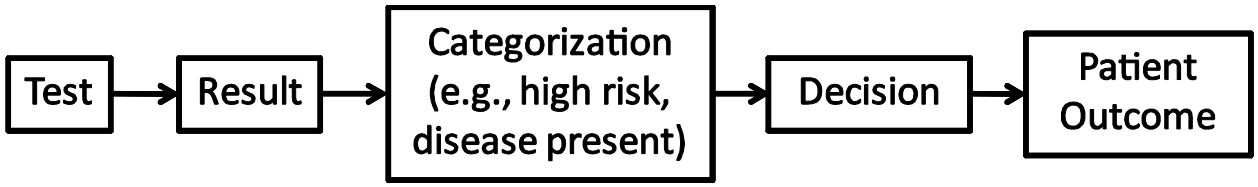
In developing this *Medical Test Methods Guide*, we sought to apply theory and empirical evidence, supplemented by personal experience and judgment, and to maintain consistency as much as possible with the principles described in AHRQ's *General Methods Guide*. We were guided by two fundamental tenets: (1) Evaluation of the value of a medical test must always be linked to the context of use; and (2) systematic reviews of medical test studies are ultimately aimed at informing the use of those tests to improve the health outcomes of patients, in part by guiding clinicians to make rational decisions and judgments.

The first tenet stands in contradiction to the common assumption that medical test results are neutral reporters of reality, independent of context. The notion that tests are "signal detectors" with invariant performance characteristics (i.e., sensitivity and specificity), likely reflects the way that the Bayes rule has been introduced to the medical community—as a pedagogical tool for transmitting the insight that a test for a condition must be interpreted in light of the likelihood of the condition before the test was performed (prior probability). Such teaching assumes that the performance characteristics of a medical test (like those of electronic receivers and similar devices) are constant over all relevant situations. There are clearly circumstances where this is true enough for practical purposes. However, the possibility that it may not be true across all relevant applications highlights the importance of context, which can affect not only sensitivity and specificity but also the clinical implications of a particular test result. Thus, throughout this document the authors return to the theme of clarifying the context in which the test under evaluation is to be used.

The second tenet is that medical tests (and therefore assessments of those tests) are about improving patient outcomes, often by guiding clinicians' judgments. Unfortunately, the vast majority of published literature on medical tests does not address the clinical impact of tests, focusing instead on test development and test performance characteristics. Indeed, test performance characteristics have been treated as sufficient criteria of test value (i.e., if the performance characteristics are good, then the test should be promoted). However, performance characteristics may not in fact be sufficient: a test with sensitivity and specificity in the high 90-

percent range may not improve the likelihood of a good patient outcome if the prevalence of the underlying condition or risk is low, or if the treatment options are of marginal efficacy or high risk. This *Medical Test Methods Guide* promotes the centrality of patient outcomes by recommending that one of the first steps in a review must be to establish a link between the use of a test and the outcomes patients and clinicians care about. This link can also be expounded through the use of visual representations such as the causal chain diagram, illustrated in a simplified form in Figure 1–1.

Figure 1–1. Causal chain diagram



In rare but ideal cases, a test is evaluated in a comprehensive clinical trial in which every relevant outcome is assessed in a representative group of patients in typical practice settings. More often, however, a systematic review may appropriately focus on only one link in this chain, as when the test is being compared with an established test known to improve outcomes. Ideally, the entire chain should be considered and evidence regarding each link assembled, evaluated, and synthesized.

Unique Challenges of Medical Tests

Of the many tools available to clinicians caring for patients, medical tests are among the most commonly employed. (Note that here "medical tests" is used as an umbrella term, to denote any test used in a health care context, irrespective of type—e.g., chemistry, genetic, radiological—or role—e.g., screening, diagnosis, or prognosis.) Tests can be used to screen for the likelihood of a disorder currently or in the future, or to diagnose the actual presence of disease. Medical tests may also be used to assess immediate or future response to treatment, including the probability of desirable or undesirable consequences. While medical tests are often thought of as something performed in the laboratory or radiology suite, the term also encompasses the traditional patient history and physical examination, as well as scored questionnaires intended, for example, for screening or to assess likely prognosis or response to therapy.

Assessing the impact of a treatment is generally more straightforward than assessing the impact of a medical test. This is the case primarily because most treatments lead directly to the intended result (or to adverse effects), whereas there may be several steps between the performance of a test and the outcome of clinical importance.[2] One consequence of this indirect relationship is that medical tests tend to be evaluated in isolation, in terms of their ability to discern an analyte or a particular anatomic condition, rather than in terms of their impact on overall health outcomes.[3]

In light of these challenges, the question we address directly in this *Medical Test Methods Guide* is: "How do we evaluate medical tests in a way that is clear (i.e., involves a process that can be reproduced), consistent (i.e., similar across reports), tractable (i.e., capable of being performed within resource constraints), and useful (i.e., addresses the information needs of the report recipients)?"

To answer this question, we might refer to the literature on evaluation of therapies. Arguably, the most robust empirical demonstration of the utility of a medical test is a properly designed randomized controlled trial (RCT)[4–7] that compares patient management outcomes of the test to the outcomes of one or more alternative strategies. In practice, such trials are not routinely performed because they are often deemed unattainable.

Recurrent Themes in the Test Evaluation Literature

In recognition of the unique challenges to evaluation presented by medical tests, a body of test evaluation literature has emerged over the past six decades. Two recurrent themes emerge from this literature. The first is the recognition that a medical test used to discriminate between the presence or absence of a specific clinical condition can be likened to an electronic signal detector.[8–10] This has opened the way to applying signal detection theory, including the notions of sensitivity, specificity, and the application of the Bayes rule, to calculate disease probabilities for positive or negative test results.[8–10]

The second theme reflected in the historical record is that medical test evaluation studies tend to fall along a continuum related to the breadth of the study objectives—from assessing a test's ability to conform to technical specifications, to the test's ability to accurately classify patients into disease states or prognostic levels, to the impact of the test on thought, action, or outcome. Various frameworks have been developed to describe the different outcomes of the study. Table 1–1 below consolidates these terms, with relevant examples, into four basic categories. Further descriptions of the various frameworks are included in the following sections.

Table 1–1. Different objectives of medical test evaluation studies

Study Objective	Terms Used	Examples
Ability of a test to conform to technical specifications	Technical efficacy	Technical quality of a radiological image
	Analytic validity	Accuracy of a chemical assay for the target analyte Concordance of a commercial genetic test with the true genotype
Ability of a test to classify a patient into a disease/phenotype or prognosis category	Diagnostic accuracy efficacy Clinical validity Test accuracy Test performance Performance characteristics Operating characteristics	Sensitivity and specificity Positive and negative likelihood ratios Positive and negative predictive value Test yield Receiver operating characteristic (ROC) curve
Ability of a test to direct clinical management and improve patient outcomes	Diagnostic thinking efficacy Therapeutic efficacy Patient outcome efficacy Clinical utility	Impact on mortality or morbidity Impact on clinician judgment about diagnosis/prognosis Impact on choice of managment
Ability of a test to benefit society as a whole	Societal efficacy	Incremental cost-effectiveness

Analytic Frameworks

While the preceding discussion provides a way to classify test evaluation studies according to their objectives, it does not offer the reviewer an explicit strategy for summarizing an often complex literature in a logical way in order to respond to key questions. In 1988, Battista and Fletcher applied "causal pathways" for the United States Preventive Services Task Force (USPSTF) in the study of evaluating preventive services, as a test for understanding and

evaluating the strength of support for the use of a preventive measure.[11] Such a framework is useful in maintaining an orderly process, clarifying questions, and organizing evidence into relevant categories. This value has been reiterated in other recommendations for reviewers.[12–14] In 1991, Woolf described a conceptual model that he termed the "Evidence Model,"[15] and in 1994, he described this same model as the "analytic framework."[16]

These points were reiterated in the most recent Procedure Manual for the USPSTF:

> The purpose of analytic frameworks is to present clearly in graphical format the specific questions that need to be answered by the literature review in order to convince the USPSTF that the proposed preventive service is effective and safe (as measured by outcomes that the USPSTF considers important). The specific questions are depicted graphically by linkages that relate interventions and outcomes. These linkages serve the dual purpose of identifying questions to help structure the literature review and of providing an "evidence map" after the review for the purpose of identifying gaps and weaknesses in the evidence.[17]

Two key components of the analytic framework are: (1) a typology for describing the context in which the test is to be used, and (2) some form of visual representation of the relationship between the application of the test or treatment and the outcomes of importance for decisionmaking. Visual display of essential information for defining key questions will also explicitly define the population, intervention, comparator and outcomes, which makes analytic frameworks consistent with the current standard approach to classifying contexts, the PICOTS typology, which is further described below. (For more information on PICOTS, see Chapter 2.)

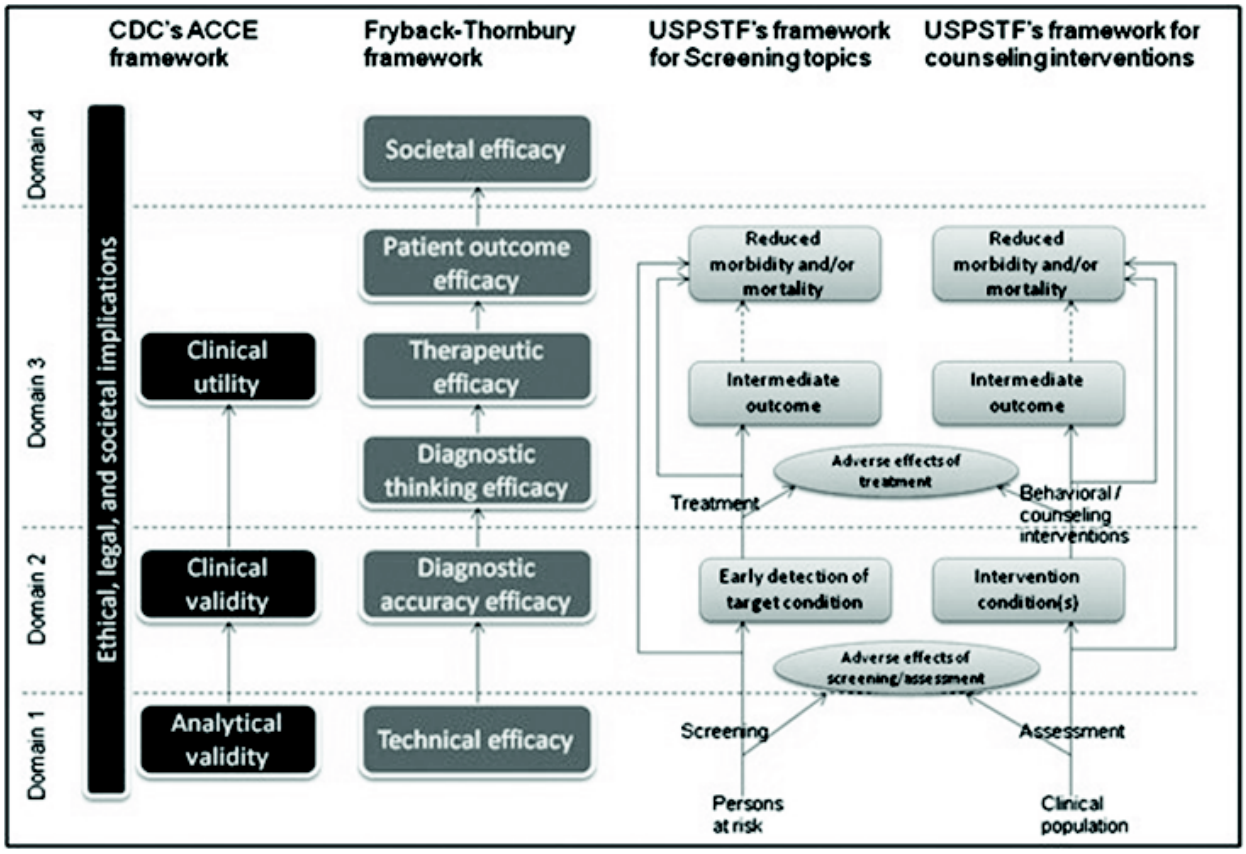
In addition to using the analytic framework in reviews to support clinical practice guidelines and the USPSTF, the AHRQ EPC Program has promoted the use of analytic frameworks in systematic reviews of effectiveness or comparative effectiveness of non-test interventions.[1] Although not specifically recommending a visual representation of the framework, the Cochrane Collaboration also organizes key questions using a similar framework.[18]

A Note on Terminology

With the evolution of the field, there has been a proliferation of terms used to describe identical or similar concepts in medical test evaluation. In this *Medical Test Methods Guide,* we have attempted to identify similar terms and to be consistent in our use of terminology. For example, throughout this document, we use terms for different categories of outcomes (Table 1–1) that are rooted in various conceptual frameworks for test evaluation (hereafter referred to as "organizing frameworks," although elsewhere referred to as "evaluative" or "evaluation" frameworks). There have been many different organizing frameworks; these have recently been systematically reviewed by Lijmer and colleagues.[5] Each framework uses slightly different terminology, yet each maps to similar concepts.

To illustrate this point, Figure 1–2 shows the relationship between three representative organizing frameworks: (1) The "ACCE" model of Analytic validity, Clinical validity, Clinical utility, and Ethical, legal and social implications,[19–20] (2) the Fryback and Thornbury model, one of the most widely used and well known of all the proposed organizing frameworks,[21] and (3) the USPSTF model for assessing screening and counseling interventions.[22] Since the key concepts are similar, unless another framework is especially apt for a particular review task, our principle of achieving consistency would argue for use of the USPSTF (See Chapter 2.)

Figure 1–2. A mapping across three major organizing frameworks for evaluating clinical tests



Notes: Used with permission of the ECRI Institute. The ECRI institute created this figure based on the specified evaluation frameworks. For a detailed description of each included framework, the reader is referred to the original references.[16–19] Domain 1—analytical validity; Domain 2—clinical validity; Domain 3—clinical utility; Domain 4—ethical, legal and societal implications.

PICOTS Typology

A typology that has proven extremely useful for the evaluation of therapies, and which also applies to the evaluation of medical tests, is called PICOTS. This typology—Patient population, Intervention, Comparator, Outcomes, Timing, Setting—is a tool established by systematic reviewers to describe the context in which medical interventions might be used, and is thus important for defining the key questions of a review and assessing whether a given study is applicable or not.[23]

The EPC Program, reflecting the systematic review community as a whole, occasionally uses variations of the PICOTS typology (Table 1–2). The standard, unchanging elements are the PICO, referring to the Patient population, Intervention, Comparator, and Outcomes. Timing refers to the timing of outcome assessment and thus may be incorporated as part of Outcomes or as part of Intervention. Setting may be incorporated as part of Population or Intervention, but it is often specified separately because it is easy to describe. For medical tests, the setting of the test has particular implications for bias and applicability in light of the spectrum effect. Occasionally, "S" may be used to refer to Study design. Other variations, not used in the present document, include a "D" that may refer to Duration (which is equivalent to Timing) or to study Design.

Table 1–2. The PICOTS typology as applied to interventions and medical tests.

Element	As Applied to Interventions	As Applied to Medical Tests	Comment
P	Patient population	Patient population; includes results of other/prior tests	Condition(s), disease severity and stage, comorbidities, patient demographics
I	Intervention	Index test; includes clinical role of index strategy in relation to comparator, and test-and-treat strategy in relation to clinical outcomes	Description of index test; includes administrator training, technology specifications, specific application issues Three main clinical roles in relation to comparator: replacement, add-on, triage Desciption of index test performance and interpretation; how results of index test lead to management decisions/actions
C	Comparator	Comparator test-and-treat strategy	Desciption of comparator test performance and interpretation; how results of comparator test lead to management decisions/actions
O	Outcomes	Relevant clinical outcomes; includes any intermediate outcomes of interest	Patient health outcomes; includes morbidity (including adverse effects of test and treatment), mortality, quality of life; intermediate outcomes: includes technical specifications, accuracy, decisional, therapeutic impact
T	Timing	Timing of outcome assessment	Duration of followup; single or multiple followup assessments
S	Setting	Setting of test assessment	Ambulatory settings (including primary, specialty care) and inpatient settings

Organization of This *Medical Test Methods Guide*

As noted above, this *Medical Test Methods Guide* complements AHRQ's *General Methods Guide*,[1] which focuses on methods to assess the effectiveness of treatments and other non-test interventions. The present document applies the principles used in the *General Methods Guide* to the specific issues and challenges of assessing medical tests, and highlights particular areas where the inherently different qualities of medical tests necessitate a variation of the approach used for a systematic review of treatments. We provide guidance in stepwise fashion for those conducting a systematic review.

Chapters 2 and 3 consider the tasks of developing the topic, structuring the review, developing the key questions, and defining the range of decision-relevant effects. Developing the topic and structuring the review—often termed "scoping"—are fundamental to the success of a report that assesses a medical test. Success in this context means not only that the report is deemed by the sponsor to be responsive but also that it is actually used to promote better quality care. In this *Medical Test Methods Guide,* we introduce various frameworks to help determine and organize the questions. While there is not a specific section on developing inclusion and exclusion criteria for studies, many of the considerations at this stage are highlighted in chapters 2 and 3, which describe how to determine the key questions, as well as in chapters 5 and 6, which describe how to assess the quality and applicability of studies.

Chapters 4 through 10 highlight specific issues in conducting reviews: searching, assessing quality and applicability, grading the body of evidence, and synthesizing the evidence. Searching for medical test studies (Chapter 4) requires unique strategies, which are discussed briefly. Assessing individual study quality (chapter 5) relates primarily to the degree to which the study is internally valid; that is, whether it measures what it purports to measure, in as unbiased a fashion as possible. Although much effort has been expended to rate features of studies in a way that accurately predicts which studies are more likely to reflect "the truth," this goal has proven

elusive. In Chapter 5, we note several approaches to assessing the limitations of a study of a medical test and recommend an approach.

Assessing applicability (Chapter 6) refers to determining whether the evidence identified is relevant to the clinical context of interest. Here we suggest that systematic reviewers search the literature to assess which factors are likely to affect test effectiveness. We also suggest that reviewers complement this with a discussion with stakeholders to determine which features of a study are crucial (i.e., which must be abstracted, when possible, to determine whether the evidence is relevant to a particular key question or whether the results are applicable to a particular subgroup.)

Once systematic reviewers identify and abstract the relevant literature, they may grade the body of literature as a whole (Chapter 7). One way to conceptualize this task is to consider whether the literature is sufficient to answer the key questions such that additional studies might not be necessary or would serve only to clarify details of the test's performance or utility. In Chapter 7, we discuss the challenges and applications of grading the strength of a body of test evidence.

Chapter 8 through 10 focus on the technical approach to synthesizing evidence, in particular, meta-analysis and decision modeling. Common challenges addressed include evaluating evidence when a reference standard is available (Chapter 8), and when no appropriate reference standard exists (Chapter 9). In reviewing the application of modeling in clinical test evidence reviews, we focus in chapter 10 on evaluating the circumstances under which a formal modeling exercise may be a particularly useful component of an evidence review.

Finally, in Chapters 11 and 12, we consider special issues related to the evaluation of genetic tests and prognostic tests, respectively. While both topics are represented in earlier chapters, those chapters focus on methods for evaluating tests to determine the current presence of disease, as with screening or diagnostic tests. Chapters 11 and 12 complete the guidance by addressing special considerations of assessing genetic and prognostic tests.

Summary

Evaluation of medical tests presents challenges distinct from those involved in the evaluation of therapies; in particular, the very great importance of context and the dearth of comprehensive RCTs aimed at comparing the clinical outcomes of different tests and test strategies. Available guidance provides some suggestions: (1) Use the PICOTS typology to clarify the context relevant to the review, and (2) Use an organizing framework to classify the types of medical test evaluation studies and their relationship to potential key questions. However, there is a diversity of recommendations for reviewers of medical tests and a proliferation of concepts, terms, and methods. As a contribution to the field, this *Medical Test Methods Guide* seeks to provide practical guidance to achieving the goal of clarity, consistency, tractability, and usefulness.

References

1. Agency for Healthcare Research and Quality. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville, MD: Agency for Healthcare Research and Quality; 2008–. http://www.ncbi.nlm.nih.gov/books/NBK47095. Accessed September 20, 2010.
2. Siebert U. When should decision analytic modeling be used in the economic evaluation of health care? Eur J Health Econ 2003;4(3):143-50.
3. Tatsioni A, Zarin DA, Aronson N, et al. Challenges in systematic reviews of diagnostic technologies. Ann Intern Med 2005;142(12 Pt 2):1048-55.
4. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet 2000;356:1844-7.
5. Lord SJ, Irwig L, Simes J. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need a randomized trial? Ann Intern Med 2006;144(11):850-5.
6. Lijmer JG, Leeflang M, Bossuyt PM. Proposals for a phased evaluation of medical tests. Med Decis Making 2009;29(5):E13-21.
7. Lord SJ, Irwig L, Bossuyt PM. Using the principles of randomized controlled trial design to guide test evaluation. Med Decis Making 2009;29(5):E1-E12.
8. Green DM, Swets JA. Signal detection theory and psychophysics. New York: Wiley, 1966. Reprinted with corrections and an updated topical bibliography by Peninsula Publishing, Los Altos, CA, 1988.
9. Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis. Science 1959;130:9-21.
10. Yerushalmy J. Statistical problems in assessing methods of medical diagnosis, with special reference to x-ray techniques. Public Health Rep 1947;62:1432-49.
11. Battista RN, Fletcher SW. Making recommendations on preventive practices: methodological issues. In: Battista RN, Lawrence RS, editors. Implementing Preventive Services. Suppl to Am J Prev Med 1988;4(4):53-67. New York, NY: Oxford University Press.
12. Bravata DM, McDonald KM, Shojania KG, Sundaram V, Owens DK. Challenges in systematic reviews: synthesis of topics related to the delivery, organization, and financing of health care. Ann Intern Med 2005;142(Suppl):1056-1065.
13. Mulrow CM, Langhorne P, Grimshaw J. Integrating heterogeneous pieces of evidence in systematic reviews. Ann Intern Med 1997;127(11):989-995.
14. Whitlock EP, Orleans T, Pender N, Allan J. Evaluating primary care behavioral counseling interventions: an evidence-based approach. Am J Prev Med 2002;22(4):267-284.
15. Woolf SH. Interim manual for clinical practice guideline development: a protocol for expert panels convened by the office of the forum for quality and effectiveness in health care. AHRQ Publication No. 91-0018. Rockville, MD: U.S. Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research; 1991.
16. Woolf SH. An organized analytic framework for practice guideline development: using the analytic logic as a guide for reviewing evidence, developing recommendations, and explaining the rationale. In: McCormick KA, Moore SR, Siegel RA, editors. Methodology perspectives: clinical practice guideline development. Rockville, MD: U.S. Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research; 1994. p. 105-13.
17. Agency for Healthcare Research and Quality. U.S. Preventive Services Task Force Procedure Manual. AHRQ Publication No. 08-05118-EF. Rockville, MD: Agency for Healthcare Research and Quality; July 2008. p. 22-4. Available at: http://www.uspreventiveservicestaskforce.org/uspstf08/methods/procmanual.htm. Accessed June 15, 2011.

18. O'Connor D, Green S, Higgins J. Chapter 5: Defining the review question and developing criteria for including studies. In: Higgins JPT, Green S, editors, Cochrane Handbook of Systematic Reviews of Intervention. Version 5.0.1 (updated September 2008). The Cochrane Collaboration, 2008. Available from www.cochrane-handbook.org. Accessed July 12, 2010.

19. Centers for Disease Control and Prevention (CDC) Office of Public Health Genomics. ACCE Model Process for Evaluating Genetic Tests. Available at: http://www.cdc.gov/genomics/gtesting/ACCE/index.htm. Accessed July 16, 2010.

20. National Office of Public Health Genomics. ACCE: a CDC-sponsored project carried out by the Foundation of Blood Research [Internet]. Atlanta, GA: Centers for Disease Control and Prevention (CDC); 2007 Dec 11.

21. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making 1991;11(2):88-94.

22. Harris RP, Helfand M, Woolf SH, et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med 2001;20(3 Suppl):21-35.

23. Chalmers I, Hedges LV, Cooper H. A brief history of research synthesis. Eval Health Prof 2002;25(1):12-37.

Acknowledgments: I would like to thank Research Assistant Crystal M. Riley for her help in preparing this introduction. I would also like to thank ECRI Institute for their work on carefully reviewing the historical record.

Funding: Funded by the Agency for Health Care Research and Quality (AHRQ) under the Effective Health Care Program.

Conflicts of interest: The author has no affiliations or financial involvement that conflicts with the information presented in this chapter.

Corresponding author: Dr. David B. Matchar, Health Services and Systems Research, Duke-NUS Graduate Medical School, 8 College Road, Singapore 169857. Phone: 65-6516-2584. Fax: 65-6534-8632. Email: david.matchar@duke-nus.edu.sg.

Suggested citation: Matchar DB. Introduction to the Methods Guide for Medical Test Reviews. AHRQ Publication No. 12-EHC073-EF. Chapter 1 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published as a special supplement to the Journal of General Internal Medicine, July 2012.

Chapter 2

Developing the Topic and Structuring Systematic Reviews of Medical Tests: Utility of PICOTS, Analytic Frameworks, Decision Trees, and Other Frameworks

David Samson, M.S., Blue Cross and Blue Shield Association
Karen M. Schoelles M.D., S.M., FACP,
ECRI Institute Health Technology Assessment Group

Abstract

Topic development and structuring a systematic review of diagnostic tests are complementary processes. The goals of a medical test review are: to identify and synthesize evidence to evaluate the impacts of alternative testing strategies on health outcomes and to promote informed decisionmaking. A common challenge is that the request for a review may state the claim for the test ambiguously. Due to the indirect impact of medical tests on clinical outcomes, reviewers need to identify which intermediate outcomes link a medical test to improved clinical outcomes. In this paper, we propose the use of five principles to deal with challenges: the PICOTS typology (Patient population, Intervention, Comparator, Outcomes, Timing, Setting), analytic frameworks, simple decision trees, other organizing frameworks, and rules for when diagnostic accuracy is sufficient.

Introduction

> *"[We] have the ironic situation in which important and painstakingly developed knowledge often is applied haphazardly and anecdotally. Such a situation, which is not acceptable in the basic sciences or in drug therapy, also should not be acceptable in clinical applications of diagnostic technology."*
>
> J. Sanford (Sandy) Schwartz, Institute of Medicine, 1985[1]

Developing the topic creates the foundation and structure of an effective systematic review. This process includes understanding and clarifying a claim about a test (as to how it might be of value in practice) and establishing the key questions to guide decisionmaking related to the claim. Doing so typically involves specifying the clinical context in which the test might be used. Clinical context includes patient characteristics, how a new test might fit into existing diagnostic pathways, technical details of the test, characteristics of clinicians or operators using the test, management options, and setting. Structuring the review refers to identifying the analytic strategy that will most directly achieve the goals of the review, accounting for idiosyncrasies of the data.

Topic development and structuring of the review are complementary processes. As Evidence Based Practice Centers (EPCs) develop and refine the topic, the structure of the review should become clearer. Moreover, success at this stage reduces the chance of major changes in the scope of the review and minimizes rework.

While this chapter is intended to serve as a guide for EPCs, the processes described here are relevant to other systematic reviewers and a broad spectrum of stakeholders including patients, clinicians, caretakers, researchers, funders of research, government, employers, health care payers and industry, as well as the general public. This paper highlights challenges unique to systematic reviews of medical tests. For a general discussion of these issues as they exist in all systematic reviews, we refer the reader to previously published EPC methods papers.[2,3]

Common Challenges

The ultimate goal of a medical test review is to identify and synthesize evidence that will help evaluate the impacts on health outcomes of alternative testing strategies. Two common problems can impede the achievement of this goal. One is that the request for a review may state the claim for the test ambiguously. For example, a new medical test for Alzheimer's disease may fail to specify the patients who may benefit from the test—so that the test's use ranges from a screening tool among the "worried well" without evidence of deficit, to a diagnostic test in those with frank impairment and loss of function in daily living. The request for review may not specify the range of use to be considered. Similarly, the request for a review of tests for prostate cancer may neglect to consider the role of such tests in clinical decisionmaking, such as guiding the decision to perform a biopsy.

Because of the indirect impact of medical tests on clinical outcomes, a second problem is how to identify which intermediate outcomes link a medical test to improved clinical outcomes, compared to an existing test. The scientific literature related to the claim rarely includes direct evidence, such as randomized controlled trial results, in which patients are allocated to the relevant test strategies and evaluated for downstream health outcomes. More commonly, evidence about outcomes in support of the claim relates to intermediate outcomes such as test accuracy.

Principles for Addressing the Challenges

Principle 1: Engage stakeholders using the PICOTS typology.

In approaching topic development, reviewers should engage in a direct dialogue with the primary requestors and relevant users of the review (herein denoted "stakeholders") to understand the objectives of the review in practical terms; in particular, investigators should understand the sorts of decisions that the review is likely to affect. This process of engagement serves to bring investigators and stakeholders to a shared understanding about the essential details of the tests and their relationship to existing test strategies (i.e., whether as replacement, triage, or add-on), range of potential clinical utility, and potential adverse consequences of testing.

Operationally, the objective of the review is reflected in the key questions, which are normally presented in a preliminary form at the outset of a review. Reviewers should examine the proposed key questions to ensure that they accurately reflect the needs of stakeholders and are likely to be answered given the available time and resources. This is a process of trying to

balance the importance of the topic against the feasibility of completing the review. Including a wide variety of stakeholders—such as the U.S. Food and Drug Administration (FDA), manufacturers, technical and clinical experts, and patients—can help provide additional perspectives on the claim and use of the tests. A preliminary examination of the literature can identify existing systematic reviews and clinical practice guidelines that may summarize evidence on current strategies for using the test and its potential benefits and harms.

The PICOTS typology (Patient population, Intervention, Comparator, Outcomes, Timing, Setting), defined in the Introduction to this *Medical Test Methods Guide* (Chapter 1), is a typology for defining particular contextual issues, and this formalism can be useful in focusing discussions with stakeholders. The PICOTS typology is a vital part of systematic reviews of both interventions and tests; furthermore, their transparent and explicit structure positively influences search methods, study selection, and data extraction.

It is important to recognize that the process of topic refinement is iterative and that PICOTS elements may change as the clinical context becomes clearer. Despite the best efforts of all participants, the topic may evolve even as the review is being conducted. Investigators should consider at the outset how such a situation will be addressed.[4–6]

Principle 2: Develop an analytic framework.

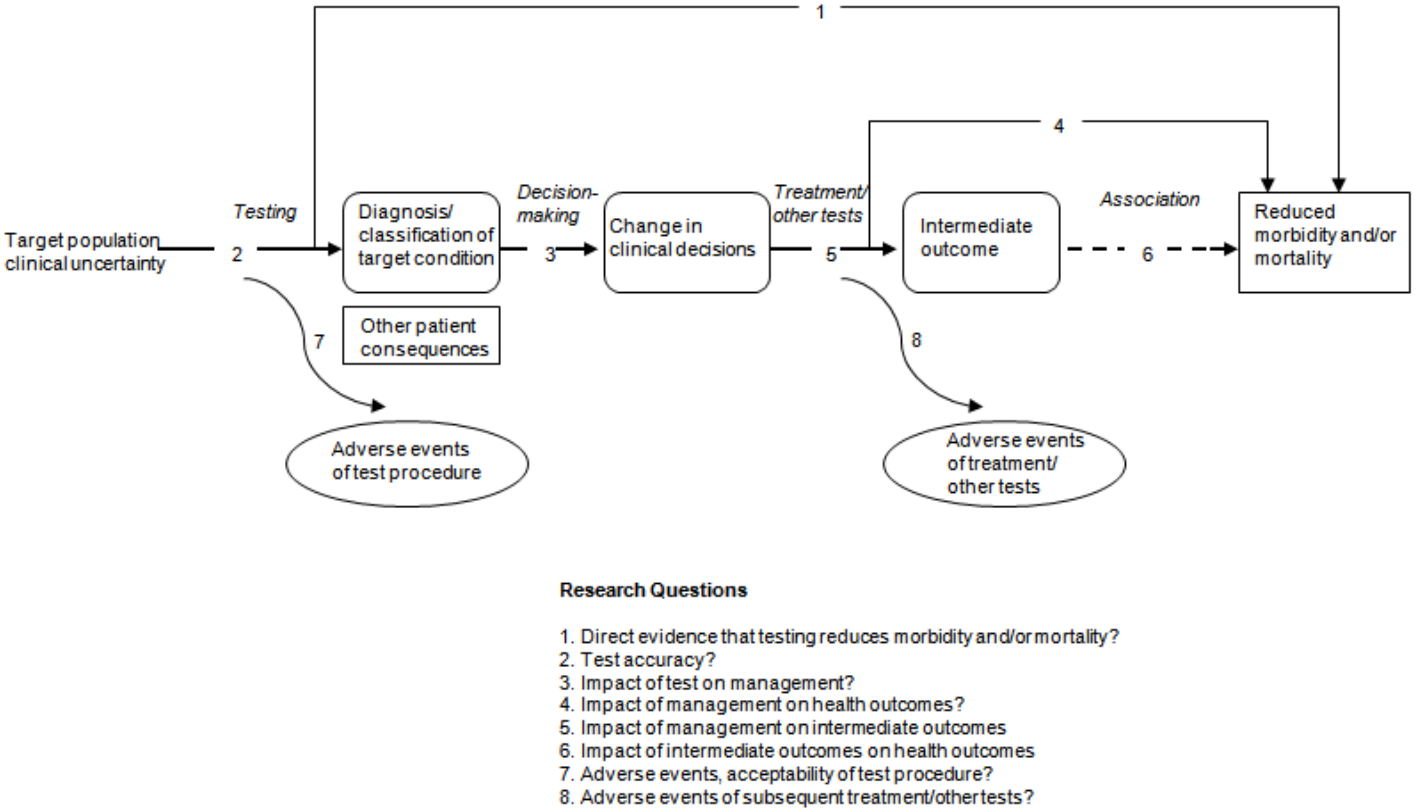
We use the term "analytic framework" (sometimes called a causal pathway) to denote a specific form of graphical representation that specifies a path from the intervention or test of interest to all-important health outcomes, through intervening steps and intermediate outcomes.[7] Among PICOTS elements, the target patient population, intervention, and clinical outcomes are specifically shown. The intervention can actually be viewed as a test-and-treat strategy as shown in links 2 through 5 of Figure 2–1. In the figure, the comparator is not shown explicitly, but is implied. Each linkage relating test, intervention, or outcome represents a potential key question and, it is hoped, a coherent body of literature.

The AHRQ EPC program has described the development and use of analytic frameworks in systematic reviews of interventions. Since the impact of tests on clinical outcomes usually depends on downstream interventions, analytic frameworks for systematic reviews of tests are particularly valuable and should be routinely included. The analytic framework is developed iteratively in consultation with stakeholders to illustrate and define the important clinical decisional dilemmas and thus serves to clarify important key questions further.[2]

However, systematic reviews of medical tests present unique challenge not encountered in reviews of therapeutic interventions. The analytic framework can help users to understand how the often convoluted linkages between intermediate and clinical outcomes fit together, and to consider whether these downstream issues may be relevant to the review. Adding specific elements to the analytic framework will reflect the understanding gained about clinical context.

Harris and colleagues have described the value of the analytic framework in assessing screening tests for the U.S. Preventive Services Task Force (USPSTF).[8] A prototypical analytic framework for medical tests as used by the USPSTF is shown in Figure 2–1. Each number in Figure 2–1 can be viewed as a separate key question that might be included in the evidence review.

Figure 2–1. Application of USPSTF analytic framework to test evaluation*



*Adapted from Harris et al., 2001[7]

In summarizing evidence, studies for each linkage might vary in strength of design, limitations of conduct, and adequacy of reporting. The linkages leading from changes in patient management decisions to health outcomes are often of particular importance. The implication here is that the value of a test usually derives from its influence on some action taken in patient management. Although this is usually the case, sometimes the information alone from a test may have value independent of any action it may prompt. For example, information about prognosis that does not necessarily trigger any actions may have a meaningful psychological impact on patients and caregivers.

Principle 3: Consider using decision trees.

An analytic framework is helpful when direct evidence is lacking, showing relevant key questions along indirect pathways between the test and important clinical outcomes. Analytic frameworks are, however, not well suited to depicting multiple alternative uses of the particular test (or its comparators) and are limited in their ability to represent the impact of test results on clinical decisions, and the specific potential outcome consequences of altered decisions. Reviewers can use simple decision trees or flow diagrams alongside the analytic framework to illustrate details of the potential impact of test results on management decisions and outcomes. Along with PICOTS specifications and analytic frameworks, these graphical tools represent systematic reviewers' understanding of the clinical context of the topic. Constructing decision trees may help to clarify key questions by identifying which indices of diagnostic accuracy and other statistics are relevant to the clinical problem and which range of possible pathways and

outcomes. (See Chapter 3, "Choosing the Important Outcomes for a Systematic Review of a Medical Test.") practically and logically flow from a test strategy. Lord et al. describe how diagrams resembling decision trees define which steps and outcomes may differ with different test strategies, and thus the important questions to ask to compare tests according to whether the new test is a replacement, a triage, or an add-on to the existing test strategy.[9]

One example of the utility of decision trees comes from a review of noninvasive tests for carotid artery disease.[10] In this review, investigators found that common metrics of sensitivity and specificity that counted both high-grade stenosis and complete occlusion as "positive" studies would not be reliable guides to actual test performance because the two results would be treated quite differently. This insight was subsequently incorporated into calculations of noninvasive carotid test performance.[10–11] Additional examples are provided in the illustrations below. For further discussion on when to consider using decision trees, see Chapter 10 in this *Medical Test Methods Guide*, "Deciding Whether To Complement a Systematic Review of Medical Tests With Decision Modeling."

Principle 4: Sometimes it is sufficient to focus exclusively on accuracy studies.

Once reviewers have diagrammed the decision tree whereby diagnostic accuracy may affect intermediate and clinical outcomes, it is possible to determine whether it is necessary to include key questions regarding outcomes beyond diagnostic accuracy. For example, diagnostic accuracy may be sufficient when the new test is as sensitive and as specific as the old test *and* the new test has advantages over the old test such as causing fewer adverse effects, being less invasive, being easier to use, providing results more quickly, or costing less. Implicit in this example is the comparability of downstream management decisions and outcomes between the test under evaluation and the comparator test. Another instance when a review may be limited to evaluation of sensitivity and specificity is when the new test is as sensitive as, but more specific than, the comparator, allowing avoidance of harms of further tests or unnecessary treatment. This situation requires the assumptions that the same cases would be detected by both tests and that treatment efficacy would be unaffected by which test was used.[12]

Particular questions to consider when reviewing analytic frameworks and decision trees to determine if diagnostic accuracy studies alone are adequate include:

1. Are the extra cases detected by the new, more sensitive test similarly responsive to treatment as are those identified by the older test?
2. Are trials available that selected patients using the new test?
3. Do trials assess whether the new test results predict response?
4. If available trials selected only patients assessed with the old test, do extra cases identified with the new test represent the same spectrum or disease subtypes as trial participants?
5. Are tests' cases subsequently confirmed by same reference standard?
6. Does the new test change the definition or spectrum of disease (e.g., by finding disease at an earlier stage)?
7. Is there heterogeneity of test accuracy and treatment effect (i.e., do accuracy and treatment effects vary sufficiently according to levels of a patient characteristic to change the comparison of the old and new test)?

When the clinical utility of an older comparator test has been established, and the first five questions can all be answered in the affirmative, then diagnostic accuracy evidence alone may be sufficient to support conclusions about a new test.

Principle 5: Other frameworks may be helpful.

Various other frameworks (generally termed "organizing frameworks," as described briefly in the Introduction to this *Medical Test Methods Guide* [Chapter 1]) relate to categorical features of medical tests and medical test studies. Lijmer and colleagues reviewed the different types of organizational frameworks and found 19 frameworks, which generally classify medical test research into 6 different domains or phases, including technical efficacy, diagnostic accuracy, diagnostic thinking efficacy, therapeutic efficacy, patient outcome, and societal aspects.[13]

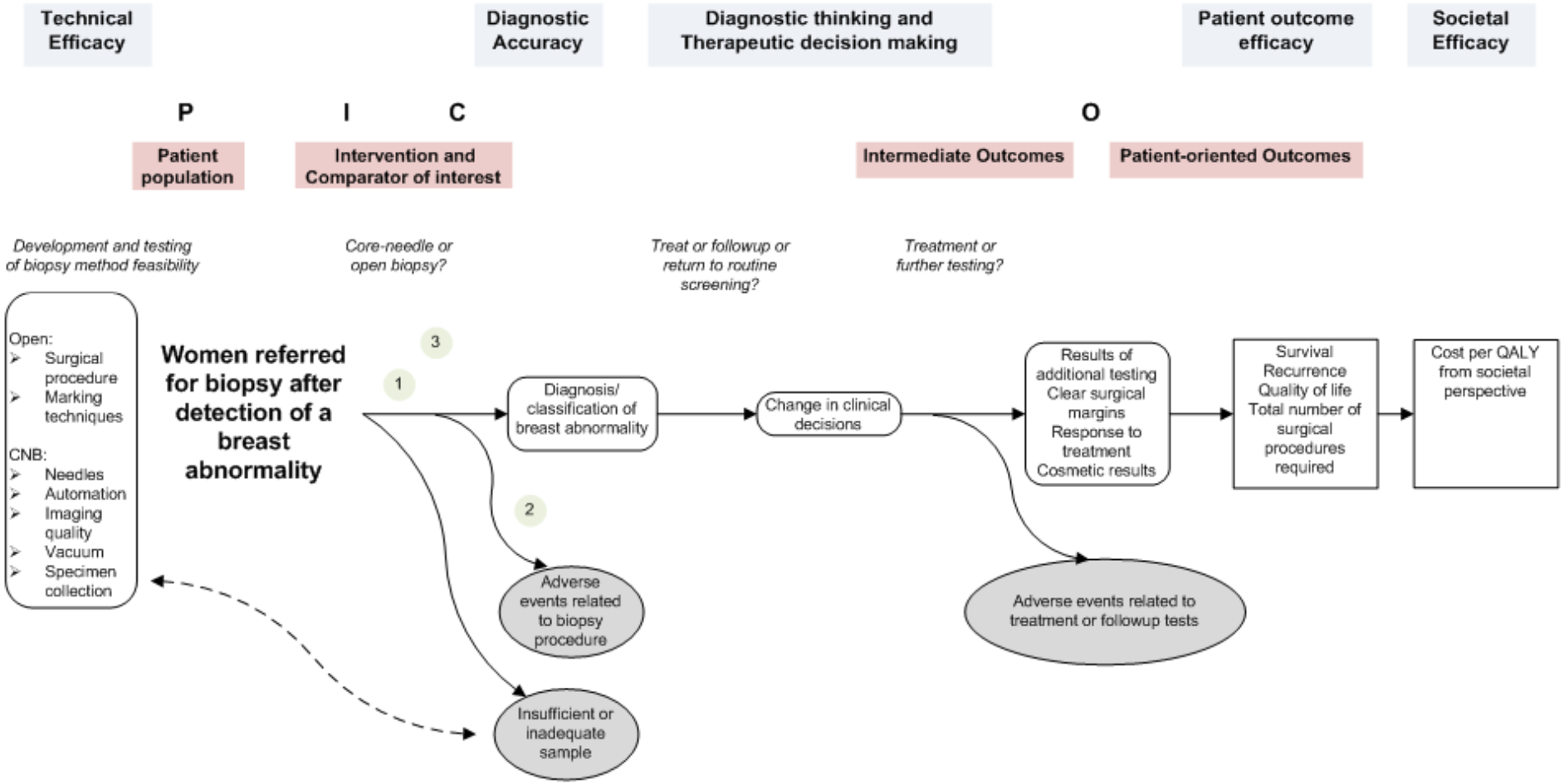
These frameworks serve a variety of purposes. Some researchers, such as Van den Bruel and colleagues, consider frameworks as a hierarchy and a model for how medical tests should be studied, with one level leading to the next (i.e., success at each level depends on success at the preceding level).[14] Others, such as Lijmer and colleagues, have argued that "The evaluation frameworks can be useful to distinguish between study types, but they cannot be seen as a necessary sequence of evaluations. The evaluation of tests is most likely not a linear but a cyclic and repetitive process."[13]

We suggest that rather than being a hierarchy of evidence, organizational frameworks should categorize key questions and suggest which types of studies would be most useful for the review. They may guide the clustering of studies, which may improve the readability of a review document. No specific framework is recommended, and indeed the categories of most organizational frameworks at least approximately line up with the analytic framework and the PICO(TS) elements as shown in Figure 2–2.

Illustrations

To illustrate the principles above, we describe three examples. In each case, the initial claim was at least somewhat ambiguous. Through the use of the PICOTS typology, the analytic framework, and simple decision trees, the systematic reviewers worked with stakeholders to clarify the objectives and analytic approach (Table 2–1). In addition to the examples described here, the AHRQ Effective Health Care Program Web site (http://effectivehealthcare.ahrq.gov/) offers free access to ongoing and completed reviews containing specific applications of the PICOTS typology and analytic frameworks.

Figure 2–2. Example of an analytical framework within an overarching conceptual framework in the evaluation of breast biopsy techniques*



*The numbers in the figure depict where the three key questions are located within the flow of the analytical framework.

Table 2–1. Examples of initially ambiguous claims that were clarified through the process of topic development

	Full-Field Digital Mammography	HER2	PET
General topic	FFDM to replace SFM in breast cancer screening (Figure 2-3)	HER2 gene amplication assay as add-on to HER2 protein expression assay (Figure 2-4)	PET as triage for breast biopsy (Figure 2-5)
Initial ambiguous claim	FFDM may be a useful alternative to SFM in screening for breast cancer.	HER2 gene amplification and protein expression assays may complement each other as means of selecting patients for targeted therapy.	PET may play an adjunctive role to breast examination and mammography in detecting breast cancer and selecting patients for biopsy.
Key concerns suggested by PICOTS, analytic framework, and decision tree	Key statistics: sensitivity, diagnostic yield, recall rate; similar types of management decisions and outcomes for index and comparator test-and-treat strategies	Key statistics: proportion of individuals with intermediate/ equivocal HER2 protein expression results who have HER2 gene amplification; key outcomes are related to effectiveness of HER2-targeted therapy in this subgroup.	Key statistics: negative predictive value; key outcomes to be contrasted were benefits of avoiding biopsy versus harms of delaying initiation of treatment for undetected tumors.
Refined claim	In screening for breast cancer, interpretation of FFDM and SFM would be similar, leading to similar management decisions and outcomes; FFDM may have a similar recall rate and diagnostic yield at least as high as SFM; FFDM images may be more expensive, but easier to manipulate and store .	Among individuals with localized breast cancer, some may have equivocal results for HER2 protein overexpression but have positive HER2 gene amplification, identifying them as patients who may benefit from HER2-targeted therapy but otherwise would have been missed.	Among patients with a palpable breast mass or suspicious mammogram, if FDG PET is performed before biopsy, those with negative scans may avoid the adverse events of biopsy with potentially negligible risk of delayed treatment for undetected tumor.
Reference	Blue Cross and Blue Shield Association Technology Evaluation Center, 2002[15]	Seidenfeld et al., 2008[16]	Samson et al., 2002[17]

FDG = fluorodeoxyglucose; FFDM = full-field digital mammography; HER2 = human epidermal growth factor receptor 2; PET = positron emission tomography; PICOTS = Patient population, Intervention, Comparator, Outcomes, Timing, Setting; SFM = screen-film mammography

The first example concerns full-field digital mammography (FFDM) as a replacement for screen-film mammography (SFM) in screening for breast cancer; the review was conducted by the Blue Cross and Blue Shield Association Technology Evaluation Center.[15] Specifying PICOTS elements and constructing an analytic framework were straightforward, with the latter resembling Figure 2–2 in form. In addition, with stakeholder input a simple decision tree was drawn (Figure 2–3) which revealed that the management decisions for both screening strategies were similar, and that therefore downstream treatment outcomes were not a critical issue. The decision tree also showed that the key indices of test performance were sensitivity, diagnostic yield, and recall rate. These insights were useful as the project moved to abstracting and synthesizing the evidence, which focused on accuracy and recall rates. In this example, the reviewers concluded that FFDM and SFM had comparable accuracy and led to comparable outcomes; that, however, storing and manipulating images was much easier for FFDM than for SFM.

Figure 2–3. Replacement test example: full-field digital mammography versus screen-film mammography*

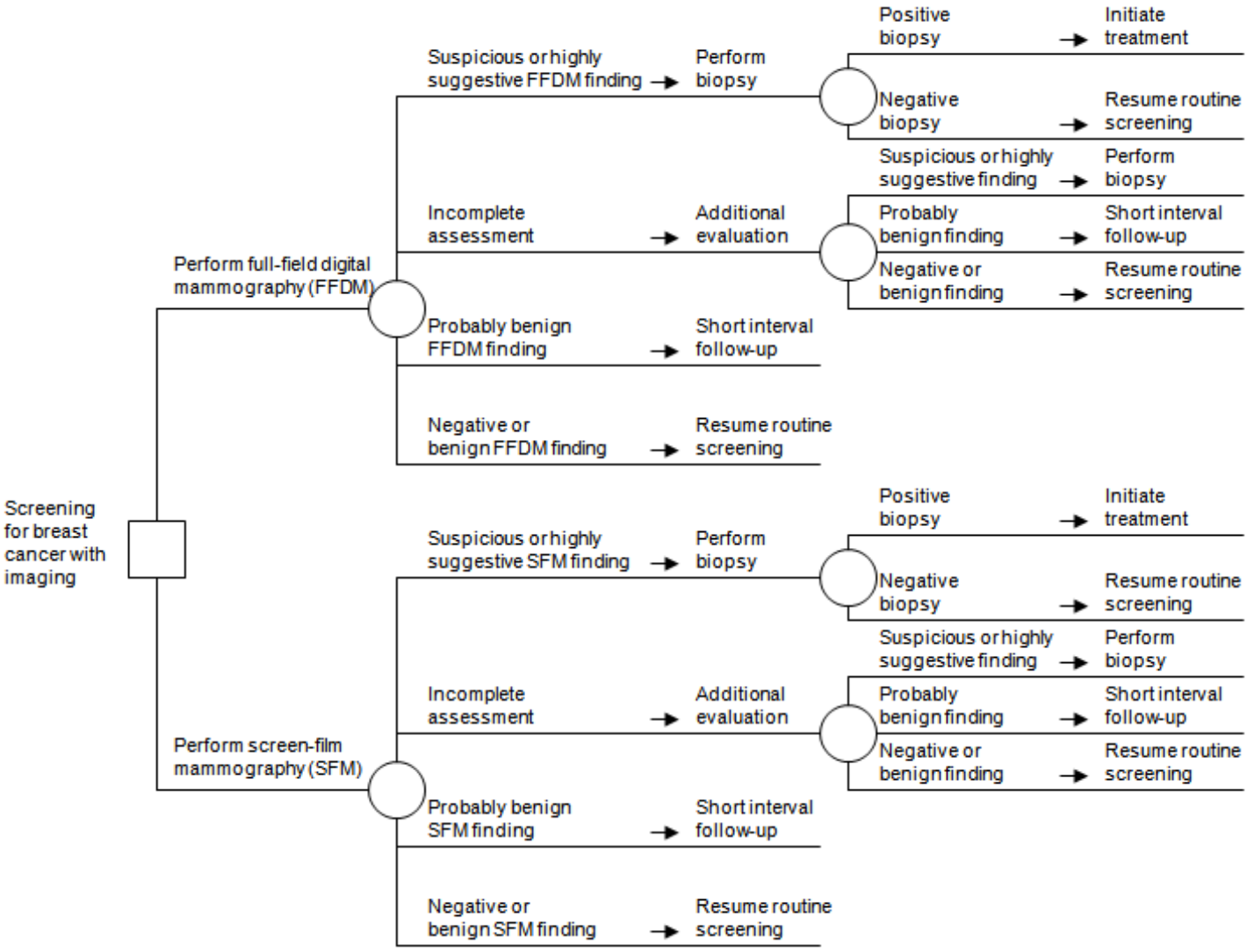
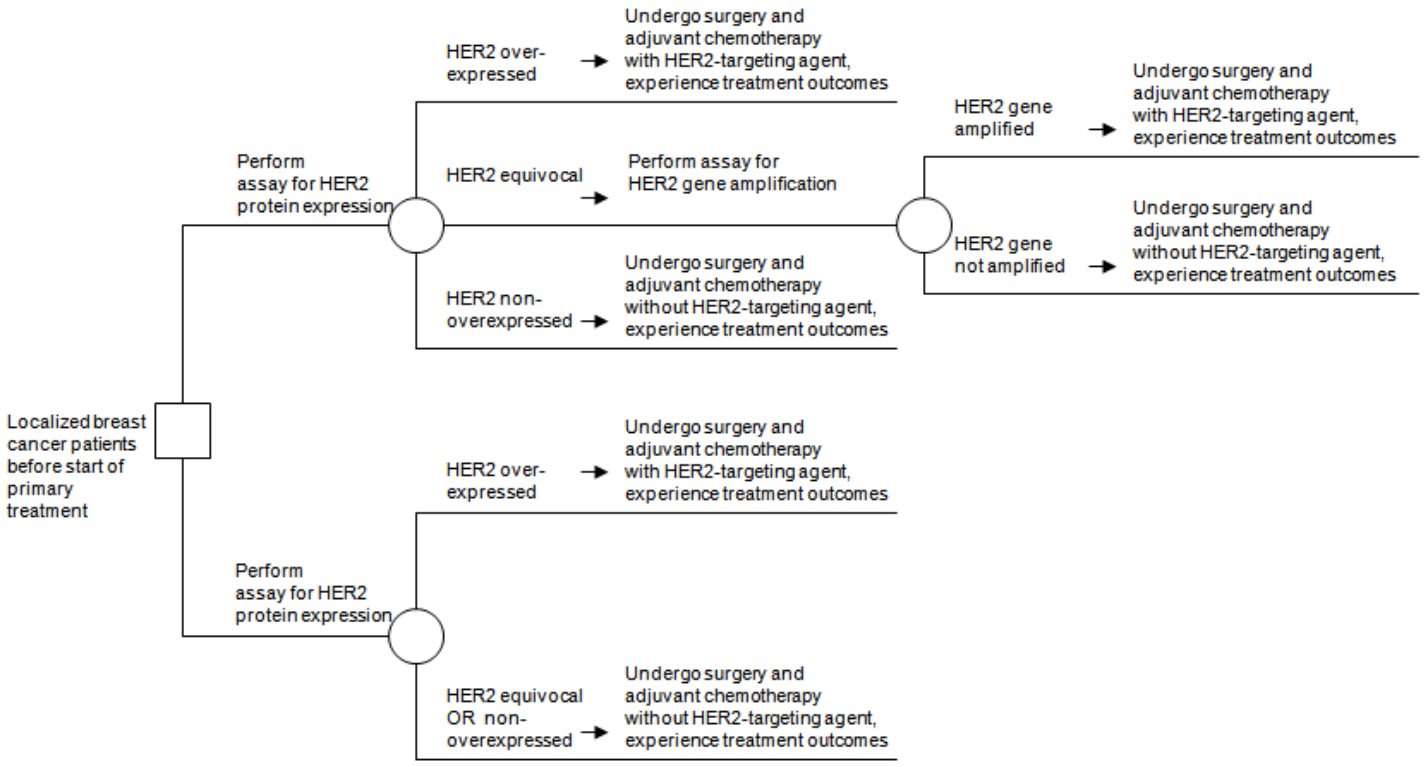


*Figure taken from Blue Cross and Blue Shield Association Technology Evaluation Center, 2002.[14]

The second example concerns use of the human epidermal growth factor receptor 2 (HER2) gene amplification assay after the HER2 protein expression assay to select patients for HER2-targeting agents as part of adjuvant therapy among patients with localized breast cancer.[16] The HER2 gene amplification assay has been promoted as an add-on to the HER2 protein expression assay. Specifically, individuals with equivocal HER2 protein expression would be tested for amplified HER2 gene levels; in addition to those with increased HER2 protein expression, patients with elevated levels by amplification assay would also receive adjuvant chemotherapy that includes HER2-targeting agents. Again, PICOTS and an analytic framework were developed, establishing the basic key questions. In addition, the authors constructed a decision tree (Figure 2–4) that made it clear that the treatment outcomes affected by HER2 protein and gene assays were at least as important as the test accuracy. While in the first case the reference standard was actual diagnosis by biopsy, here the reference standard is the amplification assay

itself. The decision tree identified the key accuracy index as the proportion of individuals with equivocal HER2 protein expression results who have positive amplified HER2 gene assay results. The tree exercise also indicated that one key question must be whether HER2-targeted therapy is effective for patients who had equivocal results on the protein assay but were subsequently found to have positive amplified HER2 gene assay results.

Figure 2–4. Add-on test example: HER2 protein expression assay followed by HER2 gene amplification assay to select patients for HER2-targeted therapy*

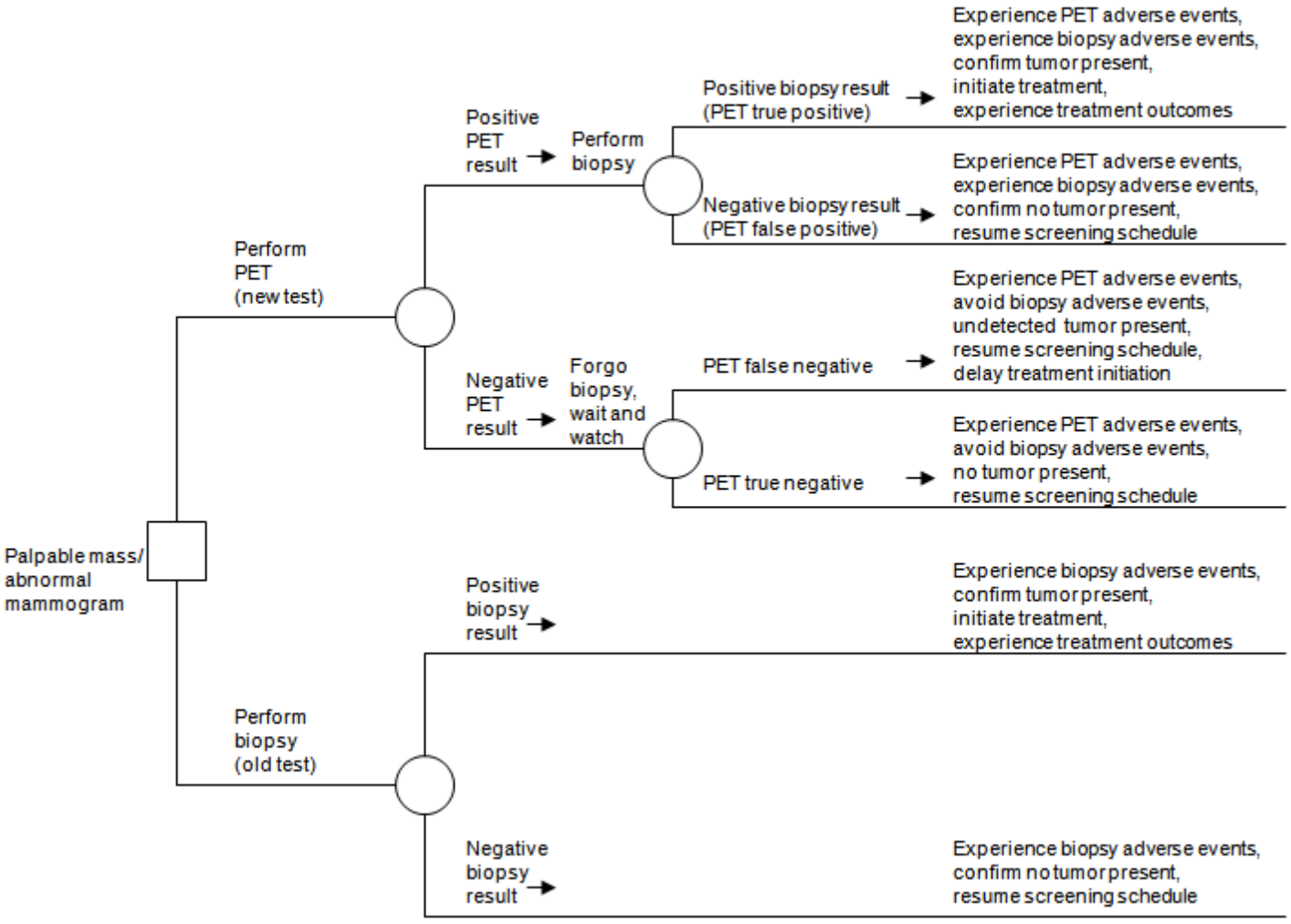


HER2 = human epidermal growth factor receptor 2
*Figure taken from Seidenfeld et al., 2008.[15]

The third example concerns use of fluorodeoxyglucose positron emission tomography (FDG PET) as a guide to the decision to perform a breast biopsy on a patient with either a palpable mass or an abnormal mammogram.[17] Only patients with a positive PET scan would be referred for biopsy. Table 2–1 shows the initial ambiguous claim, lacking PICOTS specifications such as the way in which testing would be done. The analytic framework was of limited value, as several possible relevant testing strategies were not represented explicitly in the framework. The authors constructed a decision tree (Figure 2–5). The testing strategy in the lower portion of the decision tree entails performing biopsy in all patients, while the triage strategy uses a positive PET finding to rule in a biopsy and a negative PET finding to rule out a biopsy. The decision tree illustrates that the key accuracy index is negative predictive value: the proportion of negative PET results that are truly negative. The tree also reveals that the key contrast in outcomes involves any harms of delaying treatment for undetected cancer when PET is falsely negative versus the benefits of safely avoiding adverse effects of the biopsy when PET is truly negative. The authors concluded that there is no net beneficial impact on outcomes when PET is used as a

triage test to select patients for biopsy among those with a palpable breast mass or suspicious mammogram. Thus, estimates of negative predictive values suggest that there is an unfavorable trade-off between avoiding the adverse effects of biopsy and delaying treatment of an undetected cancer.

Figure 2–5. Triage test example: positron emission tomography (PET) to decide whether to perform breast biopsy among patients with a palpable mass or abnormal mammogram*



PET = positron emission tomography
*Figure taken from Samson et al., 2002.[17]

This case illustrates when a more formal decision analysis may be useful, specifically when a new test has higher sensitivity but lower specificity than the old test, or vice versa. Such a situation entails tradeoffs in relative frequencies of true positives, false negatives, false positives, and true negatives, which decision analysis may help to quantify.

Summary

The immediate goal of a systematic review of a medical test is to determine the health impacts of use of the test in a particular context or set of contexts relative to one or more alternative strategies. The ultimate goal is to produce a review that promotes informed decisionmaking.

Key points are:

- Reaching the above-stated goals requires an interactive and iterative process of topic development and refinement aimed at understanding and clarifying the claim for a test. This work should be done in conjunction with the principal users of the review, experts, and other stakeholders.
- The PICOTS typology, analytic framework, simple decision trees, and other organizing frameworks are all tools that can minimize ambiguity, help identify where review resources should be focused, and guide the presentation of results.
- Sometimes it is sufficient to focus only on accuracy studies. For example, diagnostic accuracy may be sufficient when the new test is as sensitive and specific as the old test *and* the new test has advantages over the old test such as having fewer adverse effects, being less invasive, being easier to use, providing results more quickly or costing less.

References

1. Institute of Medicine, Division of Health Sciences Policy, Division of Health Promotion and Disease Prevention, Committee for Evaluating Medical Technologies in Clinical Use. Assessing medical technologies. Washington, DC: National Academy Press; 1985. Chapter 3: Methods of technology assessment. pp. 80-90.

2 Helfand M and Balshem H. AHRQ Series Paper 2: Principles for developing guidance: AHRQ and the effective health-care program. J Clin Epidemiol 2010;63(5):484-90.

3. Whitlock EP, Lopez SA, Chang S, et al. AHRQ Series Paper 3: Identifying, selecting, and refining topics for comparative effectiveness systematic reviews: AHRQ and the Effective Health-Care program.J Clin Epidemiol 2010;63(5):491-501.

4. Matchar DB, Patwardhan M, Sarria-Santamera A, et al. Developing a Methodology for Establishing a Statement of Work for a Policy-Relevant Technical Analysis. Technical Review 11. (Prepared by the Duke Evidence-based Practice Center under Contract No. 290-02-0025.) AHRQ Publication No. 06-0026. Rockville, MD: Agency for Healthcare Research and Quality. January 2006. Available at: http://www.ahrq.gov/downloads/pub/evidence/pdf/statework/statework.pdf. Accessed January 10, 2012.

5. Sarria-Santamera A, Matchar DB, Westermann-Clark EV, et al. Evidence-based practice center network and health technology assessment in the United States: bridging the cultural gap. Int J Technol Assess Health Care 2006;22(1):33-8.

6. Patwardhan MB, Sarria-Santamera A, Matchar DB, et al. Improving the process of developing technical reports for health care decision makers: using the theory of constraints in the evidence-based practice centers. Int J Technol Assess Health Care 2006;22(1):26-32.

7. Woolf SH. An organized analytic framework for practice guideline development: using the analytic logic as a guide for reviewing evidence, developing recommendations, and explaining the rationale. In: McCormick KA, Moore SR, Siegel RA, editors. Methodology perspectives: clinical practice guideline development. Rockville, MD: U.S. Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research; 1994. p. 105-13.

8. Harris RP, Helfand M, Woolf SH, et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med 2001;20(3 Suppl):21-35.

9. Lord SJ, Irwig L, Bossuyt PM. Using the principles of randomized controlled trial design to guide test evaluation. Med Decis Making 2009;29(5):E1-E12. Epub2009 Sep 22.

10. Feussner JR, Matchar DB. When and how to study the carotid arteries. Ann Intern Med 1988;109(10):805-18.

11. Blakeley DD, Oddone EZ, Hasselblad V, et al. Noninvasive carotid artery testing. A meta-analytic review. Ann Intern Med 1995;122(5):360-7.

12. Lord SJ, Irwig L, Simes J. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need a randomized trial? Ann Intern Med 2006;144(11):850-5.

13. Lijmer JG, Leeflang M, Bossuyt PM. Proposals for a phased evaluation of medical tests. Med Decis Making 2009;29(5):E13-21.

14. Van den Bruel A, Cleemput I, Aertgeerts B, et al. The evaluation of diagnostic tests: evidence on technical and diagnostic accuracy, impact on patient outcome and cost-effectiveness is needed. J Clin Epidemiol 2007;60(11):1116-22.

15. Blue Cross and Blue Shield Association Technology Evaluation Center (BCBSA TEC). Full-field digital mammography. Volume 17, Number 7, July 2002.

16. Seidenfeld J, Samson DJ, Rothenberg BM, et al. HER2 Testing to Manage Patients With Breast Cancer or Other Solid Tumors. Evidence Report/Technology Assessment No. 172. (Prepared by Blue Cross and Blue Shield Association Technology Evaluation Center Evidence-based Practice Center, under Contract No. 290-02-0026.) AHRQ Publication No. 09-E001. Rockville, MD: Agency for Healthcare Research and Quality. November 2008. Available at: www.ahrq.gov/downloads/pub/evidence/pdf/her2/her2.pdf. Accessed January 10, 2012.

17. Samson DJ, Flamm CR, Pisano ED, et al. Should FDG PET be used to decide whether a patient with an abnormal mammogram or breast finding at physical examination should undergo biopsy? Acad Radiol 2002;9(7):773-83.

Acknowledgements: We wish to thank David Matchar and Stephanie Chang for their valuable contributions.

Funding: Funded by the Agency for Health Care Research and Quality (AHRQ) under the Effective Health Care Program.

Disclaimer: The findings and conclusions expressed here are those of the authors and do not necessarily represent the views of AHRQ. Therefore, no statement should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

Accessibility: Persons using assistive technology may not be able to fully access information in this report. For assistance contact EffectiveHealthCare@ahrq.hhs.gov.

Conflicts of interest: None of the authors has any affiliations or involvement that conflict with the information in this chapter.

Corresponding author: David Samson, M.S., Director, Comparative Effectiveness Research, Technology Evaluation Center, Blue Cross and Blue Shield Association, 1310 G Street, NW, Washington, DC 20005. Telephone 202-626-4835 (voice); Fax 845-462-4786; email david.samson@bcbsa.com

Suggested citation: Samson D, Schoelles KM. Developing the topic and structuring systematic reviews of medical tests: utility of PICOTS, analytic frameworks, decision trees, and other frameworks. AHRQ Publication No. 12-EHC073-EF. Chapter 2 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.

Chapter 3

Choosing the Important Outcomes for a Systematic Review of a Medical Test

Jodi B. Segal, M.D., M.P.H., Johns Hopkins University School of Medicine

Abstract

In this chapter of the Evidence-based Practice Centers *Methods Guide for Medical Test Reviews*, we describe how the decision to use a medical test generates a broad range of outcomes, and suggest how each of these outcomes might be considered for inclusion in a systematic review. Awareness of these varied outcomes affects how a decisionmaker balances the benefits and risks of the test; therefore, a systematic review should present the evidence on their diversity. The key outcome categories include clinical management outcomes; direct health effects; emotional, social, cognitive, and behavioral responses to testing; legal and ethical outcomes; and costs. We describe the challenges of incorporating these outcomes in a systematic review, suggest a framework for generating potential outcomes for inclusion, and describe the role of stakeholders in choosing the outcomes for study. Finally, we give examples of systematic reviews that either included a range of outcomes or that might have done so. This chapter puts forward a set of key messages for systematic reviewers:

- Consider both the outcomes that are relevant to the process of testing and those relevant to the results of the test.
- Consider inclusion of outcomes in all five domains: clinical management effects; direct test effects; emotional, social, cognitive, and behavioral effects; legal and ethical effects; and costs.
- Consider to which group the outcomes of testing are most relevant.
- Given resource limitations, prioritize which outcomes to include. This decision depends on the needs of the stakeholder(s), who should be assisted in prioritizing the outcomes for inclusion.

Introduction

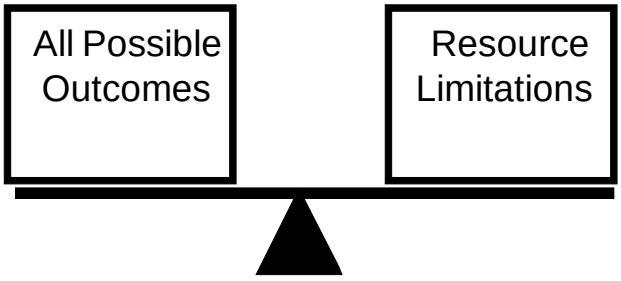
The Agency for Healthcare Research and Quality (AHRQ) requested production of a methods guide for comparative effectiveness reviews that specifically addresses the unique challenges of preparing a systematic review about the use of a medical test. This chapter describes the considerations to be taken into account when selecting the outcomes that will be included in a systematic review of a medical test. We describe the range of effects that medical tests have, and suggest how these outcomes from testing should be incorporated into a systematic review to make it maximally useful to those using the review.

We define “decision-relevant” outcomes as the outcomes that result from a testing encounter that may affect the decision to use the test. We consider a broad range of outcomes to illustrate how these may affect the balance of the benefits and risks of the test. The outcomes to be discussed in this chapter are those that are relevant to screening tests, diagnostic tests, and prognostic tests, although prognostic tests are also discussed in Chapter 11. We also address unique issues that might arise if the test in question is a genetic test, although genetic tests are explored in more detail in Chapter 10. We include a framework for generating potential outcomes for inclusion, and discuss the role of stakeholders in choosing the outcomes for study. Finally, we give examples of systematic reviews that either included a range of outcomes in the review or might have done so.

Common Challenges

Investigators are tasked with choosing the outcomes to consider in a systematic review about a medical test. Resource limitations require judicious selection from among all possible outcomes, which necessitates setting priorities for the outcomes to include. If reviewers do not *explore* the full range of outcomes at the outset of the project, the likelihood of excluding important outcomes is high; the systematic review may miss outcomes relevant to the stakeholder(s). The balance of the benefits and harms from testing will be skewed by the absence of information about key outcomes. The consequence may be that recommendations based on the systematic review are inapt when the test is used in practice. Additionally, for tests that offer modest clinical gains over another test, information on additional outcomes, for example, costs or convenience, may be essential for making decisions based on the test results,. However, we caution that if the initially broad range of outcomes is not carefully condensed, the quality of the review will be threatened by resource limitations (Figure 3–1).

Figure 3–1. Balance of outcomes against resources



Either misstep can result in a suboptimal review—the narrow review may be incomplete, and the broad review may be too superficial to provide meaningful insights.

Principles for Addressing the Challenges and Recommended Approaches for Incorporating All Decision-Relevant Outcomes

We recommend a two-step approach for choosing the outcomes for inclusion in a review about a medical test. The first step is to catalog outcomes methodically, and the second is to solicit input from the stakeholder(s). Below is a description of a conceptual approach to identifying outcomes to ensure that relevant outcomes are not overlooked.

Principle 1: Catalog outcomes methodically.

Conceptual Approach To Identifying Outcomes

The preceding chapter described frameworks for designing systematic reviews about medical tests, including the PICOTs typology (i.e., population, intervention, comparisons, outcomes, timing, and setting). Here we present another framework specifically for thinking about the outcomes from using a test in a clinical setting. In this framework, outcomes are separated into those attributable to the testing process and those attributable to knowledge of the test results. In general, outcomes attributable to the testing process are direct effects of the test; outcomes attributable to the test results are more numerous and include the patient's response to the test results and how the patient and clinician act upon the results.

Bossuyt and McCaffery described a useful framework for thinking about patient outcomes attributable to medical testing.[1] They classified outcomes into three groups: (1) outcomes that result from clinical management based on the test results; (2) the direct health effects of testing; and (3) the patients' emotional, social, cognitive, and behavioral responses to testing. We extend this model by including two additional elements to arrive at five types of outcomes: (4) the legal and ethical effects of testing, which may or may not be a consideration depending on the test under consideration; and (5) the costs of testing. These five categories of outcomes can be associated with the testing process, or with the test result, or with both.

We suggest that the relative importance of these outcomes may differ substantially depending on whether the intention of the test is screening, diagnosis, or prognosis. (Table 3–1) To illustrate, the adverse emotional effects, and the legal and ethical outcomes of testing, might be more significant for medical tests used for screening than tests used for diagnosis, due to the high prevalence of false positive test results associated with many tests used for screening purposes. Additionally, screening tests are conducted in individuals who are without symptoms of the disease of interest, so any adverse or disruptive consequences of testing may be more pronounced. Mammography is a useful example, since the emotional reaction to a false positive test may be substantial. Correspondingly, the potential legal consequences of a false negative test are substantial, as a false negative test may lead to the filing of a malpractice suit. Missed diagnoses, in particular breast cancer diagnoses, are a large category of radiology-related malpractice suits.[2]

Table 3–1. Outcomes that might be particularly consequential depending on type of medical test

Outcomes	Screening Test	Diagnostic Test	Prognostic Test
Clinical management	+	+++	+++
Direct health effects	+	++	++
Emotional, social, cognitive, behavioral responses	+++	++	++
Legal and ethical	+++	++	++
Costs	++	++	++

Systematic reviewers should remember as well that a normal test result, that is a test that has correctly excluded the presence of disease, may be as affecting as a test that has made a diagnosis, and inclusion of outcomes resulting from a negative test may be important in the review. The primary studies of the medical test may have assessed behaviors and consequences after a normal test result, which may include additional testing when a diagnosis is sought or a change in behavior in response to a normal test result (e.g., less attention to healthy lifestyle or

possibly redoubled efforts at maintaining good health). These are all appropriate outcomes for consideration for inclusion in a systematic review.

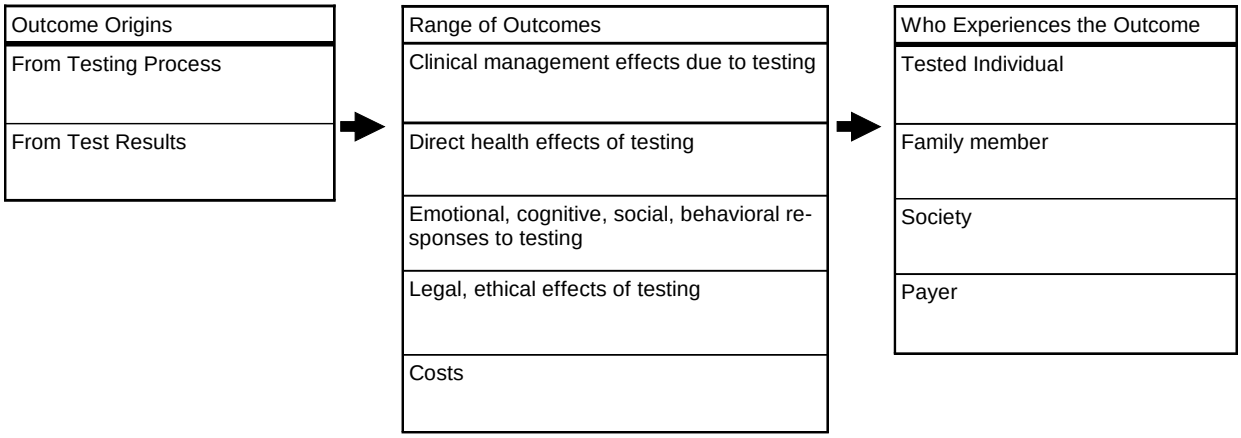
The impact of testing on clinical management is a more important consideration for reviewing diagnostic testing and less important for screening tests, where the clinical management may be quite removed from the screening step. A useful example of diagnostic testing is the use of computed tomography (CT) for detection of pulmonary embolism: a positive test will result in many months of anticoagulation therapy, an important clinical management consequence for the patient. Therefore, systematic reviews will ideally include primary literature that tests the clinical consequences resulting from the use of CT in this setting (rather than just the sensitivity and specificity and predictive values of the test). It is likely that the direct health effects of screening tests are less than in tests used for diagnosis and prognosis: screening tests are generally designed to be less invasive than tests used to make diagnoses in individuals suspected of having disease. An example is PAP testing for cervical cancer screening: there should be no direct health effects of this process.

The range of downstream activities that result from a test are also appropriate to consider for inclusion as outcomes. These may be particularly prominent in imaging tests where there is a high likelihood of identifying unexpected findings that necessitate further evaluation (e.g., unexpected adrenal masses seen during abdominal imaging), or in imaging tests that identify unexpected findings that worry the patient (e.g., degenerative spine changes seen on chest imaging). In selecting outcomes in these situations, one might consider the emotional and cognitive outcomes of unexpected findings, or the monetary costs of the downstream evaluation of incidentally identified abnormalities.

Additional cost outcomes might be considered if appropriate to the systematic review. In addition to the direct costs of the test, one might consider the downstream costs triggered by the results of the testing, which may include confirmatory testing following a positive result, treatment costs resulting from detecting disease, and costs for treatment of adverse effects of the testing (including direct harms of the test and downstream harms resulting from additional testing or treatment, or from evaluation of incidental findings.) Other costs to consider might be the costs to society from direction of funds to testing and away from other services. As an example, one might include, in a systematic review of universal newborn screening, the impact of diverting funding away from other childhood programs such as vaccination.

In addition to consideration of the consequences of testing, we suggest that reviewers also consider an additional axis; namely, who experiences the outcome. The individual being tested is not the only one who can experience outcomes from the testing process. Outcomes may be experienced by family members, particularly in the case of testing an index person for heritable conditions. Outcomes may be experienced by the population *away* from which resources are diverted by a screening activity, e.g., widespread newborn screening which diverts resources away from population-based smoking cessation activities. Society as a whole may experience some outcomes, as when a test of an individual leads to a public health intervention, e.g. prophylactic antibiotics or quarantine after exposure to an infectious individual, or diversion of resources in order to pay for testing of other individuals. Payers are affected if they need to pay for a treatment of a newly diagnosed condition (Figure 3–2).

Figure 3–2. Mapping outcomes to the testing process and to the test results



In summary, a wide range of outcomes could be included in a systematic review of a test. We encourage investigators doing systematic literature reviews to think through this range of outcomes, considering the testing process, the test results, the range of associated outcomes, and the parties that may experience the outcomes. As we discuss below, these considerations may differ depending on the type of test under consideration, and will differ importantly by the specific test and the question being addressed by the systematic review.

Principle 2: Solicit input from stakeholders.

Stakeholders' Role in Defining the Outcomes and Guidance From the Reviewers

Because the range of outcomes that a reviewer might include is broad, expecting such reviews to include "all possible outcomes" is unrealistic. The AHRQ *Methods Guide for Effectiveness and Comparative Effectiveness Reviews* (also referred to as the *General Methods Guide)* recommends that stakeholders be involved at several steps in the systematic review process.[3] We describe additional considerations regarding the role of stakeholders in reviews of medical tests, as their inputs are particularly relevant to the choice of outcomes for inclusion.

Little to no empiric evidence exists regarding what outcomes are most essential for inclusion in a systematic review. If the systematic reviewers knew that some outcomes are universally valued by users of reports, these would be routinely included. It is likely, however, that the choice of outcomes depends largely on the needs of stakeholders and how they intend to use the review. Clinicians and patients are frequently the primary users of the results of systematic reviews and therefore are important stakeholders in the outcomes selection process. An understanding of what evidence the patient or clinician needs in order to make a decision about use of the test is vital in selecting outcomes for inclusion in a review. Certainly the health effects of testing and the emotional or behavioral, social, or cognitive outcomes are directly relevant to the patient; a comprehensive review must include outcomes that are important to patients and would influence their use of a test.

To give an example of another type of stakeholder, the *Evaluation of Genomic Applications in Practice and Prevention* (EGAPP) group of the Centers for Disease Control and Prevention (CDC) has sponsored several EPC reports.[4–6] EGAPP uses these reports to generate guidelines that the CDC issues about genetic testing. EGAPP's interests are broad; it aims to maximize the effectiveness of genetic testing at a societal level. Understandably, the outcomes that it considers

relevant are also broad, and range from the analytic validity of the test to the impact of the testing process on family members. When the possible outcomes for inclusion are many, the investigators have a responsibility to work with the stakeholder to refine the questions carefully so that the task can be accomplished.

Other stakeholders, like professional societies such as the American College of Physicians, may be most interested in evidence reports that can be used to generate recommendations or guidelines for practicing clinicians. Therefore, as stakeholders, they may be more focused on how clinical outcomes vary as a result of medical testing, and perhaps less interested in outcomes that may be more relevant to payers, such as cost-shifting to accommodate costs of testing and downstream costs.

Not infrequently, the primary users of systematic reviews are Federal agencies such as the Centers for Medicare & Medicaid Services (CMS). This agency is responsible for decisions regarding coverage of their beneficiaries' medical care, including medical tests. Therefore, CMS may specify that the outcome most relevant to its coverage decision is the analytic validity of the test, as it would not want to cover a test that inadequately identifies the condition of interest.

The researchers doing comprehensive systematic reviews have a role in helping stakeholders understand the breadth of outcomes. The researchers might assist stakeholders with mapping the range of outcomes depicted in Figure 3–2. This will allow the stakeholders to review the breadth of outcomes and characterize the outcomes as being more or less vital depending on the intended use of the review.

Illustrations of the Principles

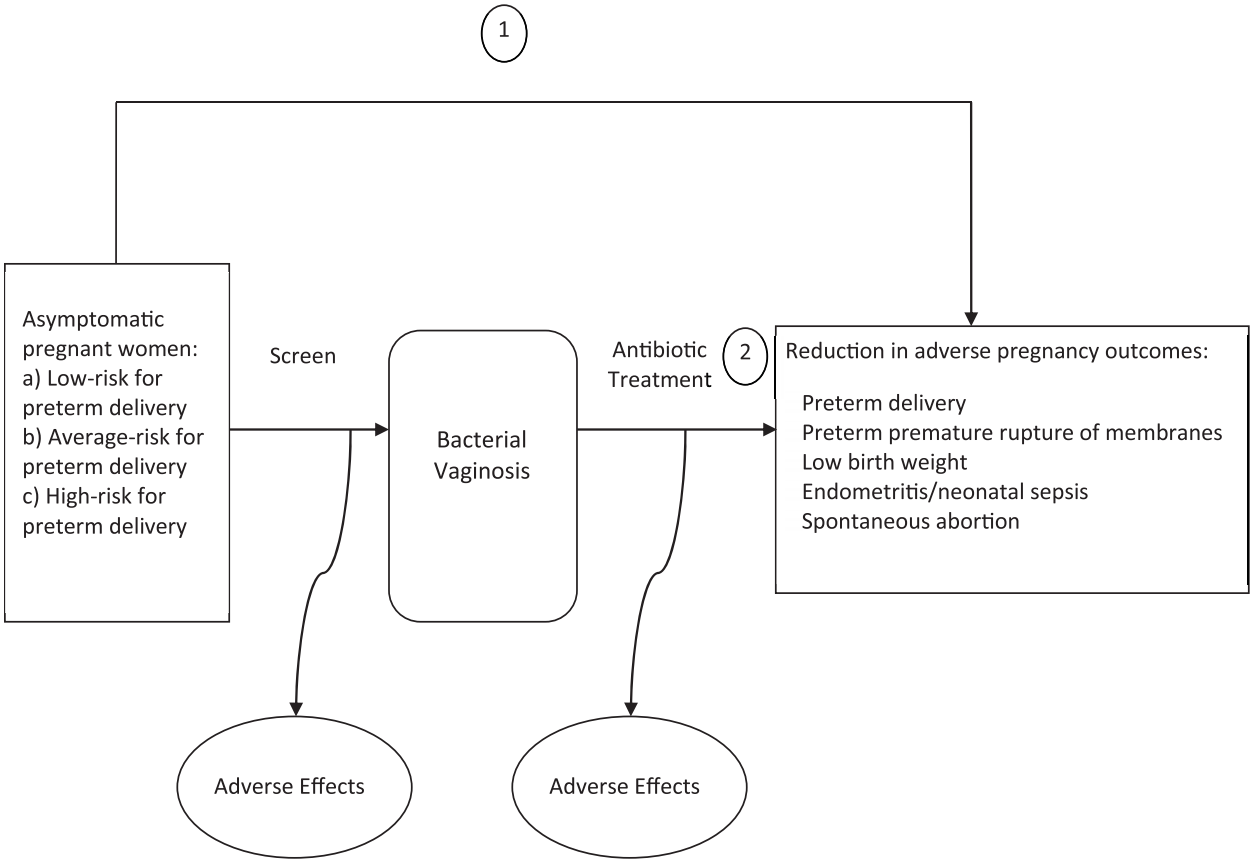
To explain these points in more detail, we use three examples: one each of a screening test, a diagnostic test, and a prognostic test. In discussing these examples, we consider both outcomes that result from the process of testing or are associated with the results of testing, and outcomes that affect the tested individual and others. We conclude with a discussion of additional considerations when the test is a genetic test.

Example of a Screening Test

Screening tests are used to detect disease in asymptomatic individuals or individuals with unrecognized symptoms.[7] Screening tests should be able to separate individuals with the disease of interest from those without, and should be employed when there is a treatment available and where early treatment improves outcomes. The U.S. Preventive Services Task Force (USPSTF) develops recommendations for use of clinical preventive services in the United States. An EPC is sometimes tasked with preparing the supporting review of the evidence.[8,9] Other stakeholders have interest in screening tests as well, including professional organizations involved in guideline preparation for their practitioners; cases in point are recommendations made by the American College of Obstetrics and Gynecology regarding cervical cancer screening[10] and the American Cancer Society's recommendations for early cancer detection.[11]

To illustrate outcomes in a systematic review of a screening test, we present the example of a systematic review about screening for bacterial vaginosis in pregnant women.[12] This systematic review was first done for the USPSTF in 2001 and was later updated. Figure 3–3 depicts the analytic framework developed by the authors.

Figure 3–3. Screening example: bacterial vaginosis



Clinical management effects. The authors addressed whether screening for bacterial vaginosis during pregnancy in asymptomatic women reduces adverse pregnancy outcomes. They included a review of the clinical management effects that would result from antibiotic treatment based on screening results. These included adverse effects of therapies and beneficial effects—reduction in adverse pregnancy outcomes such as preterm delivery. The authors might also have explicitly included an outcome that examines whether the screening leads to receipt of antibiotic treatment—that is, whether screening leads to a change in clinical management. This would be a relevant intermediate outcome on the path between screening and the outcomes attributable to therapy.

Direct test effects. Appropriately, the authors of this review did not include outcomes that are a direct result of the testing process because direct test effects are unlikely in this example: a vaginal swab will not cause any injury. Similarly, the test does not confer any direct benefit either, except perhaps contact with clinicians.

Emotional, social, cognitive, or behavioral effects. The authors might have also looked at the emotional, social, cognitive, or behavioral effects of the screening process or the screening test results. It might have been appropriate to consider outcomes that are associated with screening but are not the result of antibiotic therapy. Consideration might have been given to the effects of testing positive for bacterial vaginosis, such as emotional responses to a diagnosis of infection. leading to either healthier or riskier prenatal activities, or maternal worry as an outcome.

As with any measure, the systematic review team might require that the instrument used to measure emotional response be a validated and appropriate instrument.

Legal and ethical effect of testing. Although specifying ethical issues in screening for bacterial vaginosis (which is not a sexually transmitted infection) may seem unnecessary, bacterial vaginosis testing may be done as part of an infectious disease screening for reportable diseases such as syphilis or HIV. Therefore, a review of the effects of testing should consider whether the test being reviewed might be administered with concurrent screening tests that could themselves raise ethical issues.

Costs of the test. The authors of this review did not consider the costs of the test to the patient as an outcome. Widespread initiation of screening programs, such as on a population level, may have profound cost implications.

The authors of this review considered the effects of screening on the mother and on the fetus or infant. However, they might have also considered other relevant parties; these might include the mother's partner and society, as antibiotic resistance is a conceivable outcome from widespread testing and treatment of bacterial vaginosis.

Example of a Diagnostic Test

We differentiate diagnostic tests from screening tests largely by the population being tested. Whereas a diagnostic test is applied to confirm or refute disease in a symptomatic person, a screening test is used in an asymptomatic or pre-symptomatic person. The USPSTF mostly makes recommendations about screening tests that may be used in the general population; other organizations are more concerned with ensuring safe use of diagnostic tests in patient populations. Payers are also interested in optimizing use of diagnostic tests, as many are costly.

We discuss a review that addressed the diagnostic value of 64-slice computed tomography (CT) in comparison to conventional coronary angiography.[13] Stating that their review concerned the "accuracy" of CT, the authors aimed to assess whether 64-slice CT angiography might replace some coronary angiography for diagnosis and assessment of coronary artery disease. A broader review may consider the effectiveness of CT angiography, and the investigators would consider the full range of outcomes as below.

Clinical management effects. Numerous clinical management effects might follow testing for coronary artery disease with CT. The authors of the review focused exclusively on detection of occluded coronary arteries and not on any downstream outcomes from their identification. Individuals diagnosed with coronary artery disease are subjected to many clinical management changes; these include medications, recommendations for interventions such as angioplasty or bypass surgery, and recommendations for lifestyle changes; each of these changes has associated benefits and harms. All of these may be appropriate outcomes to include in evaluating a diagnostic test. If one test under consideration identifies more coronary artery disease than another, this difference will be reflected in clinical management and the consequences of the management choices.

Other conceivable clinical management effects relate to the impact of testing on other health maintenance activities. For example, a patient might defer other necessary testing (e.g., bone densitometry) to proceed with the CT. We would expect, however, that this would also be the case in the comparison arm. Family members may be affected as well by testing; for instance,

they may be called upon to assist the diagnosed patient with future appointments, which may necessitate time away from work and cause emotional stress.

Direct test effects. The test under consideration is a radiographic test. It confers no direct benefit itself (unlike the comparison procedure in which an intervention can be performed at the time of conventional diagnostic angiography). The testing process poses potential harms, including allergic reaction to the intravenous contrast material, renal failure from the contrast material, and radiation exposure. These are all outcomes that could be considered for inclusion. In this example, the comparison test carries comparable or greater risks.

Emotional, social, cognitive, or behavioral effects. The testing process itself is unlikely to have significant emotional consequences, as it is not an invasive test and is generally comfortable for the tested individual. The results of testing could indeed have emotional or behavioral consequences. An individual diagnosed with coronary disease might alter his or her lifestyle to reduce disease progression. On the other hand, an individual might become depressed by the results and engage in less self-care or riskier behavior. These behavioral effects are likely to affect the family members as well. However, in this example the emotional or behavioral effects are expected to be similar for both CT and conventional angiography and therefore may not be relevant for this particular review. In contrast, they would be relevant outcomes if CT angiography were being compared with no testing.

Legal and ethical effects of testing. Testing could have legal consequences if the tested individual is in a profession that requires disclosure of health threats for the safety of the public; this might arise if, e.g., the tested person is an airline pilot. However again, this outcome is not expected to differ between CT and conventional angiography.

Costs of the test. The relative costs of the two tests to the insurer and the patient, and the costs of diverting equipment away from other uses, could also be of interest to some stakeholders.

Outcomes Unique to Prognostic Tests

A prognostic test is a test used in individuals with known disease to predict outcomes. The procedure itself may be identical to a procedure that is used as a screening test or a medical test, but the results are applied with a different purpose. Given these differences, additional considerations for outcomes should be included in reviews. For example, consider the use of spirometry for predicting prognosis in individuals with chronic obstructive pulmonary disease (COPD). The test is commonly used to make the diagnosis of COPD and to monitor response to treatment, but the question has been raised as to whether it might also predict survival. In 2005, the Minnesota EPC did a systematic review of this topic on behalf of the American Thoracic Society, American College of Physicians, American Academy of Family Physicians, and American Academy of Pediatrics.[14] The discussion below focuses on one of their key questions: whether prediction of prognosis with spirometry, with or without clinical indicators, is more accurate than prediction based on clinical indicators alone. The sponsoring organizations were interested in predicting patients' survival free of premature death and disability.

Clinical management effects. The results from prognostic testing will have effects on clinical management. Although the prognoses for some diseases are minimally modifiable with current treatments, most prognostic information can be used to alter the course of treatment. In this

example, spirometry may suggest a high likelihood of progressing to respiratory failure and prompt interventions to avert this (e.g., pulmonary rehabilitation efforts, changes in medication, avoidance of some exposures). Conversely, the prognostic information may be used to make decisions regarding other interventions. If the likelihood of dying from respiratory failure is high, patients and their physicians may choose to refrain from colonoscopy and other screening procedures from which the patient is unlikely to benefit. Similarly, treatments of other conditions may be of less interest if life expectancy is short.

Direct test effects. Spirometry has few direct test effects, although patients can have adverse reactions to testing, particularly if challenged with methacholine as part of the test. In general, it is unlikely that tests used for prognosis are more or less likely to have direct test effects than tests used for other purposes.

Emotional, social, cognitive, or behavioral effects. We doubt that many emotional or cognitive effects would arise in response to the testing process. Spirometry is a noninvasive test that most patients tolerate well. Emotional effects in response to the results of testing are possible; emotional effects could even be more pronounced for prognostic tests than for screening or medical tests if the test yields more specific information about mortality risk than is usual from a diagnostic test. There could be a range of effects on behavior including efforts to alter prognosis, like quitting smoking. Test results with prognostic information would be expected to affect family members as well.

Legal and ethical effects of testing. Results of tests that provide prognostic information could have legal outcomes, too, especially if the tested individual acts in ways that belie the information he has received (e.g., entering into a contract or relationship that he is unlikely to fulfill). In this present example, it is unlikely that the prognostic information from spirometry would actually raise these issues, but in other cases, such as a test that demonstrates widely metastatic cancer, this could be an issue. These legal and ethical effects of testing may reach beyond the tested individual and affect society if many individuals have substantial concealed information that influences their actions.

Costs of the test. The relative costs of the test to the insurer and the patient, relative to the costs of collecting information from a history and physical examination, may all be of interest to stakeholders.

If the Test Is a Genetic Test

Chapter 10 of this guide describes in detail unique issues regarding evaluation of genetic tests. With respect to relevant outcomes, we note a few considerations here. Most prominent is the effect on family members. Genetic information about the tested individual has direct bearing on family members who share genes. This may affect emotional and behavioral outcomes, and ethical outcomes, if family members feel pressured to proceed with testing to provide better information for the rest of the family. A second issue is possible impact on health insurance eligibility. Recent legislation in the United States prohibits the use of genetic test results to exclude an individual from health insurance coverage, making this less a relevant outcome than in the past. This policy varies worldwide, however, and may be a relevant consideration in some countries.

Summary

In specifying and setting priorities for outcomes to address in their systematic reviews, investigators should:

- Consider both outcomes that are relevant to the process of testing and those that are relevant to the results of the test.
- Consider inclusion of outcomes in all five domains: clinical management effects, direct test effects; emotional, social, cognitive and behavioral effects; legal and ethical effects; and costs.
- Consider to which group the outcomes of testing are most relevant.
- Given resource limitations, prioritize which outcomes to include. This decision depends on the needs of the stakeholder(s), who should be assisted in prioritizing the outcomes for inclusion.

References

1. Bossuyt PM, McCaffery K. Additional patient outcomes and pathways in evaluations of testing. Med Decis Making 2009;29:E30-E38.
2. Berlin L, Berlin JW. Malpractice and radiologists in Cook County, IL: trends in 20 years of litigation. AJR Am J Roentgenol 1995;165:781-788.
3. Agency for Healthcare Research and Quality. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville, MD: Agency for Healthcare Research and Quality; 2008–. http://www.ncbi.nlm.nih.gov/books/NBK47095. Accessed April 25, 2012.
4. Matchar DB, Thakur ME, Grossman I, et al. Testing for Cytochrome P450 Polymorphisms in Adults with Non-Psychotic Depression Treated With Selective Serotonin Reuptake Inhibitors (SSRIs). Evidence Report/Technology Assessment No. 146. AHRQ Publication No. 07-E002. Rockville, MD: Agency for Healthcare Research and Quality; Janurary 2007. http://www.ahrq.gov/downloads/pub/evidence/pdf/cyp450/cyp450.pdf Accessed April 25, 2012.
5. Bonis PA, Trikalinos TA, Chung M, et al. Hereditary Nonpolyposis Colorectal Cancer: Diagnostic Strategies and Their Implications. Evidence Report/Technology Assessment No. 150. AHRQ Publication No. 07-E008. Rockville, MD: Agency for Healthcare Research and Quality; May 2001. http://www.ncbi.nlm.nih.gov/books/NBK38285. Accessed April 25, 2012.
6. Segal JB, Brotman DJ, Emadi A, et al. Outcomes of Genetic Testing in Adults with a History of Venous Thromboembolism. Evidence Report/Technology Assessment No. 180. AHRQ Publication No. 09-E011. Rockville, MD: Agency for Healthcare Research and Quality; June 2009. http://www.ncbi.nlm.nih.gov/books/NBK44591 Accessed April 25, 2012
7. Wilson JM, Jungner YG. Principles and practice of mass screening for disease. WHO Chronicle 1968;22:473.
8. Hillier TA, Vesco KK, Pedula KL, Beil TL, Whitlock EP, Pettitt DJ. Screening for gestational diabetes mellitus: a systematic review for the U.S. Preventive Services Task Force. Ann Intern Med 2008;148:766-775.
9. Whitlock EP, Lin JS, Liles E, Beil TL, Fu R. Screening for colorectal cancer: a targeted, updated systematic review for the U.S. Preventive Services Task Force. Ann Intern Med 2008;149:638-658.
10. Waxman AG. Guidelines for cervical cancer screening: history and scientific rationale. Clin Obstet Gynecol 2005;48:77-97.
11. Smith RA, Cokkinides V, Brawley OW. Cancer screening in the United States, 2008: a review of current American Cancer Society guidelines and cancer screening issues. CA Cancer J Clin 2008;58:161-179.

12. Nygren P, Fu R, Freeman M, Bougatsos C, Guise JM. Screening and Treatment for Bacterial Vaginosis in Pregnancy: Systematic Review to Update the 2001 U.S. Preventive Services Task Force Recommendation. Evidence Synthesis No. 57. AHRQ Publication No. 08-05106-EF-1. Rockville, Rockville, MD: Agency for Healthcare Research and Quality: January 2008.

13. Mowatt G, Cook JA, Hillis GS et al. 64-Slice computed tomography angiography in the diagnosis and assessment of coronary artery disease: systematic review and meta-analysis. Heart 2008;94:1386-1393.

14. Wilt TJ, Niewoehner D, Kim C et al. Use of spirometry for case finding, diagnosis, and management of chronic obstructive pulmonary disease (COPD). Evid Rep Technol Assess (Summ) 2005;1-7.

Funding: Funded by the Agency for Health Care Research and Quality (AHRQ) under the Effective Health Care Program.

Disclaimer: The findings and conclusions expressed here are those of the authors and do not necessarily represent the views of AHRQ. Therefore, no statement should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

Accessibility: Persons using assistive technology may not be able to fully access information in this report. For assistance contact EffectiveHealthCare@ahrq.hhs.gov.

Conflict of interest: The author has no affiliations or financial involvement that conflicts with the information presented in this chapter.

Corresponding author: Jodi B. Segal, M.D, M.P.H., 1830 E. Monument St., Room 8047, Baltimore, MD 21287. Phone (410) 955-9866; Fax (410) 502-6952; email jsegal@jhmi.edu

Suggested citation: Segal JB. Choosing the important outcomes for a systematic review of a medical test. AHRQ Publication No. 12-EHC075-EF. Chapter 3 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.

Chapter 4

Effective Search Strategies for Systematic Reviews of Medical Tests

Rose Relevo, M.L.I.S, M.S., Oregon Health & Science University

Abstract

This chapter discusses appropriate techniques for developing search strategies for systematic reviews of medical tests. It offers general advice for searching for systematic reviews and also addresses issues specific to systematic reviews of medical tests. Diagnostic search filters are currently not sufficiently developed for use when searching for systematic reviews. Instead, authors should construct a highly sensitive search strategy that uses both controlled vocabulary and text words. A comprehensive search should include multiple databases and sources of grey literature. A list of subject-specific databases is provided.

Introduction

Locating all published studies relevant to the key questions is a goal of all systematic reviews. Inevitably, systematic reviewers encounter variation in whether or how a study is published and in how the elements of a study are reported in the literature or indexed by organizations such as the National Library of Medicine. A systematic search must attempt to overcome these issues in order to identify all relevant studies, taking into account the usual constraints on time and resources.

Although I have written this chapter of the *Methods Guide for Medical Test Reviews* (also referred to as the *Medical Test Methods Guide*) as guidance for Evidence-based Practice Centers (EPCs), I hope it will also serve as a useful resource for other investigators interested in conducting systematic reviews on medical tests; and in particular, for the librarian or information specialist conducting the search. Searching for genetic tests and prognostic studies is covered in chapters 11 and 12 of this *Medical Test Methods Guide*.

While this chapter will discuss issues specific to systematic reviews of medical tests, (including screening, diagnostic, and prognostic tests), it is important to remember that general guidance on searching for systematic reviews[1] also applies. Literature searches will always seek a balance between recall (how much of the relevant literature is located) and precision (how much of the retrieved literature is relevant). The optimal balance depends on context. Within the context of comparative effectiveness research, the goal is to have a comprehensive (if not exhaustive) search while still trying to minimize the resources necessary for review of the retrieved citations.

In general, bibliographic searches for systematic reviews in health care should always include MEDLINE® and the Cochrane Central Register of Controlled Trials. Additional databases that are often useful to search include EMBASE®, CINAHL® and PsychINFO®. When

constructing the searches in these bibliographic databases, it is important to use both controlled and uncontrolled vocabulary and to tailor the search for each individual database. Limits such as age and language should not be used unless a specific case can be made for their use.

Working closely with the research team and reviewing the analytic framework and inclusion and exclusion criteria will help to develop the search strategy. Reading the references of all included studies is a useful technique to identify additional studies, as is using a citation database such as Scopus® or Web of Science® to find articles that have cited key articles that have already been retrieved. In addition to published literature, a comprehensive search will include looking for unpublished or "grey literature." In the context of comparative effectiveness research, regulatory information, clinical trial registries, and conference proceedings/abstracts are the most useful sources for identifying data.

Common Challenges

Systematic reviews of test strategies for a given condition require a search for each of the relevant test strategies under consideration. In conducting the search, systematic reviewers may use one of two approaches. Either the reviewers may search on all possible tests used to evaluate the given disease, which requires knowing all the possible test strategies available, or they may search on the disease or condition and then focus on medical test evaluation for that disease.

When a review focuses on specific named tests, searching is relatively straightforward. The names of the tests can be used to locate studies, and a specific search for the concept of diagnosis, screening or prognosis may not be necessary.[2, 3] But because testing strategies are constantly evolving, using the strategy of relying on specific named tests may risk missing emerging approaches. Tests that measure a gene product may be associated with multiple diseases, so searching by test name alone may be insufficient. It is often advisable to search for the target illness in addition to known test names, or for the target illness alone if specific tests are unknown. However, searches for a disease or condition are broader searches and greatly increase the burden of work in filtering down to the relevant studies on medical test evaluation.

Principles for Addressing the Challenges

Principle 1: Do not rely on search filters alone.

Several search filters (sometimes called "hedges"), which are pre-prepared and tested searches that can be combined with searches on a particular disease or condition, have been developed to aid systematic reviewers evaluating medical tests. Most of these filters have been developed for MEDLINE.[2–6] One filter in particular[7] is used in the PubMed® Clinical Queries for diagnosis (Table 4–1). Search filters have also been developed specifically for diagnostic imaging[8] and for EMBASE.[9,10]

Table 4–1. Diagnosis clinical query for PubMed

Category	Optimization	Sensitivity/Specificity	PubMed Search String
Diagnosis	Sensitivity/breadth	98%/74%	(sensitiv*[Title/Abstract] OR sensitivity and specificity[MeSH Terms] OR diagnos*[Title/Abstract] OR diagnosis[MeSH:noexp] OR diagnostic* [MeSH:noexp] OR diagnosis,differential[MeSH:noexp] OR diagnosis[Subheading:noexp])
	Specificity/ narrowness	64%/98%	(specificity[Title/Abstract])

Unfortunately, although these search filters are useful for the casual searcher who simply needs some good articles on diagnosis, they are inappropriate for use in systematic reviews of clinical effectiveness. Several researchers[6,11–14] have reported that using these filters for systematic reviews may result in relevant studies being missed. Vincent found that most of the available filters perform better when they are being evaluated than when they are used in the context of an actual systematic review;[13] this finding is particularly true for studies published before 1990 because of non-standardized reporting and indexing of medical test studies.

In recent years, improved reporting and indexing of randomized controlled trials (RCTs) have made such trials much easier to find. There is reason to believe that reporting and indexing of medical test studies will similarly improve in the future.[12] In fact, Kastner and colleagues[15] recently reviewed 22 systematic reviews of diagnostic accuracy published in 2006 to determine whether the PubMed Clinical Queries Filter for diagnosis would be sufficient to locate all the primary studies that the 22 systematic reviews had identified through traditional search strategies. Using these filters in MEDLINE and EMBASE, the authors found 99 percent of the articles in the systematic reviews they examined, and they determined that the missed articles would not have altered the conclusions of the systematic reviews. The authors therefore concluded that filters may be appropriate when searching for systematic reviews of medical test accuracy. However, until more evidence of their effectiveness is found, we recommend that searchers not rely on them exclusively.

Principle 2: Do not rely on controlled vocabulary (subject headings) alone.

When searching, it is important to use all known variants of the test name such as abbreviations, generic and proprietary names, as well as international terms and spellings, and these may not all be controlled vocabulary terms. Because reporting and indexing of studies of medical tests is so variable, one cannot rely on controlled vocabulary terms alone.[3]

Using textwords for particular medical tests will help to identify medical test articles that have not yet been indexed or that have not been indexed properly.[2] Filters may suggest the sort of textwords that may be appropriate. Michel[16] discusses appropriate MeSH headings and other terminology useful for searching for medical tests.

Principle 3: Search in multiple locations.

Always—but in particular with searches for studies of medical tests—we advise systematic reviewers to search more than one database and to tailor search strategies to each individual database.[17] Because there can be little overlap between many databases,[18–20] failure to search additional databases carries a risk of bias.[21–23] For more information on potentially appropriate databases to use see Table 4–2.

Table 4–2. Specalized databases

Database	URL	Topic Coverage
Free Databases		
The Campbell Library	http://www.campbellcollaboration.org/library.php	Library of Systematic Reviews, Protocols, Reviews of Reviews and Trials for Social Sciences (Similar to Cochrane Library)
ERIC (Education Resources Information Center)	http://www.eric.ed.gov	Education, including the education of health care professionals as well as educational interventions for patients
IBIDS (International Bibliographic Information on Dietary Supplements)	http://ods.od.nih.gov/Health_Information/IBIDS.aspx	Dietary supplements
ICL (Index to Chiropractic Literature)	http://www.chiroindex.org	Chiropractic
NAPS (new Abstracts and Papers in Sleep)	http://www.websciences.org/bibliosleep/naps/default.html	Sleep
OTseeker (Occupational Therapy Systematic Evaluation of Evidence)	http://www.otseeker.com	Occupational therapy
PEDRo (Physiothrarpy Evidence Database)	http://www.pedro.org.au/	Physical therapy
PILOTS	http://www.ptsd.va.gov/ptsd_adv_search.asp	PTSD and traumatic stress
PopLine	http://www.popline.org	Population, family planning & reproductive health
PubMed	http://www.ncbi.nlm.nih.gov/pubmed	Biology and health sciences
RDRB (Research and Development Resource Base)	http://www.rdrb.utoronto.ca/about.php	Medical education
RehabData	http://www.naric.com/research/rehab	Rehabilitation
Social Care Online	http://www.scie-socialcareonline.org.uk	Social care including: healthcare, social work & mental health
TOXNET	http://toxnet.nlm.nih.gov	Toxicology, Environmental health adverse effects
TRIS (Transportation Research Information Service)	http://ntlsearch.bts.gov/tris/index.do	Transportation research
WHO Global Health Library	http://www.who.int/ghl/medicus/en/	International biomedical topics. Global Index Medicus
Subscription Databases		
AgeLine	http://www.csa.com/factsheets/ageline-set-c.php	Aging, health topics of interest to people over 50
AMED (Allied and Complimentary Medicine Database)	http://www.ovid.com/site/catalog/DataBase/12.jsp	Complementary medicine and allied health
ASSIA (Applied Social Science Index and Abstracts)	http://www.csa.com/factsheets/assia-set-c.php	Applied social sciences including: anxiety disorders, geriatrics, health, nursing, social work and substance abuse

Table 4–2. Specialized databases (continued)

Database	URL	Topic Coverage
Subscription Databases (continued)		
BNI (British Nursing Index)	http://www.bniplus.co.uk/about_bni.html	Nursing and midwifery
ChildData	http://www.childdata.org.uk/	Child-related topics, including child health
CINAHL (Cumulative Index to Nursing and Allied Health)	http://www.ebscohost.com/cinahl	Nursing and allied health
CommunityWISE	http://www.oxmill.com/communitywise/	Community issues, including community health
EMBASE	http://www.embase.com	Biomedical, with and emphases on drugs and pharmaceuticals, more non-U.S. coverage than MEDLINE
EMCare	http://www.elsevier.com/wps/find/bibliographicdatabasedescription.cws_home/708272/description#description	Nursing and allied health
Global Health	http://www.cabi.org/datapage.asp?iDocID=169	International health
HaPI (Health and Psychosocial Instruments)	http://www.ovid.com/site/catalog/DataBase/866.jsp	Health and psychosocial testing instruments
IPA (international Pharmaceutical Abstracts)	http://www.csa.com/factsheets/ipa-set-c.php	Drugs and pharmaceuticals
MANTIS (Manual Alternative and Natural Therapy Index System)	http://www.healthindex.com/MANTIS.aspx	Osteopathy, chiropractic, and alternative medicine
PsycINFO	http://www.apa.org/pubs/databases/psycinfo/index.aspx	Psychological literature
Sociological Abstracts	http://www.csa.com/factsheets/socioabs-set-c.php	Sociology, including: health and medicine and the law, social psychology, and substance abuse and addiction
Social Services Abstracts	http://www.csa.com/factsheets/ssa-set-c.php	Social services, including: mental health services, gerontology, and health policy

Until reporting and indexing are improved and standardized, a combination of highly sensitive searches and brute force article screening will remain the best approach for systematically searching the medical test literature.[6, 11–13] However, this approach is still likely to miss relevant articles; therefore authors should search additional sources of information. Tracking citations, reading references of relevant articles, and identifying articles that cite key studies are important ways to find additional citations.[24] Table 4–3 lists databases that are appropriate for tracking citations.

Table 4-3. Citation tracking databases

Database	URL	Subscription Status
Google Scholar	http://scholar.google.com	Free
PubFocus	http://pubfocus.com	Free
PubReMiner	http://bioinfo.amc.uva.nl/human-genetics/pubreminer	Free
Scopus	http://info.scopus.com	Subscription required
Web of Science	http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science	Subscription required

In addition to bibliographic databases and citation analysis, regulatory documents are another potential source of information for systematic reviews of medical reviews. The FDA regulates many medical tests as devices. The regulatory documents for diagnostic tests are available on the FDA's Device website: http://www.accessdata.fda.gov/scripts/cdrh/devicesatfda/.

Illustration: Contrasting Search Strategies

Two contrasting search strategies may help illustrate these principles. In the AHRQ report, *Testing for BNP and NT-proBNP in the Diagnosis and Prognosis of Heart Failure*,[25] the medical tests in question were known. Therefore, the search consisted of all possible variations on the names of these tests and did not need to include a search string to capture the diagnostic testing concept. By contrast, in the AHRQ report, *Effectiveness of Noninvasive Diagnostic Tests for Breast Abnormalities*,[26] all possible diagnostic tests were not known. For this reason, the search strategy included a search string meant to capture the diagnostic testing concept, and this relied heavily on textwords. The actual search strategy used in PubMed to capture the concept of diagnostic tests was as follows: diagnosis OR diagnose OR diagnostic OR di[sh] OR "gold standard" OR "ROC" OR "receiver operating characteristic" OR sensitivity and specificity[mh] OR likelihood OR "false positive" OR "false negative" OR "true positive" OR "true negative" OR "predictive value" OR accuracy OR precision.

Summary

Key points to keep in mind when developing a search strategy for medical test reviews:

- Diagnostic search filters—or, more specifically, the reporting and indexing of medical test studies upon which these filters rely—are not sufficiently well developed to be depended upon exclusively for systematic reviews.
- If the full range of tests is known, one may not need to search for the concept of diagnostic testing; searching for the specific test using all possible variant names may be sufficient.
- Combining highly sensitive searches utilizing textwords with hand searching and acquisition and review of cited references in relevant papers is currently the best way to identify all or most relevant studies for a systematic review.
- Do not rely on controlled vocabulary alone.
- Check Devices@FDA (http://www.accessdata.fda.gov/scripts/cdrh/devicesatfda/).

References

1. Relevo R, Balshem H. Finding evidence for comparing medical interventions: Agency for Healthcare Research and Quality (AHRQ) and the Effective Health Care program. J Clin Epidemiol. 2011;64(11):1168-77.
2. Deville WL, Bezemer PD, BouterLM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. J Clin Epidemiol. 2000. 53(1): 65-9.
3. van der Weijden T et al. Identifying relevant diagnostic studies in MEDLINE. The diagnostic value of the erythrocyte sedimentation rate (ESR) and dipstick as an example. Family Practice. 1997;14(3):204-8.
4. Bachmann LM et al. Identifying diagnostic studies in MEDLINE: reducing the number needed to read. Journal of the American Medical Informatics Association. 2002; 9(6):653-8.
5. Haynes RB et al. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. Journal of the American Medical Informatics Association 1994. 1(6):447-58.
6. Ritchie G, Glanville J, Lefebvre C. Do published search filters to identify diagnostic test accuracy studies perform adequately? Health Information and Libraries Journal. 2007; 24(3):188-92.
7. Haynes RB, Wilczynski NL. Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. BMJ. 2004;328(7447):1040.
8. Astin MP, Brazzelli MG, Fraser CM, et al. Developing a sensitive search strategy in MEDLINE to retrieve studies on assessment of the diagnostic performance of imaging techniques. Radiology. 2008. 247(2):365-73.
9. Bachmann LM et al. Identifying diagnostic accuracy studies in EMBASE. Journal of the Medical Library Association. 2003;91(3):341-6.
10. Wilczynski NL, Haynes RB. EMBASE search strategies for identifying methodologically sound diagnostic studies for use by clinicians and researchers. BMC Medicine. 2005;3:7.
11. Leeflang MM, Scholten RJ, Rutjes AW, et al. Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. J Clin Epidemiol. 2006;59(3):234-40.
12. Doust JA et al. Identifying studies for systematic reviews of diagnostic tests was difficult due to the poor sensitivity and precision of methodologic filters and the lack of information in the abstract. J Clin Epidemiol. 2005;58(5):444-9.
13. Vincent S, Greenley S, Beaven O. Clinical Evidence diagnosis: Developing a sensitive search strategy to retrieve diagnostic studies on deep vein thrombosis: a pragmatic approach. Health Information and Libraries Journal. 2003;20(3):150-9.
14. Whiting P et al. Inclusion of methodological filters in searches for diagnostic test accuracy studies misses relevant studies. J Clinical Epidemiol. 2011;64(6):602-7.
15. Kastner M, Wilczynski NL, McKibbon AK, et al. Diagnostic test systematic reviews: Bibliographic search filters ("clinical queries") for diagnostic accuracy studies perform well. J Clinical Epidemiol. 2009;62(9):974-81.
16. Michel P, Mouillet E, Salmi LR. Comparison of Medical Subject Headings and standard terminology regarding performance of diagnostic tests. J Med Libr Assoc. 2006;94(2):221-3.
17. Honest H, Bachmann LM, and Khan K. Electronic searching of the literature for systematic reviews of screening and diagnostic tests for preterm birth. European Journal of Obstetrics & Gynecology and Reproductive Biology. 2003;107(1):19-23.
18. Conn VS Isaramalai SA, Rath S, et al. Beyond MEDLINE for literature searches. J Nurs Scholarsh., 2003;35(2):177-82.
19. Suarez-Almazor ME, Belsek E, Homick J, et al. Identifying clinical trials in the medical literature with electronic databases: MEDLINE alone is not enough. Control Clin Trials. 2000;21(5):476-487.

20. Betrán AP, Say L, Gülmezoglu AM, et al. Effectiveness of different databases in identifying studies for systematic reviews: experience from the WHO systematic review of maternal morbidity and mortality. BMC Med Res Methodol. 2005;5(1):6.

21. Sampson M, Barrowman NJ, Moher D, et al., Should meta-analysts search Embase in addition to Medline? [see comment]. J Clin Epidemiol. 2003;56(10):943-55.

22. Zheng MH, Zhang X, Ye Q, et al. Searching additional databases except PubMed are necessary for a systematic review. Stroke, 2008. 39(8):e139; author reply e140. [Comment on Stroke 2008;39(6):1911-9.]

23. Stevinson C, Lawlor DA. Searching multiple databases for systematic reviews: added value or diminishing returns? Complement Ther Med. 2004;12(4):228-32.

24. Whiting P, Westwood M, Burke M, et al. Systematic reviews of test accuracy should search a range of databases to identify primary studies. J Clin Epidemiol. 2008;61(4):357-364.

25. Balion, C., et al., Testing for BNP and NT-proBNP in the Diagnosis and Prognosis of Heart Failure. Evidence Report/Technology Assessment No. 142. (Prepared by the McMaster University Evidence-based Practice Center under Contract No. 290-02-0020). AHRQ Publication No. 06-E014. Rockville, MD: Agency for Healthcare Research and Quality; September 2006. Available at: http://www.ahrq.gov/clinic/tp/bnptp.htm. Accessed August 7, 2011.

26. Bruening W, Launders J, Pinkney N,et al, Effectiveness of Noninvasive Diagnostic Tests for Breast Abnormalities. Comparative Effectiveness Review No. 2. (Prepared by ECRI Evidence-based Practice Center under Contract No. 290-02-0019.) Rockville, MD: Agency for Healthcare Research and Quality; February 2006. Available at: http://effectivehealthcare.ahrq.gov/repFiles/BrCADx%20Final%20Report.pdf. Accessed August 7, 2011.

Funding: Funded by the Agency for Health Care Research and Quality (AHRQ) under the Effective Health Care Program.

Conflicts of interest: The author has no affiliations or financial involvements that conflict with the information presented in this chapter.

Corresponding author: Rose Relevo, M.L.I.S., M.S. Oregon Health & Science University, 3181 SW Sam Jackson Park Rd., Portland OR 97217. Phone: 503-220-8262x51318. Fax:503-346-6815. Email: relevo@ohsu.edu.

Suggested citation: Relevo R. Effective Search Strategies for Systematic Reviews of Medical Tests**.** AHRQ Publicaton No. 12-EHC076-EF. Chapter 4 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.

Chapter 5

Assessing Risk of Bias as a Domain of Quality in Medical Test Studies

P. Lina Santaguida, B.Sc.P.T., Ph.D., M.Sc., McMaster University
Crystal M. Riley, M.A., Nanyang Technological University, Singapore
David B. Matchar, M.D., Duke University;
Duke-NUS Graduate Medical School, Singapore

Abstract

Assessing methodological quality is a necessary activity for any systematic review, including those evaluating the evidence for studies of medical test performance. Judging the overall quality of an individual study involves examining the size of the study, the direction and degree of findings, the relevance of the study, and the risk of bias in the form of systematic error, lack of internal validity, and other study limitations. In this chapter of the *Methods Guide for Medical Test Reviews*, we focus on the evaluation of risk of bias in the form of systematic error in an individual study as a distinctly important component of quality in studies of medical test performance, specifically in the context of estimating test performance (sensitivity and specificity). We make the following recommendations to systematic reviewers: (1) When assessing study limitations that are relevant to the test under evaluation, reviewers should select validated criteria that examine the risk of systematic error, (2) categorizing the risk of bias for individual studies as "low," "medium," or "high" is a useful way to proceed, and (3) methods for determining an overall categorization for the study limitations should be established *a priori* and documented clearly.

Introduction

Medical tests are indispensable for clinicians in that they provide information that goes beyond what is available by clinical evaluation alone. Systematic reviews that attempt to determine the utility of a medical test are similar to other types of reviews, for example, those that examine clinical and system interventions. In particular, a key consideration in a review is how much influence a particular study should have on the conclusions of the review. This chapter complements the original *Methods Guide for Effectiveness and Comparative Effectiveness Reviews* (hereafter referred to as the *General Methods Guide*),[1] and focuses on issues of particular relevance to medical tests, especially the estimation of test performance (sensitivity and specificity).

The evaluation of study features that might influence the relative importance of a particular study has often been framed as an assessment of quality. Quality assessment—a broad term used to encompass the examination of factors such as systematic error, random error, adequacy of reporting, aspects of data analysis, applicability, whether ethics approval is specified, and

whether sample size estimates are detailed—has been conceptualized in a variety of ways.[2–3] In addition, some schemes for quality assessment apply to individual studies and others to a body of literature. As a result, many different tools have been developed to formally evaluate the quality of studies of medical tests; however, there is no empirical evidence that any sort of score based on quantitative weights of individual study features can predict the degree to which a study is more or less "true." In this context, systematic reviewers have not yet achieved consensus on the optimal criteria to assess study quality.

Two overarching questions that arise in considering quality in the sense of "value for judgement making" are: (1) Are the results for the population or test in the study accurate and precise (also referred to globally as the study's "internal validity"), and (2) is the study applicable to the patients who are relevant to the review (an assessment of "external validity" with regard to the purpose of the review)? The first question relates to both systematic error (lack of accuracy, here termed bias) and random error (lack of precision). The second question distinguishes the relevance of the study not only to the population of interest in the study (which relates to the potential for bias) but, most importantly for a systematic review, the relevance of the study to the population represented in the key questions established at the outset of the review (i.e., applicability).

This chapter is part of the *Methods Guide for Medical Test Reviews* (also referred to as the *Medical Test Methods Guide*) produced by the Agency for Healthcare Research and Quality (AHRQ) Evidence-Based Practice Centers (EPC) for AHRQ and the Journal of General Internal Medicine. Similar to the *General Methods Guide,*[1] assessment of the major features that influence the importance of a study to key review questions are assessed separately. Chapter 6 of this *Medical Test Methods Guide* considers the evaluation of the applicability of a particular study to a key review question. Chapter 7 details the assessment of the quality of a body of evidence, and Chapter 8 covers the issue of random error, which can be addressed when considering all relevant studies through the use, if appropriate, of a summary measure combining study results. Thus, this chapter highlights key issues when assessing risk of bias in studies evaluating medical tests—systematic error resulting from design, conduct, or reporting that can lead to over- or under-estimation of test performance.

In conjunction with the *General Methods Guide*[1] and the other eleven chapters in this *Medical Test Methods Guide,* the objective is to provide a useful resource for authors and users of systematic reviews of medical tests.

Evidence for Biases Affecting Medical Test Studies

Before considering risk of systematic bias, it is useful to consider the range of limitations in medical test studies. In a series of studies of bias in the context of medical test literature, Whiting et al. reviewed studies of the impact of a range of specific sources of error in diagnostic test studies conducted from 1966 to 2000.[3–5] In the review, the term "'test" was defined broadly to include traditional laboratory tests, clinical examinations, imaging tests, questionnaires, pathology, and measures of health status (e.g., the presence of disease or different stages/severity of a disease).[6] Each test included in the analysis was compared to a reference standard, defined as the best comparator test to diagnose the disease or health condition in question. The results of this analysis indicated that no conclusions could be drawn about the direction or relative magnitude of effects for these specific biases. Although not definitive, the reviews showed that bias does occur and that some sources of bias—including spectrum bias, partial verification bias, clinical review bias, and observer or instrument variation—are particularly common in studies of

diagnostic accuracy.[3] As a guide to further work, the authors summarized the range of quality issues arising in the reviewed articles (Table 5–1).

Table 5-1. Commonly reported sources of systematic bias in studies of medical test performance

Source of Systematic Bias	Description
Population	
Spectrum effect	Tests may perform differently in various samples. Therefore, demographic features or disease severity may lead to variations in estimates of test performance.
Context bias	Prevalence of the target condition varies according to setting and may affect estimates of test performance. Interpreters may consider test results to be positive more frequently in settings with higher disease prevalence, which may also affect estimates of test performance.
Selection bias	The selection process determines the composition of the study sample. If the selection process does not aim to include a patient spectrum similar to the population in which the test will be used, the results of the study may not accurately portray the results for the identified target population.
Test Protocol: Materials and Methods	
Variation in test execution	A sufficient description of the execution of index and reference standards is important because variation in measures of diagnostic accuracy result from differences in test execution.
Variation in test technology	When the characteristics of a medical test change over time as a result of technological improvement or the experience of the operator of the test, estimates of test performance may be affected.
Treatment paradox	Occurs when treatment is started on the basis of the knowledge of the results of the index test, and the reference standard is applied after treatment has started.
Disease progression bias	Occurs when the index test is performed an unusually long time before the reference standard, so the disease is at a more advanced stage when the reference standard is performed.
Reference Standard and Verification Procedure	
Inappropriate reference standard	Errors of imperfect reference standard bias the measurement of diagnostic accuracy of the index test.
Differential verification bias	Part of the index test results is verified by a different reference standard.
Partial verification bias	Only a selected sample of patients who underwent the index test is verified by the reference standard.
Interpretation	
Review bias	Interpretation of the index test or reference standard is influenced by knowledge of the results of the other test. Diagnostic review bias occurs when the results of the index test are known when the reference standard is interpreted. Test review bias occurs when results of the reference standard are known while the index test is interpreted.
Clinical review bias	Availability of clinical data such as age, sex, and symptoms, during interpretation of test results may affect estimates of test performance.
Incorporation bias	The result of the index test is used to establish the final diagnosis.
Observer variability	The reproducibility of test results is one determinant of the diagnostic accuracy of an index test. Because of variation in laboratory procedures or observers, a test may not consistently yield the same result when repeated. In two or more observations of the same diagnostic study, intraobserver variability occurs when the same person obtains different results, and interobserver variability occurs when two or more people disagree.
Analysis	
Handling of indeterminate results	A medical test can produce an uninterpretable result with varying frequency depending on the test. These problems are often not reported in test efficacy studies; the uninterpretable results are simply removed from the analysis. This may lead to biased assessment of the test characteristics.
Arbitrary choice of threshold value	The selection of the threshold value for the index test that maximizes the sensitivity and specificity of the test may lead to over-optimistic measures of test performance. The performance of this cutoff in an independent set of patients may not be the same as in the original study.

Elements of study design and conduct that may increase the risk of bias vary according to the type of study. For trials of tests with clinical outcomes, criteria should not differ greatly from those used for rating the quality of intervention studies.[1] However, medical test performance studies differ from intervention studies in that they are typically cohort studies that have the potential for important sources of bias (e.g., incomplete ascertainment of true disease status, inadequacy of reference standard, and spectrum effect). The next section focuses on some additional challenges in assessing the risk of bias in individual studies of medical test performance.

Common Challenges

Several common challenges exist in assessing the risk of bias in studies of medical test performance. The first challenge is to identify the appropriate criteria to use. A number of instruments are available for assessing many different aspects of individual study quality—not just the potential for systematic error, but also the potential for random error, applicability, and adequacy of reporting.[3] The challenge is to determine which of the existing instruments or which combination of criteria from these instruments are best suited to the task at hand.

A second common challenge is how to apply each criterion in a way that is appropriate to the goals of the review. For example, a criterion that is straightforward for the evaluation of laboratory studies may be less helpful when evaluating components of the medical history or physical examination. Authors must ensure that the review remains true to the spirit of the criterion and that the conclusions are sufficiently clear to be reproducible by others.

Inadequacy of reporting, a third common challenge, does not in itself lead to systematic bias but limits the adequate assessment of important risk of bias criteria. Thus, fairly or unfairly, studies with less meticulous reporting may be assessed as having been less meticulously performed and therefore not deserving the degree of attention given to well reported studies. In such cases, when a study is otherwise judged to make a potentially important contribution, reviewers may need to contact the study's authors to obtain additional information.

Principles for Addressing the Challenges

Principle 1: Use validated criteria to address relevant sources of bias.

In selecting criteria for assessing risk of bias, multiple instruments are available, and reviewers must choose the one most appropriate to the task. Two systematic reviews have evaluated quality assessment instruments specifically in the context of diagnostic accuracy. West et al.[7] evaluated 18 tools (six scales, nine guides, and three EPC rating systems). All of the tools were intended for use in conjunction with other tools relevant for judging the design-specific attributes of the study (for example, quality of RCTs or observational studies). Three scales met all six criteria considered important: (1) the Cochrane Working Group checklist,[8] (2) the tool of Lijmer et al.,[9] and (3) the National Health and Medical Research Council checklist.[10]

In 2005, Whiting et al. undertook a systematic review and identified 91 different instruments, checklists, and guidance documents.[4] Of these 91 quality-related tools, 67 were designed specifically for diagnostic accuracy studies and 21 provided guidance for interpretation, conduct, reporting, or lists of criteria to consider when assessing diagnostic accuracy studies. The majority of these 91 tools did not explicitly state a rationale for inclusion or exclusion of items; neither had the majority of these scales and checklists been subjected to formal test-retest reliability

evaluation. Similarly, the majority did not provide a definition of the components of quality considered in the tool. These variations are a reflection of inconsistency in understanding quality assessment within the field of evidence-based medicine. The authors did not recommend any particular checklist or tool, but rather used this evaluation as the basis to develop their own checklist, the Quality Assessment of Diagnostic Accuracy Studies (QUADAS).

The QUADAS checklist attempted to incorporate the sources of bias and error that had some empirical basis and validity.[6,11–12] This tool contains elements of study limitations beyond those concerned with risk of systematic bias; it also includes questions related to reporting. An updated version of this scale, called QUADAS-2, identifies four key domains (patient selection, index test(s), reference standard, and flow and timing), which are each rated in terms of risk of bias.[13] The updated checklist is shown in Table 5–2.

Table 5–2. QUADAS-2 questions for assessing risk of bias in diagnostic accuracy studies*

Domain 1: Patient Selection
Was a consecutive or random sample of patients enrolled? (Yes/No/Unclear)
Was a case-control design avoided? (Yes/No/Unclear)
Did the study avoid inappropriate exclusions? (Yes/No/Unclear)
Could the selection of patients have introduced bias? Risk: Low/High/Unclear
Domain 2: Index Test(s) (complete for each index test used)
Were the index test results interpreted without knowledge of the reference standard? (Yes/No/Unclear)
If a threshold was used, was it pre-specified? (Yes/No/Unclear)
Could the conduct or interpretation of the index test have introduced bias? Risk: Low/High/Unclear
Domain 3: Reference Standard
Is the reference standard likely to correctly classify the target condition? (Yes/No/Unclear)
Were the reference standard results interpreted without knowledge of the results of the index test? (Yes/No/Unclear)
Could the reference standard, its conduct, or its interpretation have introduced bias? Risk: Low/High/Unclear
Domain 4: Flow and Timing
Was there an appropriate interval between index test(s) and reference standard? (Yes/No/Unclear)
Did all patients receive a reference standard? (Yes/No/Unclear)
Did all patients receive the same reference standard? (Yes/No/Unclear)
Were all patients included in the analysis? (Yes/No/Unclear)
Could the patient flow have introduced bias? Risk: Low/High/Unclear

*Questions related to assessing applicability were excluded here. See the original reference for the complete scale.[13]

We recommend that reviewers use criteria for assessing the risk of systematic error that have been validated to some degree from an instrument like QUADAS-2. Chapters 6 and 8 discuss applicability and random error, which are other important aspects of quality assessment. In addition to disregarding irrelevant items, systematic reviewers may also need to add additional criteria from other standardized checklists such as the Standards for Reporting of Diagnostic Accuracy (STARD)[14] or the Strengthening the Reporting of Genetic Association Studies (STREGA),[15]—an extension of the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE).[16]

Principle 2: Standardize the application of criteria.

In order to maintain objectivity in an otherwise subjective process, it is useful to standardize the application of criteria. There is little empirical evidence to inform decisions about this process. Thus, we recommend that the review team establish clear definitions for each criterion. This approach is demonstrated in the "illustration" section below. In addition, it can be useful to pilot-test the criteria definitions with at least two reviewers (with a third to serve as arbitrar). In

this way, reviewers can revise unreliable terms and measure the reliability of the ultimate criteria.

Consistent with previous EPC guidance and other published recommendations,[2] we suggest summarizing study limitations across multiple items for a single study into simple categories. Building on the guidance given in AHRQ's *General Methods Guide*,[1] we propose using the terms "low," "medium," and "high," to rate risk of bias. Table 5–3 illustrates the application of these three categories in the context of diagnostic accuracy studies. It is useful to have two reviewers independently assign studies to categories, and to reconcile disagreements by discussion. A crucial point is that whatever definitions are used, reviewers should establish the definitions in advance of the final review (*a priori*) and should report them explicitly.

Systematic reviewers will be asked to assign an overall summary category of low, medium, or high to an individual study for the most important outcomes. Reviewers can also conclude an overall summary of unclear risk for poorly reported studies. A study's overall rating for risk of bias can also differ depending on the outcome of interest; this could be attributed to several factors including variation in completeness of data, or blinding of the outcome assessor. Summarizing risk of bias based on several outcomes within a study is not recommended. Finally, a clear rationale and the steps taken to summarize risk of bias must be specified.

Table 5–3. Categorizing individual studies into general quality classes*

Category	Application to Randomized Controlled Trials	Application to Medical Test Performance Studies
Low. No major features that risk biased results	The study avoids problems such as failure to apply true randomization, selection of a population unrepresentative of the target patients, low dropout rates, or analysis by intention-to-treat. Key study features are described clearly, including the population, setting, interventions, comparison groups, outcome measurements, and reasons for dropouts.	RCTs are considered a high-quality study design, but studies that include consecutive patients representative of the intended sample for whom diagnostic uncertainty exists may also meet this standard. A "low risk" study avoids the multiple biases to which medical test studies are subject (e.g., use of an inadequate reference standard, verification bias), and key study features are clearly described, including the comparison groups, outcomes measurements, and characteristics of patients who failed to have actual state (diagnosis or prognosis) verified.
Medium. Susceptible to some bias, but flaws not sufficient to invalidate the results	The study does not meet all the criteria required for a rating of low risk, but no flaw is likely to cause major bias. The study may be missing information, making it difficult to assess limitations and potential problems.	Application of this category to medical test performance studies is similar to application to RCTs.
High. Significant flaws imply biases of various types that may invalidate the results	The study has large amounts of missing information, discrepancies in reporting, or serious errors in design, analysis, and/or reporting.	The study has significant biases determined a priori to be major or "fatal" (i.e., likely to make the results either uninterpretable or invalid).

*See text.

Principle 3: Decide when inadequate reporting constitutes a fatal flaw.

Reviewers must also carefully consider how to handle inadequate reporting. Inadequate reporting, in and of itself, does not introduce systematic bias, but it does limit the reviewers' ability to assess the risk of bias. Some systematic reviewers may take a conservative approach by assuming the worst, while others may be more liberal by giving the benefit of the doubt.

When a study otherwise makes a potentially important contribution to the review, reviewers may resolve issues of reporting by contacting study authors. When it is not possible to obtain these details, reviewers should document that a study did not adequately report a particular criterion.

More importantly, it must be determined *a priori* whether failure to report some criterion might represent a "fatal flaw" (i.e., likely to make the results either uninterpretable or invalid). For example, if a review is intended to apply to older individuals yet there was no reporting of age, this could represent a flaw that would cause the study to be excluded from the review, or included and assessed as "high" with regard to risk of bias. Reviewers should identify their proposed method of handling inadequate reporting *a priori* and document this carefully.

Illustration

A recent AHRQ systematic review evaluated the accuracy of the reporting of family history and the factors that were likely to affect accuracy.[17–18] The index test was patients' self-reports of their family history, and the reference standard test could include verification of the relatives' status from either medical records or disease or death registries. The methods chapter identified a single instrument (QUADAS) to evaluate quality of the eligible studies. The reviewers provided a rationale for their selection of items from within this tool; they excluded 4 of 14 items, and gave their justifications for doing so in an appendix. Additionally, the reviewers provided contextual examples of how each QUADAS item had been adapted for the review. As noted in Table 5–4, partial verification bias was defined in the context of self-reported family history as the index test, and verification by the relatives (through either direct contact, health record, or disease/death registry) was the reference test. The authors provided explicit rules for rating this quality criterion as "yes," "no," or "unclear."

Table 5–4. Interpretation of partial verification bias: the example of family history[17–18*]

Modified QUADAS Item (*Topic/Bias*)	Interpretation
5. Did the whole sample or a random selection of the sample receive verification using a reference standard of diagnosis? *(Partial verification bias)*	This item concerns **partial verification bias**, which is a form of selection bias that occurs when not all of the study participants receive the reference standard (in our context, confirmation of the TRUE disease status of the relative). Sometimes the reason only part of the sample receives the reference standard is that knowledge of the index test results **influence** the decision to perform the reference standard. (Note that in the context of family history, the reference standard can only be applied to family members or relatives. The self report by the probands or informants is the "index test.") We consider the whole sample to be ALL relatives for which the proband or informant provided information (including "don't know" status). **YES:** All relatives that the proband identifies/reports upon represent the whole sample of relatives. As such, some form of verification is attempted for all identified relatives. **NO:** Not all relatives receive verification via the reference standard. As such, we consider partial verification bias to be present in the following situations: 1. Knowledge of the index test will determine which relatives are reported to have the disease status. Often **UNAFFECTED relatives** do not have their disease status verified by any method (assume proband/informant report is the true disease status); in this case, the disease status is verified in the **AFFECTED relatives** only. In this situation, the outcomes of sensitivity and specificity cannot be computed. 2. Relatives for which the proband/informant indicates "don't know status" are excluded and do not have their disease status verified (no reference standard testing). 3. Relatives who are **DECEASED** are excluded from having any verification undertaken (no reference standard testing). 4. Relatives who are **UNABLE TO PARTICIPATE** in interviews or further clinical testing are excluded from having any verification method (no reference standard testing). **UNCLEAR:** Insufficient information to determine whether partial verification was present.

QUADAS = Quality Assessment of Diagnostic Accuracy Studies
*See text.

The systematic reviewer can choose to present ratings of individual QUADAS criteria in tabular form as a percentage of the studies that scored "yes," "no," or "unclear" for each criterion. The developers of the tool do not recommend using composite scores.[6]

Summary

An assessment of methodological quality is a necessary activity for authors of systematic reviews; this assessment should include an evaluation of the evidence for studies of medical test performance. Judging the overall quality of an individual study involves examining the size of the study, the direction and degree of findings, the relevance of the study, and the risk of bias in the form of systematic error, lack of internal validity, and other study limitations. In this chapter of the *Medical Test Methods Guide*, we focus on the evaluation of systematic bias in an individual study as a distinctly important component of quality in studies of medical test performance.

Key Points

- When assessing limitations in studies of medical tests, systematic reviewers should select validated criteria that examine the risk of systematic error.
- Systematic reviewers should categorize the risk of bias for individual studies as "low," "medium," or "high."
- Two reviewers should independently assess individual criteria as well as global categorization.
- Reviewers should establish methods for determining an overall categorization for the study limitations *a priori* and document these decisions clearly.

References

1. Agency for Healthcare Research and Quality. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville, MD: Agency for Healthcare Research and Quality; 2008–. http://www.ncbi.nlm.nih.gov/books/NBK47095. Accessed September 20, 2011.
2. Higgins JPT, Altman DG, Sterne JAC on behalf of the Cochrane Statistical Methods Group and the Cochrane Bias Methods Group. Chapter 8: Assessing risk of bias in included studies. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 (updated March 2011). The Cochrane Collaboration, 2011. Available at: http://www.cochrane-handbook.org. Accessed September 19, 2011.
3. Whiting P, Rutjes AWS, Reitsma JB, et al. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med. 2004;140(3):189-202.
4. Whiting P, Rutjes AWS, Dinnes J, et al. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. J Clin Epidemiol. 2005;58:1-12.
5. Whiting P, Rutjes AWS, Dinnes J, et al. Development and validation of methods for assessing the quality of diagnostic accuracy studies. Health Technol Assess. 2004;8(25):iii, 1-234.
6. Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. 2003;3(25).

7. West S, King V, Carey TS, et al. Systems to rate the strength of scientific evidence. (Prepared by the Research Triangle Institute – University of North Carolina Evidence-based Practice Center under Contract No. 290-97-0011.) AHRQ Publication No. 02-E016. Rockville, MD: Agency for Healthcare Research and Quality. April 2002. Available at: http://www.thecre.com/pdf/ahrq-system-strength.pdf. Accessed September 19, 2011.

8. Cochrane Methods Working Group on Systematic Review of Screening and Diagnostic Tests. Recommended Methods; 1996.

9. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999;282(11):1061-6.

10. National Health and Medical Research Council (NHMRC). How to Review the Evidence: Systematic Identification and Review of the Scientific Literature. Canberra, Australia: NHMRC; 2000.

11. Leeflang MMG, Deeks JJ, Gatsonis C, Bossuyt PMM on behalf of the Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. Ann Intern Med. 2008;149(12):889-97.

12. Centre for Reviews and Dissemination. Systematic Reviews: CRD's Guidance for Undertaking Reviews in Health Care. Centre for Reviews and Dissemination: York, UK; 2009. Available at: http://www.york.ac.uk/inst/crd/pdf/Systematic_Reviews.pdf. Accessed September 19, 2011.

13. Whiting P, Rutjes A, Sterne J, et al. QUADAS-2. (Prepared by the QUADAS-2 Steering Group and Advisory Group). Available at: http://www.bris.ac.uk/quadas/resources/quadas2.pdf. Accessed September 12, 2011.

14. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med. 2003;138(1):40-4.

15. Little J, Higgins JPT, Ioannidis JPA, et al. STrengthening the REporting of Genetic Association studies (STREGA) - an extension of the STROBE statement. Eur J Clin Invest. 2009;39:247-66.

16. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007;370: 1453-7.

17. Qureshi N, Wilson B, Santaguida P, et al. Family History and Improving Health. Evidence Report/Technology Assessment No. 186. (Prepared by the McMaster University Evidence-based Practice Center, under Contract No. HHSA 290-2007-10060-I.) AHRQ Publication No. 09-E016. Rockville, MD: Agency for Healthcare Research and Quality. August 2009. Available at: http://www.ahrq.gov/downloads/pub/evidence/pdf/famhistory/famhimp.pdf. Accessed February 28, 2011.

18. Wilson BJ, Qureshi N, Santaguida P, et al. Systematic review: family history in risk assessment for common diseases. Ann Intern Med. 2009;151(12):878-85.

Acknowledgement: Sean R. Love assisted in the editing and preparation of this manuscript.

Funding: Funded by the Agency for Health Care Research and Quality (AHRQ) under the Effective Health Care Program.

Conflict of interest: None of the authors has any affiliations or involvements that conflict with the information in this chapter.

Corresponding author: Dr. David B. Matchar, M.D., Health Services and Systems Research Duke-NUS Graduate Medical School, 8 College Road, Singapore 169857. Phone: +65-6516-2584. Fax: +65-6534-8632. Email: david.matchar@duke-nus.edu.sg

Suggested citation: Santaguida PL, Riley CR, Matchar DB. Assessing risk of bias as a domain of quality in medical test studies. AHRQ Publication No. 12-EHC077-EF. Chapter 5 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.

Chapter 6

Assessing Applicability of Medical Test Studies in Systematic Reviews

Katherine E. Hartmann, M.D., Ph.D., Institute for Medicine and Public Health, Vanderbilt University, Nashville, Tennessee
David B. Matchar, M.D., Center for Clinical Health Policy Research, Duke University, and Duke University Medical Center, Durham, North Carolina
Stephanie M. Chang, M.D., M.P.H., Center for Outcomes and Evidence, Agency for Healthcare Research and Quality, Rockville, Maryland

Abstract

Use of medical tests should be guided by research evidence about the accuracy and utility of those tests in clinical care settings. Systematic reviews of the literature about medical tests must address applicability to real-world decisionmaking. Challenges for reviews include: (1) lack of clarity in key questions about the intended applicability of the review, (2) numerous studies in many populations and settings, (3) publications that provide too little information to assess applicability, (4) secular trends in prevalence and spectrum of the condition for which the test is done, and (5) changes in the technology of the test itself. We describe principles for crafting reviews that meet these challenges and capture the key elements from the literature necessary to understand applicability.

Introduction

Most systematic reviews are conducted for a practical purpose: to support clinicians, patients, and policymakers—decisionmakers—in making informed decisions. To make informed decisions about medical tests, whether diagnostic, prognostic, or those used to monitor the course of disease or treatment, decisionmakers need to understand whether a test is worthwhile in a specific context. For example, decisionmakers need to understand whether a medical test has been studied in patients and care settings similar to those in which they are practicing, and whether the test has been used as part of the same care management strategy they plan to use. They may also want to know whether a test is robust over a wide range of scenarios for use, or relevant only to a narrow set of circumstances.

The Agency for Healthcare Research and Quality (AHRQ) Evidence-based Practice Centers (EPCs) review scientific literature on topics including clinical care and medical tests to produce evidence reports and technology assessments to inform coverage decisions, quality measures, educational materials and tools, guidelines, and research agendas. The EPCs use four principles for assessing and reporting applicability of individual studies and the overall applicability of a body of evidence. These principles may provide a useful framework for other investigators conducting systematic review of medical tests:

- Determine the most important factors that affect applicability.
- Systematically abstract and report key characteristics that may affect applicability.
- Make and report judgements about major limitations to applicability of individual studies.
- Consider and summarize the applicablity of the body of evidence.

Comprehensive information about the general conduct of reviews is available in the AHRQ Evidence-based Practice Center *Methods Guide for Comparative Effectiveness Reviews*.[1] In this report we highlight common challenges in reviews of medical tests and suggest strategies that enhance interpretation of applicability.

Common Challenges

Unclear Key Questions

Key questions guide the presentation, analysis, and synthesis of data, and thus the ability to judge applicability. Key questions should provide clear context for determining the applicability of a study. Lack of specificity in key questions can result in reviews of larger scope than necessary, failure to abstract relevant study features for evidence tables, less useful organization of summary tables, disorganized synthesis of results, and findings from meta-analysis that do not aggregate data in crucial groupings. In addition, key questions that do not distinguish the management context in which the test is being used can introduce misinterpretations of the literature. A common scenario for such confusion is when the research compares the accuarcy of a new test to another test (i.e., as a replacement), but in reality the test is proposed to be used as a triage test to guide further testing or as an add-on after another test.

If relevant contextual factors are not stipulated in the key questions, decisions during the review process are hindered. Which studies should be included and which excluded? If the patient population and care setting are not explicitly described, the default decision can be to broadly lump all contexts and uses of the test together. However, decisions to "lump" or "split" must be carefully considered and justified. Inappropriate lumping without careful consideration of subgroups that should be analyzed separately may result in oversimplification. Decisions about meaningful subgroupings, for instance by age of participants, by setting (hopsital versus ambulatory), or version of the test, should be made in advance.

Conducting subgroup analyses after appraising the included studies may introduce type 1 error (i.e., a hypothesis may appear true when it is false) due to *a posteriori* biases in interpretation, making it difficult to distinguish whether identified effects are spurious or real. Decisions in advance to split reporting of results for specific subgroups and contexts should be carefully considered and justified. Decisions should be based on whether there is evidence that a particular contextual factor is expected to influence the performance characteristics of the test or its effectiveness as a componenet of care.

Studies Not Specific to the Key Questions

When there is appropriate justification to "split" a review so that key questions or subquestions relate to a specific population, setting, or management strategy, the studies identified for inclusion may not reflect the same subgroups or comparisons identified in the key questions. The reviewer is faced with deciding when these deviations from ideal are minor, and when they are more crucial and likely to affect test performance, clinical decisionmaking, and health outcomes in some significant way. The conduct and synthesis of the findings will require

a method of tracking and describing how the reviewers dealt with two types of mismatches: (1) literature from other populations and contexts that does not directly address the intended context of the key question; and (2) studies that do not provide sufficient in formation about context to determine if they apply. By making annotations througout the review, in tables and synthesis, the reviewer can then note if these types of mismatch apply, how common they were, and what the expected impact is on interpreting applicability.

Rapidly Evolving Tests

A third challenge, especially relevant to medical tests, is that, even more than treatments, tests often change rapidly, in degree (enhancements in existing technologies), type (substantively new technologies) or target (new molecular targets). The literature often contains evidence about tests that are not yet broadly available or are no longer common in clinical use. Secular (i.e., long-term) trends in use patterns and market forces may shape applicability in unanticipated ways. For instance, suppose that a test is represented in the literature by dozens of studies that report on a version that provides dichotomous, qualitative results (present versus absent), and that the company marketing the test subsequently announces production of a new version that provides only a continuous, quantitative measure. Or genetic tests for traits may evolve from testing for a single-nucleotide polymorphisms to determining the gene sequence. In these situations, reviewers must weigh how best to capture data relating the two versions of the test, and must decide whether there is value in reviewing the obsolete test to provide a point of reference for expectations about whether the replacement test has any merit, or whether reviewing only the more limited, newer data better addresses the key question for contemporary practice.

Principles for Addressing the Challenges

The root cause of these challenges is that test accuracy, as well as more distal effects of test use, is often highly sensitive to context. Therefore, the principles noted here relate to clarifying context factors and, to the extent possible, using that clarity to guide study selection (inclusion/exclusion), description, analysis, and summarization. In applying the principles described below, the PICOTS typology can serve as a framework for assuring relevant factors have been systematically assessed (see Table 6–1).[2,3]

Table 6–1. Using the PICOTS framework to assess and describe applicability of medical tests

PICOTS Element	Potential Characteristics To Describe and Assess	Challenges When Assessing Studies	Example	Potential Systematic Approaches for Decisions
Population	Justification for Lumping or splitting key questions. Method of identification/selection Inclusion & exclusion criteria for the review. Demographic characteristics of those included in review. Prevalence of condition in practice and in studies. Spectrum of disease in practice and in studies.	Source of population not described. Study population poorly specified. Key characteristics not reported. Unclear whether test performance varies by population.	Education/literacy level not reported in study of pencil-and-paper functional status assessment.	Exclude *a priori* if key element crucial to assessing intended use is missing. Or include but: – Flag missing elements in tables/text. – Organize data within key questions by presence/absence of key elements. – Include presence/absence as parameter in meta-regression or sensitivity analyses. – Note need for challenge to be addressed in future research.
Intervention	Version of test used in practice and in studies. How and by whom tests are conducted in practice and in studies. Cut-off/diagnostic thresholds applied in practice and in studies. Skill of assessors when interpretation of test is required in studies.	Version/ instrumentation not specified. Training/quality control not described. Screening and diagnostic uses mixed.	Ultrasound machines and training of sonographers not described in study of fetal nuchal translucency assessment for detection of aneuploidy.	Exclude a priori if version is critical and not assessed. Or include but: – Contact authors for clarification. – Flag version of test or deficits in reporting in tables/text. – Discuss implications. – Model cut-offs and conduct sensitivity analyses.
Comparator	Use of gold standard vs. “alloy” standard in studies. Alternate or “usual” test used in the studies. How test is used as part of management strategy (e.g. triage, replacement, or add-on) in practice and in studies. In trials is comparator no testing vs. usual care with ad hoc testing.	Gold standard not applied. Correlational data only.	Cardiac CT compared with stress treadmill without use of angiography as a gold standard.	Exclude a priori if no gold standard. Or include but: – Restrict to specified comparators. – Group by comparator in tables/text.

Table 6–1. Using the PICOTS framework to assess and describe applicability of medical tests (continued)

PICOTS Element	Potential Characteristics To Describe and Assess	Challenges When Assessing Studies	Example	Potential Systematic Approaches for Decisions
Outcome of use of the test	How accuracy outcomes selected for review relate to use in practice: Accuracy of disease status classification. Sensitivity/specificity. Predictive values. Likelihood ratios. Diagnostic odds ratio. Area under curve. Discriminant capacity.	Failure to test "normals," or subset, with gold standard. Precision of estimates not provided. Tests used as part of management strategy in which exact diagnosis is less important than "ruling out" a disease.	P-value provided for mean of continuous test results by disease status but confidence bounds not provided for performance characteristics.	Exclude a priori if test results cannot be mapped to disease status (i.e., 2x2 or other test performance data cannot be extracted). Exclude if subset of "normals" not tested. Or include but: – Flag deficits in tables/text. – Discuss implications. – Assess heterogeneity in meta-analysis and comment of sources of heterogeneity in estimates.
Clinical Outcomes from test results	How studies addressed clinical outcomes selected for the review: Earlier diagnosis Earlier intervention Change in treatment given Change in sequence of other testing Change in sequence/intensity of care Improved outcomes, quality of life, costs, etc.	Populations and study designs of included studies heterogeneous with varied findings. Data not stratified or adjusted for key predictors.	Bone density testing reported in relation to fracture risk reduction without consideration of prior fracture or adjustment for age.	Exclude if no disease outcomes and outcomes key to understanding intended use case. Or include and: – Document details of deficits in tables/text. – Discuss implications. – Note need for challenge to be addressed in future research.
Timing	Timing of availability of results to care team in studies and how this might relate to practice. Placement of test in the sequence of care (e.g. relationship of test to treatment or follow-on management strategies) of studies and how this might relate to practice Timing of assessment of disease status and outcomes in studies.	Sequence of use of other diagnostics unclear. Time from results to treatment not reported. Order of testing varies across subjects and was not randomly assigned.	D-dimer studies in which it is unclear when results were available relative to DVT imaging studies.	Exclude if timing/sequence is key to understanding intended use case. Or include and: – Contact authors for information. – Flag deficits in tables/text. – Discuss implications. – Note need for challenge to be addressed in future research.

Table 6–1. Using the PICOTS framework to assess and describe applicability of medical tests (continued)

PICOTS Element	Potential Characteristics To Describe and Assess	Challenges When Assessing Studies	Example	Potential Systematic Approaches for Decisions
Setting	How setting of test in studies relate to key questions and current practice: Primary care vs. specialty care Hospital-based Routine processing vs. specialized lab or facility Specialized personnel Screening vs. diagnostic use	Resources available to providers for diagnosis and treatment of condition vary widely. Provider type/specialty vary across settings. Comparability of care in international settings unclear.	Diagnostic evaluation provided by geriatricians in some studies and unspecified primary care providers in others.	Exclude if care setting known to influence test/outcomes or if setting is key to understanding intended use case. Or include but: – Document details of setting – Discuss implications

CT = computed tomography; DVT = deep venous thromboembolism; PICOTS = population, intervention, comparator, outcome, timing, setting

Principle 1: Identify important contextual factors.

In an ideal review, all possible factors related to the impact of a test use on health outcomes should be considered. However, this is usually not practical, and some tractable list of factors must be selected before initiating a detailed review. Consider factors that could affect the causal chain of direct relevance to the key question: for instance, in assessing the accuracy of cardiac MRI for detecting atherosclerosis, slice thickness is a relevant factor in assessing applicability. It is also important to consider applicability factors that could affect a later link in the causal chain (e.g., for lesions identified by cardiac MRI vs. angiogram, what factors may impact the effectiveness of treatment?).

In pursuing this principle, consider contextual issues that are especially relevant to tests, such as patient populations (e.g., spectrum effect), management strategy, time effects, and secular or long-term trends.

Spectrum Effect

The severity or type of disease may effect the accuracy of the test. For example, cardiac MRI tests may be generally accurate in identifying cardiac anatomy and functionality, but certain factors may affect the test performance, such as arrythmias, location of lesion, or obesity. Reviews must identify these factors ahead of time and justify when to "split" questions or to conduct subgroup analyses.

Tests as Part of a Management Strategy

Studies on cardiac MRI often select patients with a relatively high pretest probability of disease (i.e. presumably prescreened with other non-invasive testing such as stess EKG) and evaluate the diagnostic accuracy when compared to a gold standard of x-ray coronary angiography. However, the test performance under these conditions does not necessarily apply when used in patients with lower pretest probability of disease, such as when screening patients with no symptoms, or when used as an initial triage test (i.e., compared to stress EKG) rather

than as an add-on test after initial screening. It is important for reviewers to clarify and distinguish the conditions in which the test is studied and in which it is likely to be used.

Methods of the Test Over Time

Diagnostics, like all technology, evolve rapidly. For example, MRI slice thickness has fallen steadily over time, allowing resolution of smaller lesions. Thus excluding studies with older technologies and presenting results of included studies by slice thickness may both be appropriate. Similarly, antenatal medical tests are being applied earlier and earlier in gestation, and studies of test performance would need to be examined by varied cutoffs for stages of gestation; and genetic tests are evolving from detection of specific polymorphisms to full gene sequences. Awareness of these changes should guide review parameters such as date range selection and eligible test type for the included literature to help categorize findings and facilitate discussion of results.

Secular Trends in Population Risk and Disease Prevalence

Direct and indirect changes in the secular setting (or differences across cultures) can influence medical test performance and the applicability of related literature. As an example, when examining the value of screening tests for gestational diabetes, test performance is likely to be affected by the average age of pregnant women, which has risen by more than a decade over the past 30 years, and by the proportion of the young female population that is obese, which has also risen steadily. Both conditions are associated with risk of type II diabetes. As a result, we would expect the underlying prevalence of undiagnosed type II diabetes in pregnancy to be increased, and the predictive values and cost-benefit ratios of testing, and even the sensitivity and specificity of the tests in general use, to change modestly over time.

Secular trends in population characteristics can have indirect effects on applicability when population characteristics change in ways that influence ability to conduct the test. For example, obesity diminishes image quality in tests, such as ultrasound for diagnosis of gallbladder disease or fetal anatomic survey, and MRI for detection of spinal conditions or joint disease. Since studies of these tests often restrict enrollment to persons with normal body habitus, current population trends in obesity mean that such studies exclude an ever-increasing portion of the population. As a result, clinical imaging experts are concerned that these tests may not perform in practice as described in the literature because the actual patient population is significantly more likely to be obese than the study populations. Expert guidance can identify such factors to be considered.

Prevalence is inexorably tied to disease definitions that may also change over time. Examples include: (1) criteria to diagnose acquired immune deficiency syndrome (AIDS), (2) the transition from cystometrically defined detrusor instability or overactivity to the symptom complex "overactive bladder," and (3) the continuous refinement of classifications of mental health conditions recorded in the *Diagnostic and Statistical Manual* updates.[4] If the diagnostic criteria for the condition change, the literature may not always capture such information; thus, expert knowledge with a historical vantage point can be invaluable.

Routine Preventive Care Over Time

Routine use of a medical test as a screening test might be considered an indirect factor that alters population prevalence. As lipid testing moved into preventive care, the proportion of individuals with cardiovascular disease available to be diagnosed for the first time with dyslipidemia and eligible to have the course of disease altered by that diagnosis has changed.

New vaccines, such as the human papilloma virus (HPV) vaccine to prevent cervical cancer, are postulated to change the distribution of viral subtypes in the population and may influence the relative prevalence of subtypes circulating in the population. As preventive practices influence the natural history of disease—such as increasing proportions of a population receiving vaccine—they also change the utility of a medical test like that for HPV detection. Knowledge of preventive care trends is an important component of understanding current practice, and these trends should be considered as a backdrop when contextualizing the applicability of a body of literature.

Treatment Trends

As therapeutics arise that change the course of disease and modify outcomes, literature about the impact of diagnostic tools on outcomes requires additional interpretation. For example, the implications of testing for carotid arterial stenosis are likely changing as treatment of hypertension and the use of lipid-lowering agents have improved.

We suggest two steps to ensure that data about populations and subgroups are uniformly collected and useful. First, refer to the PICOTS typology[2,3] (see Table 6–1) to identify the range of possible factors that might affect applicability and consider the hidden sources of limitations noted above. Second, review with stakeholders the list of applicability factors to ensure common vantage points and to identify any hidden factors specific to the test or history of its development that may influence applicability. Features judged by stakeholders to be crucial to assessing applicability can then be captured, prioritized, and synthesized in designing the process and abstracting data for an evidence review.

Principle 2: Be prepared to deal with additional factors affecting applicability.

Despite best efforts, some contextual factors relevant to applicability may only be uncovered after a substantial volume of literature has been reviewed. For example, in a meta-analysis, it may appear that a test is particularly inaccurate for older patients, although age was never considered explicitly in the key questions or in preparatory discussions with an advisory committee. It is crucial to recognize that like any relationship discovered *a posteriori*, this may reflect a spurious association. In some cases, failing to consider a particular factor may have been an oversight; in retrospect, the importance of that factor on the applicability of test results may be physiologically sensible and supported in the published literature. Although it may be helpful to revisit the issue with an advisory committee, when in doubt, it is appropriate to comment on an apparent association and clearly state that it is a hypothesis, not a finding.

Principle 3: Justify decisions to "split" or restrict the scope of a review.

In general, it may be appropriate to restrict a review to specific versions of the test, selected study methods or types, or populations most likely to be applicable to the group(s) whose care is the target of the review such as a specific group (e.g., people with arthritis, women, obese patients) or setting (e.g., primary care practice, physical therapy clinics, tertiary care neonatal intensive care units). These restrictions may be appropriate (1) when all partners are clear that a top priority of a review is applicability to a particular target group or setting, (2) when there is evidence that test performance in a specific subgroup differs from the test performance in the

broader population or setting, or that a particular version of the test performs differently than the current commonly used version. Restriction of reviews is efficient when all partners are clear that a top priority of a review is applicability to a particular target group or setting. Restriction can be more difficult to accomplish when parties differ with respect to the value they place on less applicable but nonetheless available evidence. Finally, restriction is not appropriate when fully comprehensive summaries including robust review of limitations of extant literature are desired.

Depending on the intent of the review, restricting it during the planning process to include only specific versions of the test, selected study methods or types, or populations most likely to be applicable to the group(s) whose care is the target of the review may be warranted. For instance, if the goal of a review is to understand the risks and benefits of colposcopy and cervical biopsies in teenagers, the portion of the review that summarizes the accuracy of cervical biopsies for detecting dysplasia might be restricted to studies that are about teens; that present results stratified by age; or that include teens, test for interaction with age, and find no effect. Alternatively, the larger literature could be reviewed with careful attention to biologic and health systems factors that may influence applicability to young women.

In practice, we often use a combination of inclusion and exclusion criteria based on consensus along with careful efforts to highlight determinants of applicability in the synthesis and discussion. Decisions about the intended approach to the use of literature that is not directly applicable need to be tackled early to ensure uniformity in review methods and efficiency of the review process. Overall, the goal is to make consideration of applicability a prospective process that is attended to throughout the review and not a matter for *post hoc* evaluation.

Principle 4: Maintain a transparent process.

As a general principle, reviewers should address applicability as they define their review methods and document their decisions in a protocol. For example, time-varying factors should prompt consideration of using timeframes as criteria for inclusion or careful descriptions and analyses as approprite of the possible impact of thes effects on applicability.

Transparency is essential, particularly when a review decision may be controversial. For example, after developing clear exclusion criteria based on applicability, a reviewer may find themselves “empty-handed.” In retrospect, experts—even those accepting the original exclusion criteria—may decide that some excluded evidence may indeed be relevant by extension or analogy. In this event, it may be appropriate to include and comment on this material, clearly documenting how it may not be directly applicable to key questions, but represents the limited state of the science.

An Illustration

Our work on the 2002 Cervical Cancer Screening Summary of the Evidence for the U.S. Preventive Services Task Force[5] illustrates several challenges and principles at work. The literature included many studies that did not use gold standards or testing of normals and many did not relate cytologic results to final histopathologic status. We encountered significant examples of changes in secular trends and availability and format of medical tests: liquid-based cervical cytology was making rapid inroads into practice; resources for reviewing conventional Pap smear testing were under strain from a shortage of cytotechnologists in the workforce and from restrictions on the volume of slides they could read each day; several new technologies had entered the marketplace designed to use computer systems to pre- or postscreen cervical

cytology slides to enhance accuracy; and the literature was beginning to include prospective studies of adjunct use of HPV testing to enhance accuracy or to triage which indiviudals needed evaluation with colposcopy and biopsies to evaluate for cervical dysplasia and cancer. No randomized controlled trials (RCTs) were available using, comparing, or adding new tests or technologies to prior conventional care.

Because no data were available comparing the effects of new screening tools or strategies on cervical cancer outcomes, the report focused on medical test characteristics (sensitivity, specificity, predictive values, and likelihood ratios), reviewing three computer technologies, two liquid cytology approaches, and all methods of HPV testing. Restricting the review to techologies available in the United States, and therefore most applicable, would have reduced the scope substantially. Including all the technologies to determine if there were clear differences among techniques made clear whether potentially comparable or superior methods were being overlooked or no longer offered, but may have also unnecessarily complicated the findings. Only in retrospect, after the decision to include all tests was made and the review conducted, were we able to see that this approach did not substantially add to understanding the findings because the tests that were no longer available were not meaningfully superior.

Although clearly describing the dearth of information available to inform decisions, the review was not able to provide needed information. As a means of remediation, not planned in advance, we used prior USPSTF meta-analysis data on conventional Pap medical test performance,[6] along with the one included paper about liquid cytology,[7] to illustrate the potential risk of liquid cytology overburdening care systems with detection of low-grade dysplasia while not substantively enhancing detection of severe disease or cancer.[8] The projections from the report have since been validated in prospective studies.

For two specific areas of applicabily interest (younger and older age, and hysterectomy status), we included information about underlying incidence and prevalence in order to provide context, as well as to inform modeling efforts to estimate the impact of testing. These data helped improve understanding of the burden of disease in the subgroups compared with other groups and improve understanding about the yield and costs of screening in the subgroups compared with others.

Summary

Review teams need to familiarize themselves with the availability, technology, and contemporary clinical use of the test they are reviewing. They should consider current treatment modalities for the related disease condition, the potential interplay of the disease severity and performance characteristics of the test, and the implications of particular study designs and sampling strategies for bias in the findings about applicability.

As examples throughout this report highlight, applicability of a report can be well served by restricting inclusion of marginally related or outdated studies. Applicability is rarely enhanced by uncritically extrapolating results from one context to another. For example, we could not estimate clinical usefulness of HPV testing among older women from trends among younger women. In the design and scoping phase for a review, consideration of the risks and advantages of restricting scope or excluding publications with specific types of flaws benefits from explicit guidance from clinical, medical testing, and statistical experts about applicability challenges.

Often the target of interest is intentionally large—for example, all patients within a health system, a payer group such as Medicare, or a care setting such as a primary care practice. Regardless of the path taken—exhaustive or narrow—the review team must take care to group

findings in meaningful ways. For medical tests, this means gathering and synthesizing data in ways that enhance ability to readily understand applicability. Grouping summaries of the findings using familiar structures like PICOTS can enhance the clarity of framing of the applicability issues, for instance, grouping results by the demographics of the population included: all women, women and men, by the intervention, grouping together studies that used the same version of the test, or by outcomes, grouping together those studies that report an intermediate markers versus those that measured the actual outcome of interest. This may mean that studies are presented within the review more than once, grouping findings along different "applicability axes" to provide the clearest possible picture.

Since most systematic reviews are conducted for the practical purpose of supporting informed decisions and optimal care, keeping applicability in mind from start to finish is an investment bound to pay off in a the form of a more useful review. The principles summarized in this review can assure valuable aspects of weighing applicability are not overlooked and that review efforts support evidence-based practice.

Key Points

- Early in the review planning process, systematic reviewers should identify important contextual factors that may affect test performance (Table 6–1).
- Reviewers should carefully consider and document justification for how these factors will be addressed in the review—whether through restricting key questions or from careful assessment, grouping, and description of studies in a broader review.
    - A protocol should clearly document which populations or contexts will be excluded from the review, and how the review will assess subgroups.
    - Reviewers should document how they will address challenges in including studies that may only partly fit with the key questions or inclusion/exclusion criteria or that poorly specify the context.
- The final systematic review should include a description of the test's use in usual practice and care management and how the studies fit with usual practice.

References

1. Agency for Healthcare Research and Quality. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville, MD: Agency for Healthcare Research and Quality; 2008–. http://www.ncbi.nlm.nih.gov/books/NBK47095. Accessed April 29, 2012.
2. Matchar DB. Introduction to the methods guide for medical test reviews. AHRQ Publication No. EHC073-EF. Chapter 1 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare REsearch and Quality; June 2012. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.
3. Samson D. Developing the topic and structuring the review: utility of PICOTS, analytic frameworks, decision trees, and other frameworks. AHRQ Publication No. 12-EHC074-EF. Chapter 2 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.
4. American Psychiatric Association, Task Force on DSM-IV. Diagnostic and statistical manual of mental disorders : DSM-IV-TR. 4th ed. Washington, DC: American Psychiatric Association; 2000.
5. Hartmann KE, Hall SA, Nanda K, Boggess JF, Zolnoun D. Screening for Cervical Cancer. Available at: http://www.ahrq.gov/downloads/pub/prevent/pdfser/cervcanser.pdf. Accessed November 8, 2011.
6. McCrory DC, Matchar DB, Bastian L, et al. Evaluation of cervical cytology. Evid Rep Technol Assess (Summ). 1999(5):1-6.
7. Hutchinson ML, Zahniser DJ, Sherman ME, et al. Utility of liquid-based cytology for cervical carcinoma screening: results of a population-based study conducted in a region of Costa Rica with a high incidence of cervical carcinoma. Cancer. 1999;87(2):48-55.
8. Hartmann KE, Nanda K, Hall S, Myers E. Technologic advances for evaluation of cervical cytology: is newer better? Obstet Gynecol Surv. 2001;56(12):765-74.

Funding: Funded by the Agency for Health Care Research and Quality (AHRQ) under the Effective Health Care Program.

Conflicts of interest: The author has no affiliations or financial involvement that conflicts with the information presented in this chapter.

Corresponding author: Katherine E. Hartmann, M.D., Ph.D., Professor, Obstetrics & Gynecology and Medicine, Vanderbilt University School of Medicine, 2525 West End Avenue, Suite 600, Nashville, TN 37203-8291. Phone: (615) 322-0964; Fax: (615) 936-8291; Email: katherine.hartmann@vanderbilt.edu

Suggested citation: Hartmann KE, Matchar DB, Chang S, Assessing applicability of medical test studies in systematic reviews. AHRQ Publication No. 12-EHC078-EF. Chapter 6 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.

Chapter 7

Grading a Body of Evidence on Diagnostic Tests

Sonal Singh, M.D., M.P.H., Johns Hopkins University School of Medicine and Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD
Stephanie M. Chang, M.D., M.P.H., Center for Outcomes and Evidence, Agency for Healthcare Research and Quality, Rockville, MD
David B. Matchar, M.D., Duke-NUS Medical School Singapore, Duke Center for Clinical Health Policy Research, Durham, NC
Eric B. Bass, M.D., M.P.H., Johns Hopkins University School of Medicine and Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD

Abstract

Introduction: Grading the strength of a body of diagnostic test evidence involves challenges over and above those related to grading the evidence from health care intervention studies. This chapter identifies challenges and outlines principles for grading the body of evidence related to diagnostic test performance.
Challenges: Diagnostic test evidence is challenging to grade because standard tools for grading evidence were designed for questions about treatment rather than diagnostic testing; and the clinical usefulness of a diagnostic test depends on multiple links in a chain of evidence connecting the performance of a test to changes in clinical outcomes.
Principles: Reviewers grading the strength of a body of evidence on diagnostic tests should consider the principle domains of *risk of bias, directness, consistency, and precision,* as well as *publication bias, dose response association, plausible unmeasured confounders* that would decrease an effect, and *strength of association*, similar to what is done to grade evidence on treatment interventions. Given that most evidence regarding the clinical value of diagnostic tests is indirect, an analytic framework must be developed to clarify the key questions, and strength of evidence for each link in that framework should be graded separately. However, if reviewers choose to combine domains into a single grade of evidence, they should explain their rationale for a particular summary grade and the relevant domains that were weighed in assigning the summary grade.

Introduction

"Grading" refers to the assessment of the strength of the body of evidence supporting a given statement or conclusion rather than to the quality of an individual study.[1] Grading can be valuable for providing information to decisionmakers, such as guideline panels, clinicians, caregivers, insurers and patients who wish to use an evidence synthesis to promote improved patient outcomes.[1,2] In particular, such grades allow decisionmakers to assess the degree to

which any decision can be based on bodies of evidence that are of high, moderate, or only low strength of evidence. That is, decisionmakers can make a more defensible recommendation about the use of the given intervention or test than they might make without the strength of evidence grade.

The Evidence-based Practice Center (EPC) Program supported by the Agency for Healthcare Research and Quality (AHRQ) has published guidance on assessing the strength of a body of evidence when comparing medical interventions.[1,3] That guidance is based on the principles identified by the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) working group[4–6] with minor adaptations for EPCs. It is important to distinguish between the quality of a study and the strength of a body of evidence on diagnostic tests as assessed by the GRADE and EPC approaches. EPCs consider *"The extent to which all aspects of a study's design and conduct can be shown to protect against systematic bias, nonsystematic bias, and inferential error"* as the *quality* or *internal validity* or *risk of bias* of an individual study.[7] In contrast to the GRADE approach, the EPC approach prefers to use the term "strength of evidence" instead of "quality of evidence" to describe the grade of an evidence base for a given outcome, because the latter term is often equated with the quality of individual studies without consideration of the other domains for grading a body of evidence. An assessment of the *strength* of the entire body of evidence includes an assessment of the quality of an individual study along with other domains. Although the GRADE approach can be used to make judgments about the strength of an evidence base and the strength of recommendations, this chapter considers using GRADE as a tool for assessing only the strength of an evidence base.

When assessing the strength of an evidence base, systematic reviewers should consider four principle domains—*risk of bias, consistency*, *directness,* and *precision*.[5] Additionally, reviewers may wish to consider *publication bias* as a fifth principle domain as recently suggested by the GRADE approach.[6] Additional domains to consider are *dose-response association*, *existence of plausible unmeasured confounders,* and *strength of association (i.e., magnitude of effect*). Of note, GRADE considers applicability as an element of *directness*. This is distinct from the EPC approach, which encourages users to evaluate applicability as a separate component.

EPCs grade the strength of evidence for each of the relevant outcomes and comparisons identified in the key questions addressed in a systematic review. The process of defining the important intermediate and clinical outcomes of interest for diagnostic tests is further described in a previous article.[8] Because most diagnostic test literature focuses on test performance (e.g., sensitivity and specificity), at least one key question will normally relate to that evidence. In the uncommon circumstance in which a diagnostic test is studied in the context of a clinical trial (e.g., test versus no test) with clinical outcomes as the study endpoint, the reader is referred to the *Methods Guide for Effectiveness and Comparative Effectiveness Reviews* (also referred to as the *General Methods Guide*) on evaluating interventions.[1,3] For other key questions, such as those related to analytic validity, clinical validity, and clinical utility, the principles described in the present document and the *General Methods Guide* should apply.

This chapter is meant to complement the EPC *General Methods Guide*, and not to be a complete review. Although we have written this paper to serve as guidance for EPCs, we also intend for this to be a useful resource for other investigators interested in conducting systematic reviews on diagnostic tests. In this paper, we outline the particular challenges that systematic reviewers face in grading the strength of a body of evidence on diagnostic test performance. Our focus will be on diagnostic tests, meaning tests that are used in the diagnostic and management strategy of a patient symptom or complaint, as opposed to prognostic tests, which are for

predicting responsiveness to treatment. We then propose principles for addressing the challenges entailed in conducting this type of systematic review.

Common Challenges

Diagnostic test studies commonly focus on the accuracy of the test in making a disease diagnosis, and the task of grading this body of evidence is a challenge in itself. Through discussion with EPC investigators and a review of recent EPC reports on diagnostic tests,[9–13] we identified common challenges that reviewers face when assessing the strength of a body of evidence on diagnostic test performance.

One common challenge is that standard tools for assessing the quality of a body of evidence associated with an intervention—in which the body of evidence typically relates directly to the overarching key question—are not so easily applied to a body of evidence associated with a diagnostic test—evidence which is often indirect. Indeed, this is the reason that establishing a logical chain with an analytic framework and the associated key questions is particularly important for evaluating a diagnostic test (*see* Chapter 2).[8] It is also the reason we must assess the strength of the body of evidence for each link in the chain. The strength of the body of evidence regarding the overarching question of whether a test will improve clinical outcomes depends both on the total body of evidence, as well as the body of evidence for the weakest link in this chain.

Although there is a temptation to use diagnostic accuracy as an intermediate outcome for the effect of a diagnostic test on clinical outcomes, there is often no direct linkage between the diagnostic accuracy outcome and a clinical outcome. This is particularly challenging when tests are used as a part of an algorithm. While rates of false positives and false negatives may be directly related to adverse effects or harms, other accuracy outcomes such as sensitivity or specificity may not directly correlate to effective management and treatment of disease, especially when the test under question is not directly linked to the use of an established treatment algorithm. When tests are used in regular practice not for final diagnosis and treatment, but as a triage for further testing, then accuracy of diagnosis is less important than accuracy of risk classification.

A second challenge arises in the application of the strength of evidence domains for studies of diagnostic tests. For example, in assessing the precision of estimates of test performance, it is particularly difficult to judge whether a particular confidence interval is sufficiently precise; because of the logarithmic nature of diagnostic performance measurements—such as sensitivity, specificity, likelihood ratios, and diagnostic odds ratios—even a relatively wide confidence interval suggesting that imprecision may not necessarily translate into imprecision that is clinically meaningful. Table 7–1 shows an example where a 10 percent reduction in the sensitivity of various biopsy techniques (from 98 percent to 88 percent in the far right column) changes the estimated probability of having cancer after a negative test by less than 5 percent.[11]

Table 7–1. Example of the impact of precision of sensitivity on negative predictive value*

Type of Biopsy	Post Biopsy Probability of Having Cancer After a Negative Core-Needle Biopsy Result[8]			
	Analysis Results	Analysis Overestimated Sensitivity by 1% (e.g., Sensitivity 97% Rather Than 98%)	Analysis Overestimated Sensitivity by 5% (e.g., Sensitivity 93% Rather Than 98%)	Analysis Overestimated Sensitivity by 10% (e.g., Sensitivity 88% Rather Than 98%)
Freehand automated gun	6%	6%	8%	9%
Ultrasound guidance automated gun	1%	1%	3%	5%
Stereotactic guidance automated gun	1%	1%	3%	5%
Ultrasound guidance vacuum-assisted	2%	2%	3%	6%
Stereotactic guidance vacuum-assisted	0.4%	0.8%	3%	5%

*For a woman with a BI-RADS® 4 score following mammography and expected to have an approximate prebiopsy risk of malignancy of 30 percent. Note that an individual woman's risk may be different from these estimates depending on her own individual characteristics.[11]

Principles for Addressing the Challenges

Principle 1: Methods for grading intervention studies can be adapted for studies evaluating studies on diagnostic tests with clinical outcomes.

A body of evidence evaluating diagnostic test outcomes such as diagnostic thinking, therapeutic choice, and clinical outcomes can be assessed in very much the same way as a body of evidence evaluating outcomes of therapeutic interventions. Issues relating to grading in this type of diagnostic test study are more straightforward than in studies measuring accuracy outcomes. Although this is rarely done, the effect of tests on the clinical outcomes described above can be assessed directly with trial evidence. In cases where trial evidence is available, methods of applying grading criteria such as GRADE should not significantly differ from the methods used for intervention evidence.

An unresolved issue is what to do when there is no direct evidence available linking the test to the outcome of interest. For grading intervention studies, the use of intermediate outcomes, such as accuracy outcomes, would be considered "indirect" evidence and would reduce the strength of the grade. The linkage of accuracy outcomes such as true positives and false positives to clinical outcomes depend in part upon the benefits and harms of available treatments as well as the cognitive or emotional outcomes resulting from the knowledge itself, as outlined in Chapter 3 of this *Medical Tests Methods Guide*.[14]

Currently there is no consensus for one particular approach to grading an overall body of evidence when it is entirely indirect, such as when only studies of accuracy are available. As discussed in Chapter 2 of this *Medical Test Methods Guide*,[8] there are circumstances in which accuracy outcomes may be sufficient to conclude that there is or is not a benefit on clinical outcomes.[15] In other cases in which only indirect evidence on intermediate accuracy outcomes is available, EPCs should discuss with decisionmakers and methodologists the benefits of including such indirect evidence and the specific methods to be used.

Principle 2: Consider carefully what test characteristic measures are the most appropriate intermediate outcomes for assessing the impact of a test on clinical outcomes and for assessing the test's precision in the clinical context represented by the key question.

Consistent with EPC and GRADE principles that emphasize the patient-important outcomes, reviewers should consider how any surrogates such as accuracy outcomes will lead to changes in clinical outcomes. Use of an analytic framework and decision models as described in Chapter 2,[8] help to clarify the linkage between accuracy outcomes and clinical outcomes for systematic reviewers, and users of systematic reviews alike.

If accuracy outcomes are presented as true positives, true negatives, false positives, and false negatives, then they can be easily translated into other accuracy outcomes such as sensitivity and specificity, positive predictive value (PPV) and negative predictive value (NPV). Systematic reviewers need to carefully consider which of these accuracy outcomes to assess based on which outcome will relate most directly to clinical outcomes as well as what levels of precision are necessary.

Sometimes it is more important to "rule out" a particular disease that has severe consequences if missed. In these cases, use of a triage test with high sensitivity and NPV may be what is most important, and actual diagnosis of a particular disease is less important.

When the treatment of a disease has high associated risks, multiple tests are often used to assure the highest accuracy. Tests used in isolation need to have both high sensitivity and specificity, or high PPV and NPV, but if no such test is available, clinicians may be interested in the added benefits and harms of "adding-on" a test. The accuracy outcome of interest of these tests would primarily be high specificity or PPV.

Tests that are more invasive will naturally have greater harms. Additional harms may result from misdiagnosis, so it is almost always important to consider the measurement of false positives and false negatives when assessing the harms of a diagnostic test. The degree of harms from false negatives depends on the severity of disease if there is a missed diagnosis, in addition to the risks from the testing itself (i.e., if the test is invasive and associated with risks in and of itself). The degree of harms from false positives depends on the invasiveness of further testing or treatment, as well as the emotional and cognitive effects of inaccurate disease labeling.

As a simple example, one might have compelling data regarding the value of outcomes resulting from true positive test results, as well as true negative, false positive, and false negative results. In a simple decision model it is possible to identify a threshold line for the combinations of test sensitivity and specificity for which testing versus not testing is a toss-up—where net benefits are equivalent to net harms. To the extent that the confidence intervals for sensitivity and specificity derived from the body of evidence are contained within one territory or the other

("testing better," as in this illustration), these intervals are sufficiently precise for purposes of decisionmaking.[16]

Of course, this formulation over-simplifies many situations. Tests are rarely used alone to diagnose disease and determine treatment choices, but are more often used as part of an algorithm of testing and management. The accuracy outcome of most interest depends on how the test is used in a clinical algorithm, as well as the mechanisms by which the test could improve clinical outcomes or cause harms. Whether or not one uses a decision model to help sort out these issues, considering the important test characteristics and their precision in the clinical context represented by the key question is a necessary step in the process of assessing a body of evidence.

Principle 3: The principle domains of GRADE can be adapted to assess a body of evidence on diagnostic test accuracy.

To assess a body of evidence related to diagnostic test performance, we can adapt the GRADE's principle domains of *risk of bias, consistency, directness,* and *precision* (Table 7– 2). Evaluating *risk of bias* includes considerations of how the study type and study design and conduct may have contributed to systematic bias. The potential sources of bias relevant to diagnostic test performance and strategies for assessing the risk of systematic error in such studies are discussed in Chapter 5 of this *Medical Test Methods Guide*.[17] Diagnostic tests, particularly laboratory tests, can yield heterogeneous results due to different technical methods. For example, studies may report using different antibodies for immunoassays, or may use standards with different values and units assigned to them.

Table 7–2. Required and additional domains and their definitions*

Domain	Definition and Elements	Application to Evaluation of Diagnostic Test Performance
Risk of Bias	Risk of bias is the degree to which the included studies for a given outcome or comparison have a high likelihood of adequate protection against bias (i.e., good internal validity), assessed through main elements: • Study design (e.g., RCTs or observational studies) • Aggregate quality of the studies under consideration from the rating of quality (good/fair/poor) done for individual studies	Use one of three levels of aggregate risk of bias: • Low risk of bias • Medium risk of bias • High risk of bias Well designed and executed studies of new tests compared against an adequate criterion standard are rated as "Low risk of bias."
Consistency	Consistency is the degree to which reported study results (e.g., sensitivity, specificity, likelihood ratios) from included studies are similar. Consistency can be assessed through two main elements: • The range of study results is narrow. • *Variability in study results is explained by differences in study design, patient population or test variability.*	Use one of three levels of consistency: • Consistent (i.e., no inconsistency) • Inconsistent • Unknown or not applicable (e.g., single study) Single-study evidence bases should be considered as "consistency unknown (single study)."

Table 7–2. Required and additional domains and their definitions[*] (continued)

Domain	Definition and Elements	Application to Evaluation of Diagnostic Test Performance
Directness	Directness relates to whether the evidence links the interventions directly to outcomes. For a comparison of two diagnostic tests, directness implies head-to-head comparisons against a common criterion standard. Directness may be contingent on the outcomes of interest.	Score dichotomously as one of two levels of directness: • Direct • Indirect When assessing the directness of the overarching question, if there are no studies linking the test to a clinical outcome, then evidence that only provides diagnostic accuracy outcomes would be considered indirect. If indirect, specify which of the two types of indirectness account for the rating (or both, if this is the case); namely, use of intermediate/ surrogate outcomes rather than health outcomes, and use of indirect comparisons. If the decision is made to grade the strength of evidence of an intermediate outcome such as diagnostic accuracy, then the reviewer does not need to automatically "downgrade" this outcome for being indirect.
Precision	Precision is the degree of certainty surrounding an effect estimate with respect to a given outcome (i.e., for each outcome separately). If a meta-analysis was performed, the degree of certainty will be the confidence interval around the summary measure(s) of test performance (e.g., sensitivity, or true positive).	Score dichotomously as one of two levels of precision: • Precise • Imprecise A precise estimate is an estimate that would allow a clinically useful conclusion. An imprecise estimate is one for which the confidence interval is wide enough to include clinically distinct conclusions.
Publication bias[**]	Publication bias indicates that studies may have been published selectively, with the result that the estimate of test performance based on published studies does not reflect the true effect. Methods to detect publication bias for medical test studies are not robust. Evidence from small studies of new tests or asymmetry in funnel plots should raise suspicion for publication bias.	Publication bias can influence ratings of consistency, precision, and magnitude of effect— and, to a lesser degree, risk of bias and directness). Reviewers should comment on publication bias when circumstances suggest that relevant empirical findings, particularly negative or no-difference findings, have not been published or are unavailable.
Dose-response association	This association, either across or within studies, refers to a pattern of a larger effect with greater exposure (including dose, duration, and adherence).	The dose-response association may support an underlying mechanism of detection and potential relevance for some tests that have continuous outcomes and possibly multiple cutoffs [e.g., gene expression, serum PSA (prostate-specific antigen) levels, and ventilation/perfusion scanning].

Table 7–2. Required and additional domains and their definitions[*] (continued)

Domain	Definition and Elements	Application to Evaluation of Diagnostic Test Performance
Plausible unmeasured confounding and bias that would decrease an observed effect or increase an effect if none was observed.	Occasionally, in an observational study, plausible confounding factors would work in the direction opposite to that of the observed effect. Had these confounders not been present, the observed effect would have larger. In such case the evidence can be upgraded.	The impact of plausible unmeasured confounders may be relevant to testing strategies that predict outcomes. A study may be biased to find low diagnostic accuracy via spectrum bias and yet despite this find very high diagnostic accuracy.
Strength of association (magnitude of effect)	Strength of association refers to the likelihood that the observed effect or association is large enough that it cannot have occurred solely as a result of bias from potential confounding factors.	The strength of association may be relevant when comparing the accuracy of two different medical tests, with one being more accurate than the other. It is possible that the accuracy of a test is better than the reference standard because of an imperfect reference standard. It is important to consider this possibility, and modify the analysis to take into consideration alternative assumptions about the best reference standard.

EPC = Evidence-based Practice Center
[*]Adapted from the *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*.[3]
[**]The GRADE approach is moving towards considering publication bias a GRADE principle domain.

Consistency concerns homogeneity in the direction and magnitude of results across different studies. The concept can be similarly applied to diagnostic test performance studies, although the method of measurement may differ. For example, consistency among intervention studies with quantitative data may be assessed visually with a forest plot. However, for diagnostic test performance reviews, the most common presentation format is a summary receiver operating characteristic (ROC) curve, which displays the sensitivity and specificity results from various studies. A bubble plot of true positive versus false positive rates showing spread in ROC space is one method of assessing the consistency of diagnostic accuracy among studies. As with intervention studies, the strength of evidence is reduced by unexplained heterogeneity—that is, heterogeneity not explained by different study designs, methodologic quality of studies, diversity in subject characteristics, or study context.

Directness, according to AHRQ's *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*,[3] occurs when the evidence being assessed "reflects a single, direct link between the interventions of interest [diagnostic tests] and the ultimate health outcome under consideration."[1] When assessing the directness of the overarching question, if there are no studies linking the test to a clinical outcome, then evidence that only provides diagnostic accuracy outcomes would be considered indirect. If the decision is made to grade the strength of evidence of an intermediate outcome such as diagnostic accuracy, then the reviewer does not need to automatically "downgrade" that outcome for being indirect. Of note, directness does apply to how a test is used in comparison to another test. For example, a study may compare the use of a d-dimer test as a replacement to venous ultrasound for the diagnosis of venous thromboembolism, but in actual practice the relevant question may be the comparison of d-dimer test as a triage for venous ultrasound compared to the use of ultrasound alone. It is worth noting that EPCs consider some aspects of directness separately as described in Chapter 6.[18] Although not included when EPCs assess directness or the strength of evidence, other schemas, such as GRADE, rate directness based on whether the test evaluated is not the exact test used in practice, or if the test accuracy is being calculated in a population or for a use (diagnosis, prognosis, etc.) that is different from the population or use evaluated in the report. Because EPC reports are

intended to be used by a broad spectrum of stakeholders, describing the applicability of the evidence with respect to these factors allows decisionmakers to consider how the evidence relates to their test and population.

Precision refers to the width of confidence intervals for diagnostic accuracy measurements and is integrally related to sample size.[1] Before downgrading the strength of an evidence base for imprecision, reviewers could consider how imprecision for one measure of accuracy may impact clinically meaningful outcomes. This may involve a simple calculation of posttest probabilities over a range of values for sensitivity and specificity, as shown in Table 7–1, or, as illustrated above, a more formal analysis with a decision model.[19] If the impact of imprecision on clinical outcomes is negligible or if the demonstrated precision is sufficient to make the decision, the evidence should not be downgraded.

Principle 4: Additional GRADE domains can be adapted to assess a body of evidence with respect to diagnostic test accuracy.

When grading a body of evidence about a diagnostic test, additional domains should be considered. These additional domains are summarized in Table 7–2.[1–3] These additional domains include *publication bias, dose-response association, existence of plausible unmeasured confounders,* and *strength of association.* Reviewers should comment on *publication bias* when circumstances suggest that negative or no-difference findings have not been published or are unavailable. The *dose-response association* may support an underlying mechanism of detection and potential relevance for some tests that have continuous outcomes and possibly multiple cutoffs (e.g., gene expression, serum PSA [prostate-specific antigen] levels, and ventilation/perfusion scanning). The impact of *plausible unmeasured confounders* may be relevant to testing strategies that predict outcomes. A study may be biased to find low diagnostic accuracy due to spectrum bias and nevertheless have very high diagnostic accuracy. The *strength of association* may be relevant when comparing the accuracy of two different diagnostic tests with one being more accurate than the other.

Principle 5: Multiple domains should be incorporated into an overall assessment in a transparent way.

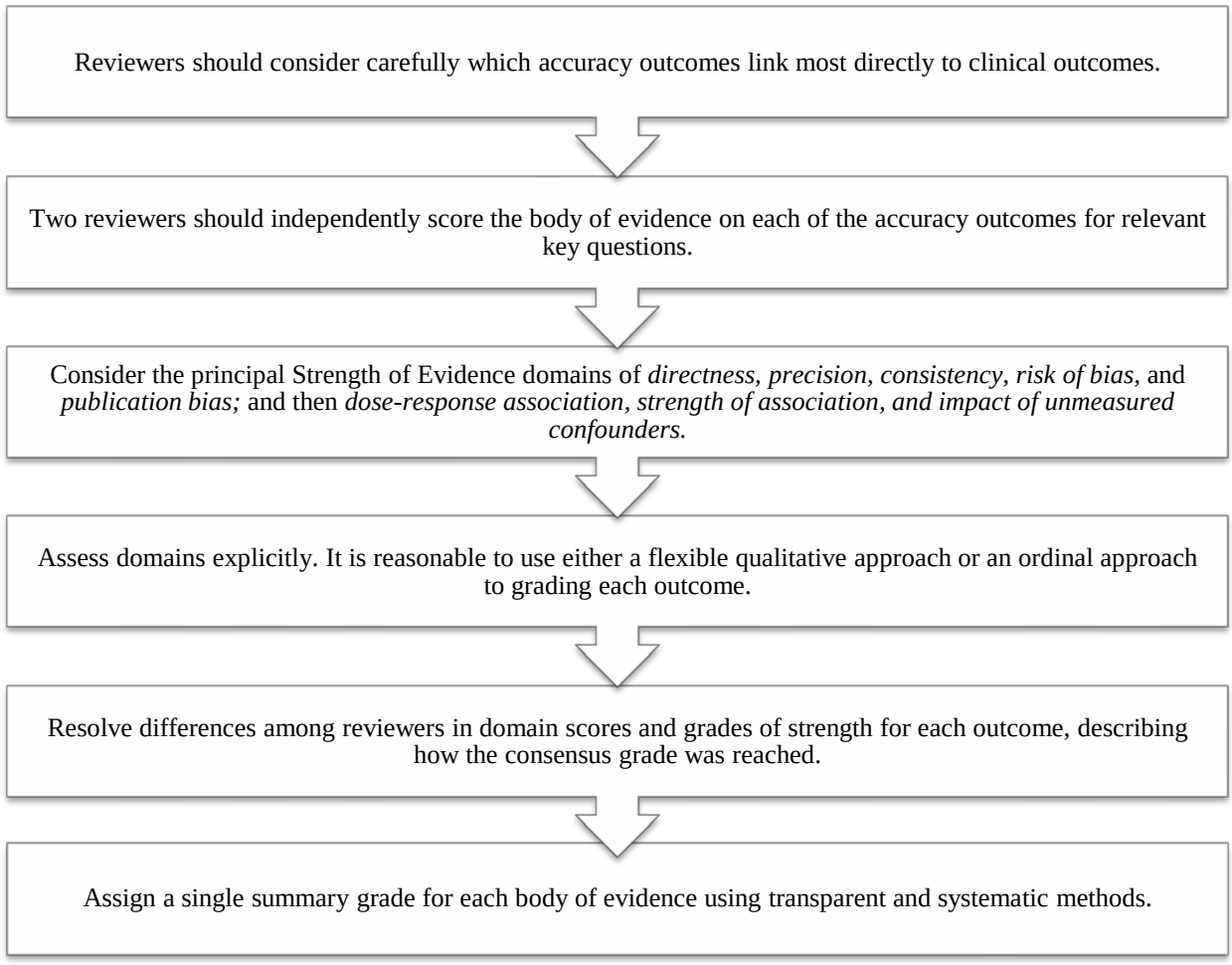
The overall strength of evidence reflects a global assessment of the principle domains and any additional domains, as needed, into an overall summary grade—high, moderate, low, or insufficient evidence. The focus should be on providing an overall grade for the relevant key question link in the analytic chain or for outcomes considered relevant for patients and decisionmakers. These should ideally be identified *a priori*. Consideration should be given on how to incorporate multiple domains into the overall assessment.

There is no empirical evidence to suggest any difference in assigning a summary grade based on qualitative versus quantitative approaches. GRADE advocates an ordinal approach with a ranking from high, moderate, or low, to very low. These "grades" or "overall ratings" are developed using GRADE's eight domains**.** The EPC approach for intervention studies described in the *General Methods Guide*[1,3] allows for more flexibility on grading the strength of evidence. Whichever approach reviewers choose for diagnostic tests, they should consider describing their rationale for which of the required domains were weighted the most in assigning the summary grades.

Illustration

An illustration in Figure 7–1 provides guidance on how reviewers should approach grading a body of evidence on diagnostic test accuracy. This is adapted from the GRADE approach and the EPC *General Methods Guide*. Reviewers should carefully consider which accuracy outcomes are linked to clinical outcomes. In choosing the accuracy outcomes, if the diagnostic test is followed by an invasive procedure, then the number of false positives may be considered most important. However when "diagnostic tests" are used as part of a management strategy, consideration should also be given to grading the positive predictive value and the negative predictive value or likelihood ratios if these additional outcomes assist decisionmakers. An additional example of grading for positive predictive value and negative predictive value is shown in the norovirus table below (Table 7–3). [20–21] This table illustrates how presentation of the same information in different ways can be helpful in considering how to link the accuracy outcomes to clinical practice and projecting how the test would impact clinical outcomes.

Figure 7–1. Steps in grading a body of evidence on diagnostic test accuracy outcomes*



*Adapted from the Methods Guide for Effectiveness and Comparative Effectiveness Reviews.[3]

Table 7–3. Illustration of the approach to grading a body of evidence on diagnostic tests- Identifying norovirus in a health care setting[*]

Outcome	Quantity and Type of Evidence	Findings	Starting GRADE	Decrease GRADE[‡]					GRADE of Evidence for Outcome	Overall GRADE[#]
				Risk of Bias[‡]	Consistency[‡]	Directness[‡]	Precision[‡]	Publication Bias[‡]		
Sensitivity[*]	1 DIAG	68%	High	0	0	0	-1	0	Moderate	Moderate
Specificity[†]	1 DIAG	99%	High	0	0	0	-1	0	Moderate	
PPV[†]	1 DIAG	97%	High	0	0	0	-1	0	Moderate	
NPV[†]	1 DIAG	82%	High	0	0	0	-1	0	Moderate	

DIAG = diagnostic; NPV = negative predictive value; PPV = positive predictive value

[*]Adapted from MacCannell T, et al.[20]

[†]These outcomes were considered the most critical by the guideline developers. From Turcios RM, et al.[21]

[‡]These modifiers can impact the GRADE by 1 or 2 points. From Turcios RM, et al.[21]

[#]Consider the additional domains of strength of association, dose-response and impact of plausible confounders if applicable. From Turcios RM, et al.[21]

Another review of the use of non-invasive imaging in addition to standard workup after recall for evaluation of a breast lesion detected on screening mammography or physical examination illustrates how accuracy does not relate to outcomes when it is being used as part of an algorithm of whether to treat versus watchful waiting.[13] This evidence review focused on the non-invasive imaging studies intended to guide patient management decisions after the discovery of a possible abnormality. The studies were intended to provide additional information to enable women to be appropriately triaged into "biopsy," "watchful waiting," or "return to normal screening intervals" care pathways. Thus the usual strategy of assuming the clinical outcome would be simply a downgrade of the surrogate doesn't always hold true. Reviewers should evaluate the surrogate in the context of the clinical outcome. As the table summary of key findings in the evidence report illustrates, despite the accuracy of the exact diagnosis being low, clinical management may be the same if the post-test probability did not cross a certain decision threshold to alter management decisions.

Two reviewers should independently score these relevant major outcomes and comparisons, within each key question. They should consider the principle domains of *directness*, *precision*, *consistency*, *risk of bias,* and *publication bias*, as well as dose response association, strength of association, and impact of unmeasured confounders. Reviewers should explicitly assess each domain to arrive at a grade for each outcome. Reviewer's choice of various accuracy outcomes to grade may affect how the various domains of *directness*, *precision*, and *consistency* are assessed. This is illustrated in the example by the GRADE working group about multisplice coronary CT scans as compared to coronary angiography.[4] Evidence was considered direct for certain accuracy outcomes such as true positives, true negatives, and false positives since there was little uncertainty about the clinical implications of these results. However, since there was uncertainty about the clinical implications of a false negative test result, this was considered indirect.[4] This resulted in a low strength of evidence grade for false negatives as compared to moderate for other accuracy outcomes.

It is reasonable to consider either a more flexible qualitative approach to grading or the standard ordinal approach ranging from high to insufficient strength of evidence. Reviewers

should resolve differences in domain assessments and grades of outcomes and should describe how the consensus score was reached (e.g., whether by discussion or by third-party adjudication). If appropriate, they should consider arriving at a single summary grade for the diagnostic test through transparent and systematic methods. If reviewers chose to assign an overall summary grade, they should consider the impact of various accuracy outcomes on the overall strength of evidence grade, and should identify which of these accuracy outcomes was considered "key."

Summary

Grading the strength of a body of diagnostic test evidence involves challenges over and above those related to grading the evidence from therapeutic intervention studies. The greatest challenge appears to be assessing multiple links in a chain of evidence connecting the performance of a test to changes in clinical outcomes. In this chapter we focused primarily on grading the body of evidence related to a crucial link in the chain—diagnostic test performance— and described less fully the challenges involved in assessing other links in the chain.

No one system for grading the strength of evidence for diagnostic tests has been shown to be superior to any other and many are still early in development. However, we conclude that, in the interim, applying the consistent and transparent system of grading using the domains described above, and giving an explicit rationale for the choice of grades based on these domains, will make EPC reports and other reports on diagnostic tests more useful for decisionmakers.

Key Points

- One can use GRADE for diagnostic tests. The outcomes one should consider are the clinical outcomes of effectiveness or harm if available for diagnostic tests.
- When intermediate accuracy outcomes are used, an analytic framework should describe how the test is related to clinical outcomes, and should then delineate the individual questions that can be answered in that framework.
- Selection of accuracy outcomes (i.e. sensitivity, specificity, positive predictive value and negative predictive value, true positives, true negatives, false positives, and false negatives) and needed levels of precision of these quantities should be based on consideration of how the test is to be used in the clinical context.
- Domains of risk of bias, directness, consistency, precision, publication bias, dose response association, and plausible unmeasured confounders can be used to grade the strength of evidence for the effect of a diagnostic test on clinical outcomes or on intermediate surrogate outcomes if selected by the EPC and key informants.
- Whether reviewers choose a qualitative or quantitative approach to combining domains into a single grade, they should consider explaining their rationale for a particular summary grade and the relevant domains that were most heavily weighted in assigning the summary grade.

References

1. Owens DK, Lohr KN, Atkins D, et al. AHRQ series paper 5: Grading the strength of a body of evidence when comparing medical interventions--Agency for Healthcare Research and Quality and the Effective Health-Care Program. J Clin Epidemiol 2010;63(5):513-23.
2. Atkins D, Fink K, Slutsky J. Better information for better health care: the Evidence-based Practice Center program and the Agency for Healthcare Research and Quality. Ann Intern Med 2005; 142(12 Pt 2):1035-41.
3. Agency for Healthcare Research and Quality. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville, MD: Agency for Healthcare Research and Quality.; 2008–. http://www.ncbi.nlm.nih.gov/books/NBK47095. Accessed April 2012
4. Schunemann HJ, Oxman AD, Brozek J, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008; 336(7653):1106-10.
5. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336(7650): 924-6.
6. Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S, Guyatt GH. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011; 64(4):401-6.
7. Lohr KN, Carey TS. Assessing "best evidence": issues in grading the quality of studies for systematic reviews. Jt Comm J Qual Improv 1999; 25:470-9.
8. Samson D, Schoelles KM. Developing the topic and structuring systematic reviews of medical tests: utility of PICOTS, analytic frameworks, decision trees, and other frameworks. AHRQ Publication No. 12-EHC074-EF. Chapter 2 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.
9. Marchionni L, Wilson RF, Marinopoulos SS, et al. Impact of Gene Expression Profiling Tests on Breast Cancer Outcomes. Evidence Report/Technology Assessment No. 160. (Prepared by The Johns Hopkins University Evidence-based Practice Center under contract No. 290-02-0018). AHRQ Publication No. 08-E002. Rockville, MD: Agency for Healthcare Research and Quality. January 2008. Available at: www.ahrq.gov/downloads/pub/evidence/pdf/brcancergene/brcangene.pdf. Accessed December 2011.
10. Ross SD, Allen IE, Harrison KJ, et al. Systematic Review of the Literature Regarding the Diagnosis of Sleep Apnea. Evidence Report/Technology Assessment No. 1. (Prepared by MetaWorks Inc. under Contract No. 290-97-0016.) AHCPR Publication No. 99-E002. Rockville, MD: Agency for Health Care Policy and Research; February 1999. Available at: http://www.ncbi.nlm.nih.gov/books/NBK32884/. Accessed December 2011.
11. Bruening W, Schoelles K, Treadwell J, et al. Comparative Effectiveness of Core-Needle and Open Surgical Biopsy for the Diagnosis of Breast Lesions. Comparative Effectiveness Review No. 19. (Prepared by ECRI Institute Evidence-based Practice Center under Contract No. 290-02-0019.) Rockville, MD: Agency for Healthcare Research and Quality. December 2009. Available at: http://effectivehealthcare.ahrq.gov/ehc/products/17/370/finalbodyforposting.pdf. Accessed December 2011.

12. Segal JB, Brotman DJ, Emadi A, et al. Outcomes of Genetic Testing in Adults with a History of Venous Thromboembolism. Evidence Report/Technology Assessment No. 180. (Prepared by Johns Hopkins University Evidence-based Practice Center under Contract No. HHSA 290-2007-10061-I). AHRQ Publication No. 09-E011. Rockville, MD. Agency for Healthcare Research and Quality. June 2009. Available at: http://www.ahrq.gov/downloads/pub/evidence/pdf/factorvleiden/fvl.pdf. Accessed December 2011.

13. Bruening W, Uhl S, Fontanarosa J, Schoelles K. Noninvasive Diagnostic Tests for Breast Abnormalities: Update of a 2006 Review. Comparative Effectiveness Review No. 47 (Prepared by the ECRI Institute Evidence-based Practice Center under Contract No. 290-02-0019.) AHRQ Publication No. 12-EHC014-EF. Rockville, MD: Agency for Healthcare Research and Quality; February 2012. http://www.ncbi.nlm.nih.gov/books/NBK84530. PMID: 22420009.

14. Segal JB. Choosing the important outcomes for a systematic review of a medical test. AHRQ Publication No. 12-ECH075-EF. Chapter 3 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published in a special supplement to the Journal of General Internal Medicine, June 2012.

15. Lord SJ, Irwig L, Simes J. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need a randomized trial? Ann Intern Med 2006; 144(11):850-5.

16. Pauker SG, Kassirer JP.The threshold approach to clinical decision making.N Engl J Med. 1980 May 15;302(20):1109-17.

17. Santaguida PL, Riley CM, Matchar DB. Assessing risk of bias as a domain of quality in medical test studies. AHRQ Publication No. 12-EHC077-EF. Chapter 5 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.

18. Hartmann KE, Matchar DB, Chang S. Assessing applicability of medical test studies in systematic reviews. AHRQ Publication No. 12-ECH078-EF. Chapter 6 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.

19. Trikalinos TA, Kulasingam S, Lawrence WF. Deciding whether to complement a systematic review of medical tests with decision modeling. AHRQ Publication No. 12-EHC082-EF. Chapter 10 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.

20. MacCannell T, Umscheid CA, Agarwal RK, Lee I, Kuntz G, Stevenson, KB, and the Healthcare Infection Control Practices Advisory Committee. Guideline for the prevention and control of norovirus gastroenteritis outbreaks in healthcare settings. Infection Control and Hospital Epidemiology. 2011; 32(10): 939-969.

21. Turcios RM, Widdowson MA, Sulka AC, Mead PS, Glass RI. Reevaluation of epidemiological criteria for identifying outbreaks of acute gastroenteritis due to norovirus: United States, 1998-2000.Clin Infect Dis. 2006;42(7):964-9.

Acknowledgments: The authors would like to acknowledge the contribution of Drs. Mark Helfand (Oregon Evidence-based Practice Center), Joseph Lau (Tufts Evidence-based Practice Center), Jonathan Treadwell (ECRI Institute Evidence-based Practice Center), Kathleen N. Lohr (RTI International), and Douglas K. Owens (Stanford University) for providing comments on a draft of the manuscript.

Funding: Funded by the Agency for Health Care Research and Quality (AHRQ) under the Effective Health Care Program.

Conflict of interest: None of the authors has any affiliations or financial involvements that conflicts with the information presented in this chapter.

Corresponding author: Sonal Singh, M.D., M.P.H., Department of Medicine, Johns Hopkins University School of Medicine, 624 N. Broadway 680 B, Baltimore, MD, 21287 USA. Telephone 410-955-9869; Fax 410-955- 0825; email ssingh31@jhu.edu.

Suggested citation: Singh S, Chang S, Matchar DB, Bass EB. Grading a body of evidence on diagnostic tests. AHRQ Publication No. 12-EHC079-EF. Chapter 7 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published as a special supplement to the Journal of General Internal Medicine, July 2012.

Chapter 8

Meta-Analysis of Test Performance When There Is a "Gold Standard"

Thomas A. Trikalinos, M.D., Tufts Evidence-based Practice Center, Boston, MA
Cynthia M. Balion, Ph.D., Hamilton General Hospital and McMaster Evidence-based Practice Center, Hamilton, Ontario, CA
Craig I. Coleman, Pharm.D. University of Connecticut/Hartford Hospital Evidence-based Practice Center, and University of Connecticut School of Pharmacy, Storrs, CT
Lauren Griffith, Ph.D., McMaster Evidence-based Practice Center
P. Lina Santaguida, B.Sc.P.T., Ph.D., M.Sc., McMaster Evidence-based Practice Center
Ben Vandermeer, M.Sc., University of Alberta Evidence-based Practice Center, Edmonton, Alberta, CA
Rongwei Fu, Ph.D., Oregon Evidence-based Practice Center, Portland, OR

Abstract

Synthesizing information on test performance metrics, such as sensitivity, specificity, predictive values and likelihood ratios, is often an important part of a systematic review of a medical test. Because many metrics of test performance are of interest, the meta-analysis of medical tests is more complex than the meta-analysis of interventions or associations. Sometimes, a helpful way to summarize medical test studies is to provide a "summary point," a summary sensitivity and a summary specificity. Other times, when the sensitivity or specificity estimates vary widely or when the test threshold varies, it is more helpful to synthesize data using a "summary line" that describes how the average sensitivity changes with the average specificity. The choice of the most helpful summary is subjective, and in some cases both summaries provide meaningful and complementary information. Because sensitivity and specificity are not independent across studies, the meta-analysis of medical tests is fundamentaly a multivariate problem, and should be addressed with multivariate methods. More complex analyses are needed if studies report results at multiple thresholds for positive tests. At the same time, quantitative analyses are used to explore and explain any observed dissimilarity (heterogeneity) in the results of the examined studies. This can be performed in the context of proper (multivariate) meta-regressions.

Introduction

The *Methods Guide to Medical Test Reviews* (also referred to as the Medical Test Methods Guide) highlights common challenges in systematic reviews of medical tests and outlines their mitigation, as perceived by researchers partaking in the Agency for Healthcare Research and Quality (AHRQ) Effective Healthcare Program.[1] Generic by their very nature, these challenges

and their discussion apply to the larger set of systematic reviews of medical tests, and are not specific to AHRQ's program.

This chapter of the *Medical Test Methods Guide* focuses on choosing strategies for meta-analysis of test "accuracy," or more preferably, test performance. Meta-analysis is not required for a systematic review, but when appropriate, it should be undertaken with a dual goal: to provide summary estimates for key quantities, and to explore and explain any observed dissimilarity (heterogeneity) in the results of the examined studies.

"Summing-up" information on test performance metrics such as sensitivity, specificity, and predictive values is rarely the most informative part of a systematic review of a medical test.[2–5] Key clinical questions driving the evidence synthesis (e.g., "Is this test alone or in combination with a test-and-treat strategy likely to improve decisionmaking and patient outcomes?") are only indirectly related to test performance *per se*. Formulating an effective evaluation approach requires careful consideration of the context in which the test will be used. These framing issues are addressed in other chapters of this *Medical Test Methods Guide*.[6–8] Further, in this chapter we assume that medical test performance has been measured against a "gold standard," that is, a reference standard considered adequate to define the presence or absence of the condition of interest. Another chapter discusses ways to summarize medical tests when such a reference standard does not exist.[9]

Syntheses of medical test data often focus on test performance, and much of the attention to statistical issues relevant to synthesizing medical test evidence focuses on summarizing test performance data; thus their meta-analysis was chosen to be the focus of this paper. We will assume that the decision to perform meta-analyses of test performance data is justified and taken, and will explore two central challenges; namely, how do we quantitatively summarize medical test performance when: (1) the sensitivity and specificity estimates of various studies do not vary widely, or (2) the sensitivity and specificity of various studies vary over a large range.

Briefly, in the first situation, it may be helpful to use a "summary point" (a summary sensitivity and summary specificity pair) to obtain summary test performance when sensitivity and specificity estimates do not vary widely across studies. This could happen in meta-analyses where all studies have the same explicit test positivity threshold (a threshold for categorizing the results of testing as positive or negative), since if studies have different explicit thresholds, the clinical interpretation of a summary point is less obvious, and perhaps less helpful. However, an explicit common threshold is neither sufficient nor necessary for opting to synthesize data with a "summary point"; a summary point can be appropriate whenever sensitivity and specificity estimates do not vary widely across studies.

In the second type of situation, when the sensitivity and specificity of various studies vary over a large range, rather than using a "summary point" it may be more helpful to describe how the average sensitivity and average specificity relate by means of a "summary line." This oft-encountered situation can be secondary to explicit or implicit variation in the threshold for a "positive" test result; heterogeneity in populations, reference standards, or the index tests; variation in study design; chance; or bias.

Of note, in many applications it may be informative to present syntheses in both ways, as the two modes convey complementary information.

Deciding whether a "summary point" or a "summary line" is more helpful as a synthesis is subjective, and no hard-and-fast rules exist. We briefly outline common approaches for meta-analyzing medical tests, and discuss principles for choosing between them. However, a detailed presentation of methods or their practical application is outside the scope of this work. In

addition, it is expected that readers are versed in clinical research methodology and familiar with methodological issues pertinent to the study of medical tests. We also assume familiarity with the common measures of medical test performance (reviewed in the Appendix, and in excellent introductory papers[10]). For example, we do not review challenges posed by methodological or reporting shortcomings of test performance studies.[11] The Standards for Reporting of Diagnostic accuracy (STARD) initiative published a 25-item checklist that aims to improve reporting of medical tests studies.[11] We refer readers to other papers in this issue[12] and to several methodological and empirical explorations of bias and heterogeneity in medical test studies.[13–15]

Nonindependence of Sensitivity and Specificity Across Studies and Why It Matters for Meta-Analysis

In a typical meta-analysis of test performance, we have estimates of sensitivity and specificity for each study, and seek to provide a meaningful summary across all studies. *Within each study*, sensitivity and specificity are independent, because they are estimated from different patients (sensitivity from those with the condition of interest, and specificity from those without). According to the prevailing reasoning, *across studies* sensitivity and specificity are likely negatively correlated: as one estimate increases the other is expected to decrease. This is perhaps more obvious when studies have different explicit thresholds for "positive" tests; and thus the term "threshold effect" has been used to describe this negative correlation. For example, the D-dimer concentration threshold for diagnosing an acute coronary event can vary from approximately 200 to over 600 ng/mL.[16] It is expected that higher thresholds would correspond to generally lower sensitivity but higher specificity, and the opposite for lower thresholds (though in this example it is not clearly evident; see Figure 8–1, especially Figure 8–1a). A similar rationale can be invoked to explain between-study variability for tests with more implicit or suggestive thresholds, such as imaging or histological tests.

Negative correlation between sensitivity and specificity across studies may be expected for reasons unrelated to thresholds for positive tests. For example, in a meta-analysis evaluating the ability of serial creatine kinase-MB (CK-MB) measurements to diagnose acute cardiac ischemia in the emergency department,[17,18] the time interval from the onset of symptoms to serial CK-MB measurements (rather than the actual threshold for CK-MB) could explain the relationship between sensitivity and specificity across studies. The larger the time interval, the more CK-MB is released into the bloodstream, affecting the estimated sensitivity and specificity. Unfortunately, the term "threshold effect" is often used rather loosely to describe the relationship between sensitivity and specificity across studies, even when, strictly speaking, there is no direct evidence of variability in study thresholds for positive tests.

Because of the above, the current thinking is that, in general, the study estimates of sensitivity and specificity do not vary independently, but jointly, and likely with a negative correlation. Summarizing the two correlated quantities is a multivariate problem, and multivariate methods should be used to address it, as they are more theoretically motivated.[19,20] At the same time there are situations when a multivariate approach is not practically different from separate univariate analyses. We will expand on some of these issues.

Figure 8–1a. Typical data on the performance of a medical test (D-dimers for venous thromboembolism)

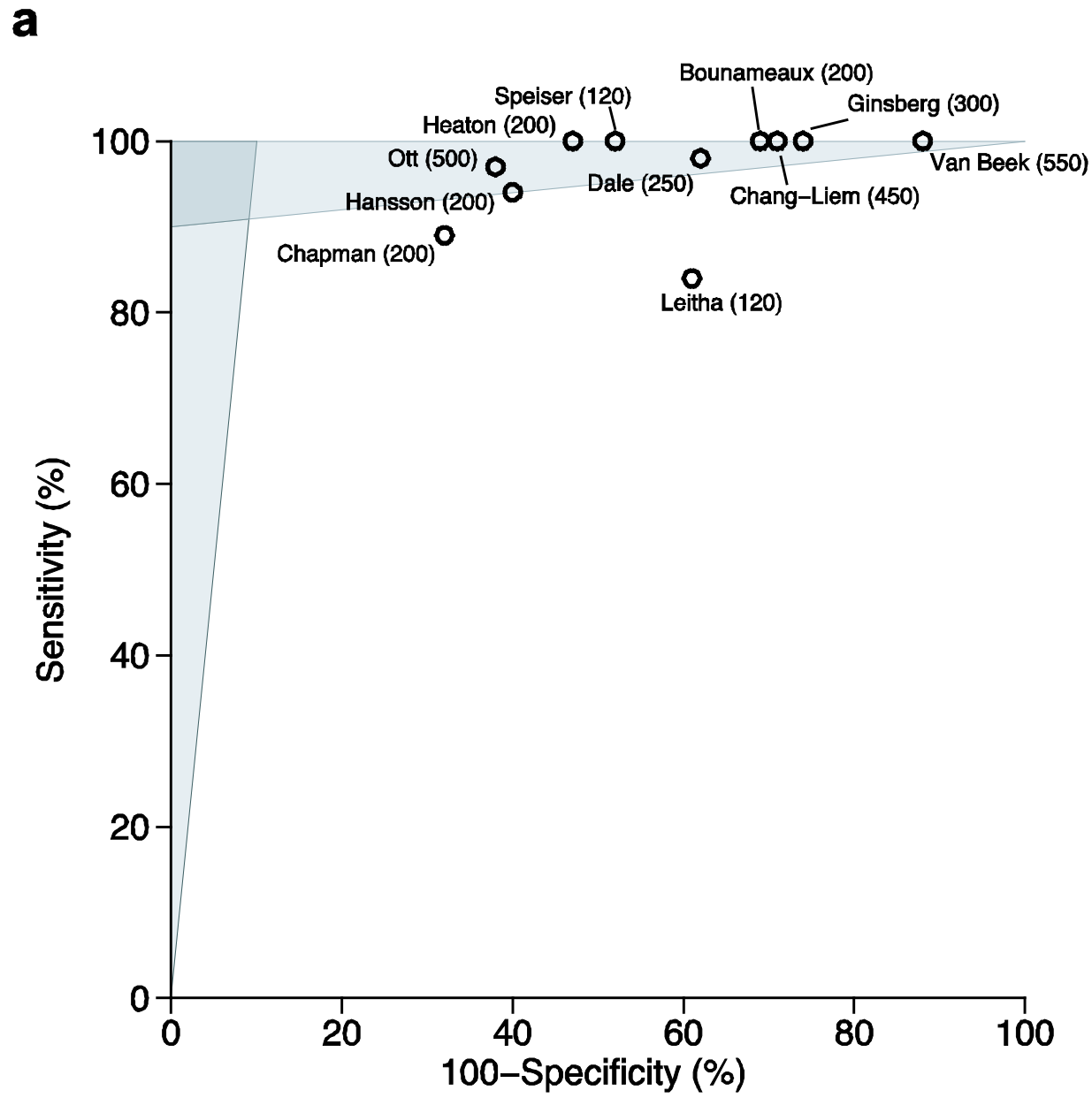
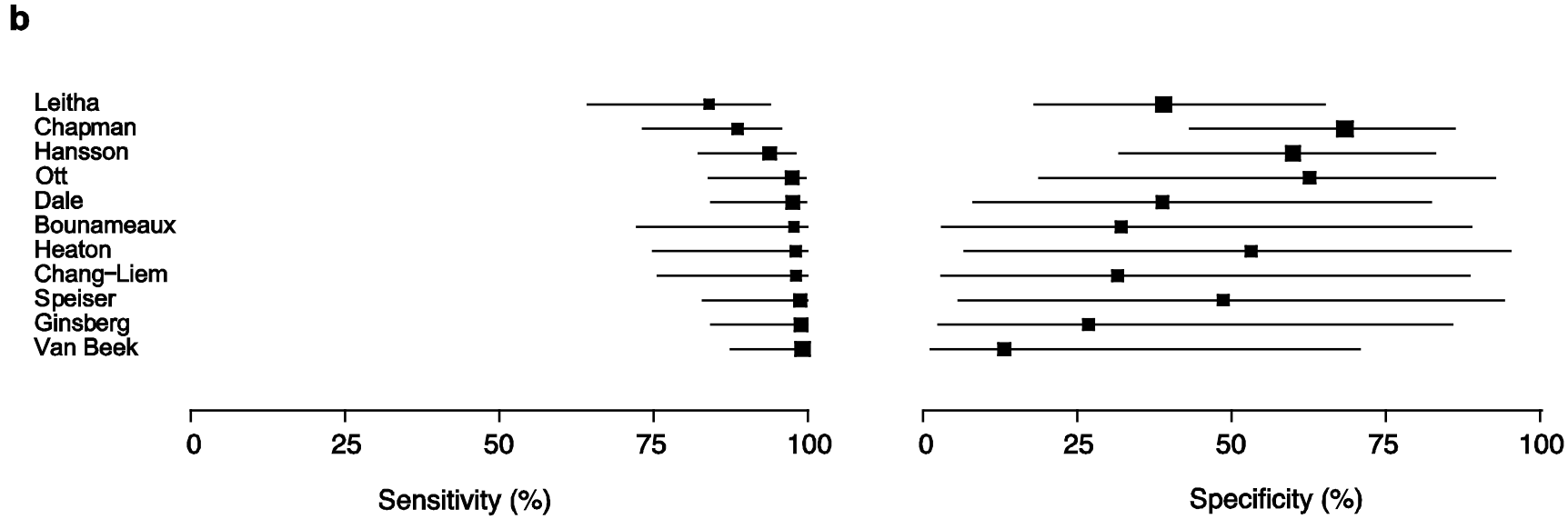


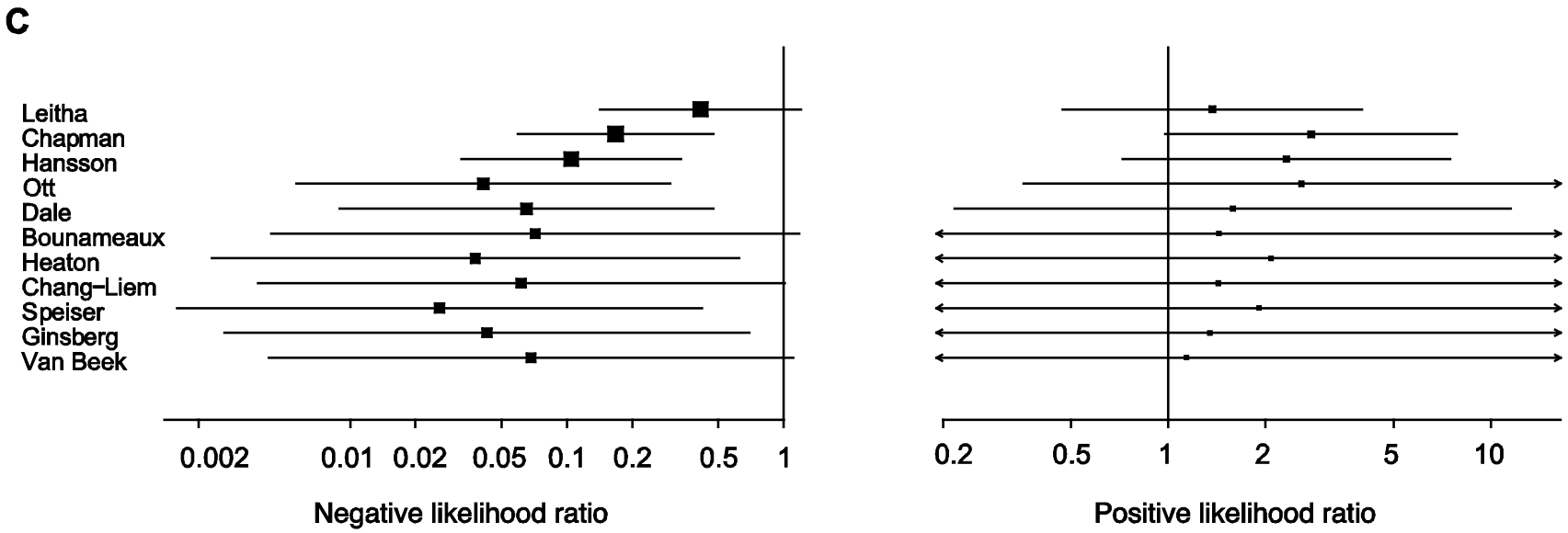
Figure 8–1a depicts studies as markers, labeled by author names and thresholds for a positive test (in ng/mL). Studies listed on the left lightly shaded area have a positive likelihood ratio of at least 10. Studies listed on the top lightly shaded area have a negative likelihood ratio of at most 0.1. Studies listed at the intersection of the gray areas (darker gray polygon) have both a positive likelihood ratio of at least 10 and a negative likelihood ratio of 0.1 or less.

Figure 8–1b. Typical data on the performance of a medical test (D-dimers for venous thromboembolism)



Eleven studies on ELISA-based D-dimer assays for the diagnosis of venous thromboembolism.[16]
Figure 8–1b shows "paired" forest plots in ascending order of sensitivity (left) along with the corresponding specificity (right). Note how sensitivity increases with decreasing specificity, which could be explained by a "threshold effect."

Figure 8–1c. Typical data on the performance of a medical test (D-dimers for venous thromboembolism)



Eleven studies on ELISA-based D-dimer assays for the diagnosis of venous thromboembolism.[16] Figure 8–1c shows the respective negative and positive likelihood ratios.

Principles for Addressing the Challenges

To motivate our suggestions on meta-analyses of medical tests, we invoke two general principles:

- Principle 1: Favor the most informative way to summarize the data. Here we refer mainly to choosing between a summary point and a summary line, or choosing to use both.
- Principle 2: Explore the variability in study results with graphs and suitable analyses, rather than relying exclusively on "grand means."

Recommended Approaches

Which Metrics to Meta-Analyze

For each study, the estimates of sensitivity, specificity, predicitive values, likelihood ratios, and prevalence are related through simple formulas (Appendix). However, if one performs a meta-analysis for each of these metrics, the summaries across all studies will generally be inconsistent: the formulas will not be satisfied *for the summary estimates*. To avoid this, we propose to obtain summaries for sensitivities and specificities via meta-analysis, and to back-calculate the overall predictive values or likelihood ratios from the formulas in the Appendix, for a range of plausible prevalences. Figure 8–2 illustrates this strategy for a meta-analysis of *K* studies. We explain the rationale below.

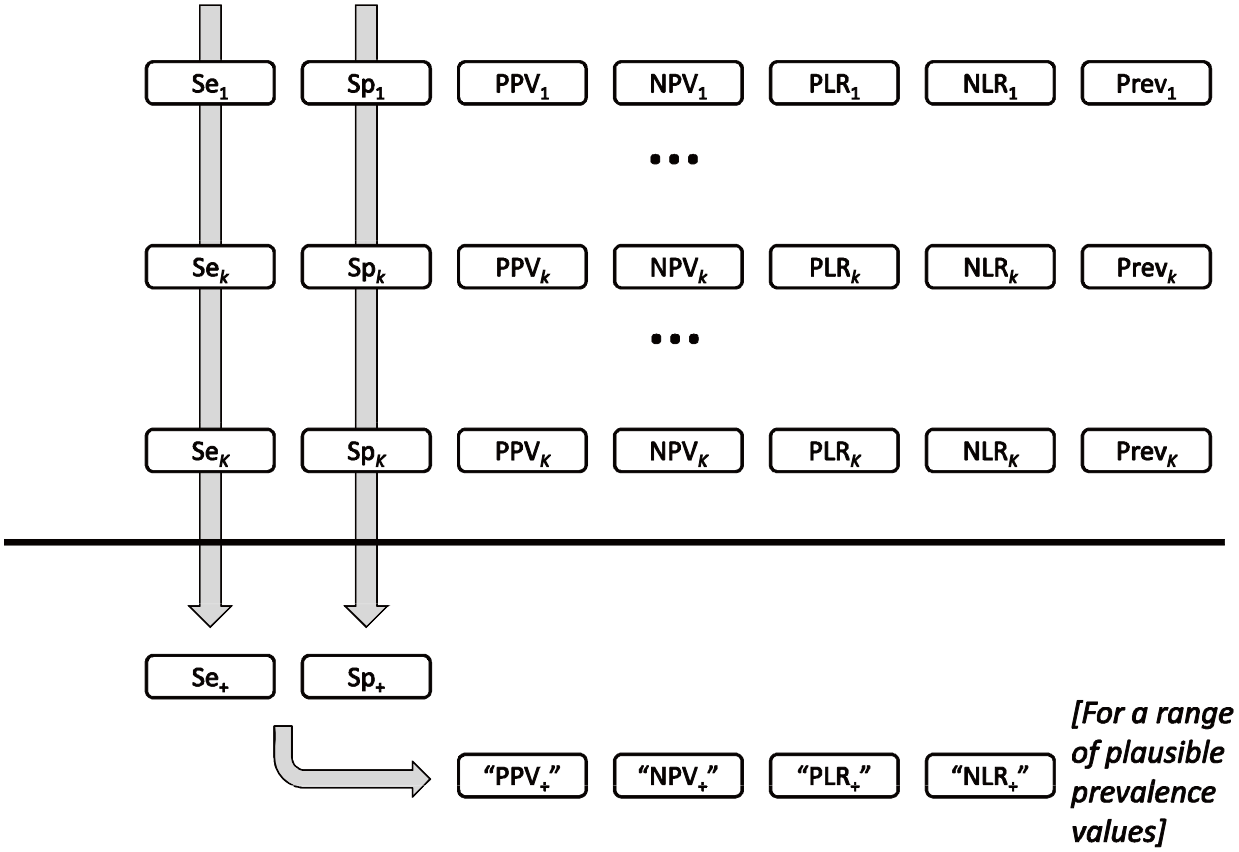
Why *does* it make sense to directly meta-analyze sensitivity and specificity?

Summarizing studies with respect to sensitivity and specificity aligns well with our understanding of the effect of positivity thresholds for diagnostic tests. Further, sensitivity and specificity are often considered independent of the prevalence of the condition under study (though this is an oversimplification that merits deeper discussion[21]). The summary sensitivity and specificity obtained by a direct meta-analysis will always be between zero and one. Because these two metrics do not have as intuitive an interpretation as likelihood ratios or predictive values,[10] we can use formulas in the Appendix to back-calculate "summary" (overall) predictive values and likelihood ratios that correspond to the summary sensitivity and specificity for a range of plausible prevalence values.

Why does it *not* make sense to directly meta-analyze positive and negative predictive values or prevalence?

Predictive values are dependent on prevalence estimates. Because prevalence is often wide ranging, and because many medical test studies have a case-control design (where prevalence cannot be estimated), it is rarely meaningful to directly combine these across studies. Instead, predictive values can be calculated as mentioned above from the summary sensitivity and specificity for a range of plausible prevalence values.

Figure 8–2. Obtaining summary (overall) metrics for medical test performance



PLR/NLR = positive (negative) likelihood ratio; PPV/NPV = positive (negative) predictive value; Prev = prevalence; Se = Sensitivity; Sp = specificity

The herein recommended approach is to perform a meta-analysis for sensitivity and specificity across the *K* studies, and then use the summary sensitivity and specificity (Se_+ and Sp_+; a row of two boxes after the horizontal black line) to back-calculate "overall" values for the other metrics (second row of boxes after the horizontal black line). In most cases it is not meaningful to synthesize prevalences (see text).

Why could directly meta-analyzing likelihood ratios be problematic?

Positive and negative likelihood ratios could also be combined in the absence of threshold variation, and in fact, many authors give explicit guidance to that effect.[22] However, this practice does not guarantee that the summary positive and negative likelihood ratios are internally consistent. Specifically, it is possible to get summary likelihood ratios that correspond to impossible summary sensitivities or specificities (outside the zero to one interval).[23] Back-calculating the summary likelihood ratios from summary sensitivities and specificities avoids this complication. Nevertheless, these aberrant cases are not common,[24] and calculations of summary likelihood ratios by directly meta-analyzing them or from back calculation of the summary sensitivity and specificity rarely results in different conclusions.[24]

What should be said about directly meta-analyzing diagnostic odds ratios?

The synthesis of diagnostic odds ratios is straightforward and follows standard meta-analysis methods.[25,26] The diagnostic odds ratio is closely linked to sensitivity, specificity, and likelihood ratios, and it can be easily included in meta-regression models to explore the impact of

explanatory variables on between-study heterogeneity. Apart from challenges in interpreting diagnostic odds ratios, a disadvantage is that it is impossible to weight the true positive and false positive rates separately.

Desired Characteristics of Meta-Analysis Methods

Over several decades many methods have been used for meta-analyzing medical test performance data. Based on the above considerations, methods should be motivated by (1) respecting the multivariate nature of test performance metrics (i.e., sensitivity and specificity); (2) allowing for the nonindependence between sensitivity and specificity across studies ("threshold effect") and (3) allowing for between-study heterogeneity. Table 8–1 lists commonly used methods for meta-analysis of medical tests. The most theoretically motivated meta-analysis approaches are based on multivariate methods (hierarchical modeling).

Table 8–1. Commonly used methods for meta-analysis of medical test performance

Method	Description or Comment	Does it have Desired Characteristics?
	Summary Point	
Independent meta-analysis of sensitivity and specificity	• Entails separate meta-analyses per metric. • Within-study variability is preferably modeled by the binomial distribution.[46]	• Ignores correlation between sensitivity and specificity. • Yields underestimates of summary sensitivity and specificity and incorrect confidence intervals.[27]
Joint (multivariate) meta-analysis of sensitivity and specificity based on hierarchical modeling	• Based on multivariate (joint) modeling of sensitivity and specificity. • Two families of models[27;31] (see text), are equivalent when there are no covariates.[19] • Modeling should preferably use binomial likelihood rather than normal approximations.[31,38,47,48]	• The generally preferred method
	Summary Line	
Moses and Littenberg model	• Summary line based on a simple regression of the difference of logit-transformed true and false positive rates versus their average.[33-35]	• Ignores unexplained variation between studies (fixed effects). • Does not account for correlation between sensitivity and specificity. • Does not account for variability in the independent variable. • Unable to weight studies optimally—yields wrong inferences when covariates are used.
Random intercept augmentation of the Moses-Littenberg model	• Regression of the difference of logit-transformed true and false positive rates versus their average with random effects to allow for variability across studies.[36–37]	• Does not account for correlation between sensitivity and specificity. • Does not account for variability in the independent variable.
Summary ROC based on hierarchical modeling	• Same as for multivariate meta-analysis to obtain a summary point – hierarchical modeling.[27,31] • Many ways to obtain a (hierarchical) summary ROC : o Rutter-Gatsonis (most common)[31] o Several alternative curves[38,39]	• Most theoretically motivated method • Rutter-Gatsonis HSROC recommended in the Cochrane handbook,[48] as it is the method with which there is most experience.

HSROC = hierarchical summary ROC; ROC = receiver operating characteristic

We will focus on the case where each study reports a single pair of sensitivity and specificity at a given threshold (although thresholds can differ across studies). Another, more complex situation arises when multiple sensitivity and specificity pairs (at different thresholds) are reported in each study. Statistical models for the latter case exist, but there is less empirical evidence on their use. These will be described briefly, as a special case.

Preferred Methods for Obtaining a "Summary Point" (Summary Sensitivity and Specificity): Two Families of Hierarchical Models

When a "summary point" is deemed a helpful summary of a collection of studies, one should ideally perform a *multivariate meta-analysis* of sensitivity and specificity, i.e., a joint analysis of both quantities, rather than separate univariate meta-analyses. This is not only theoretically motivated,[27–29] but also corroborated by simulation analyses as well.[2,28,30]

Multivariate meta-analyses require advanced hierarchical modeling. We can group the commonly used hierarchical models in two families: The so-called "bivariate model"[27] and the "hierarchical summary receiver operating characteristic " (HSROC) model.[31] Both use two levels to model the statistical distributions of data. At the first level, they model the counts of the 2×2 table within each study, which accounts for within-study variability. At the second level, they model the between-study variability (heterogeneity), allowing for the theoretically expected nonindependence of sensitivity and specificity across studies. The two families differ in their parameterization at this second level: the bivariate model uses parameters that are transformations of the average sensitivity and specificity, while the HSROC model uses a scale parameter and an accuracy parameter, which are functions of sensitivity and specificity and define an underlying hierarchical summary ROC (receiver operating characteristic) curve.

In the absence of covariates, the two families of hierarchical models are mathematically equivalent; one can use simple formulas to relate the fitted parameters of the bivariate model to the HSROC model and vice versa, rendering choices between the two approaches moot.[19] The importance of choosing between the two families becomes evident in meta-regression analyses, when covariates are used to explore between-study heterogeneity. The differences in design and conduct of the included diagnostic accuracy studies may affect the choice of the model.[19] For example, "spectrum effects," where the subjects included in a study are not representative of the patients who will receive the test in practice,[32] "might be expected to impact test accuracy rather than the threshold, and might therefore be most appropriately investigated using the HSROC approach. Conversely, between-study variation in disease severity will (likely) affect sensitivity but not specificity, leading to a preference for the bivariate approach."[19] When there are covariates in the model, the HSROC model allows direct evaluation of the difference in accuracy or threshold parameters or both. The accuracy parameter affects how much higher the summary line is from the diagonal (the line of no diagnostic information), and the threshold parameter affects the degree of asymmetry of the HSROC curve.[19] Bivariate models, on the other hand, allow for direct evaluation of covariates on sensitivity or specificity or both. Systematic reviewers are encouraged to look at study characteristics and think through how study characteristics could affect the diagnostic accuracy, which in turn might affect the choice of the meta-regression model.

Preferred Methods for Obtaining a "Summary Line"

When a summary line is deemed more helpful in summarizing the available studies, we recommend summary lines obtained from hierarchical modeling, instead of several simpler approaches (8–1).[33–37] As mentioned above, when there are no covariates, the parameters of hierarchical summary lines can be calculated from the parameters of the bivariate random effects models using formulas.[19,31,38] In fact, a whole range of HSROC lines can be constructed using parameters from the fitted bivariate model;[38,39] one proposed by Rutter and Gatsonis[31] is an example. The various HSROC curves represent alternative characterizations of the bivariate distribution of sensitivity and specificity, and can thus have different shapes. Briefly, apart from the commonly used Rutter-Gatsonis HSROC curve, alternative curves include those obtained from a regression of logit-transformed true positive rate on logit-transformed false positive rate; logit false positive rate on logit true positive rate; or the major axis regression between logit true and false positive rates.[38,39]

When the estimated correlation between sensitivity and specificity is positive (as opposed to the typical negative correlation), the latter three alternative models can generate curves that follow a downward slope from left to right. This is not as rare as once thought[38]—a downward slope (from left to right) was observed in approximately one out of three meta-analyses in a large empirical exploration of 308 meta-analyses.[40] Chappell et al. argued that in meta-analyses with evidence of positive estimated correlation between sensitivity and specificity (e.g., based on the correlation estimate and confidence interval or its posterior distribution) it is meaningless to use an HSROC line to summarize the studies,[39] as a "threshold effect" explanation is not possible. Yet, even if the estimated correlation between sensitivity and specificity is positive (i.e., not in the "expected" direction), an HSROC still represents how the summary sensitivity changes with the summary specificity. The difference is that the explanation for the pattern of the studies cannot involve a "threshold effect"; rather, it is likely that an important covariate has not been included in the analysis (see the proposed algorithm below).[39]

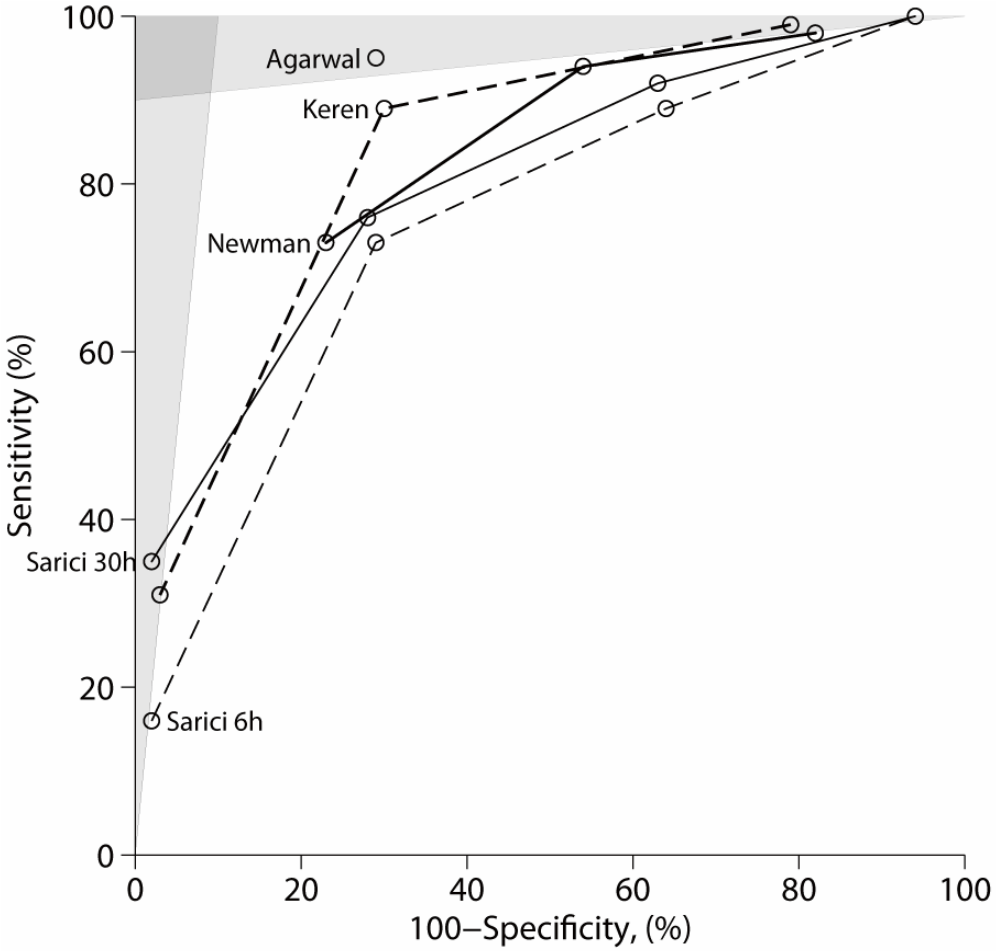
A Special Case: Joint Analysis of Sensitivity and Specificity When Studies Report Multiple Thresholds

It is not uncommon for some studies to report multiple sensitivity and specificity pairs at several thresholds for positive tests. One option is to decide on a single threshold from each study and apply the aforementioned methods. To some extent, the setting in which the test is used can guide the selection of the threshold. For example, in some cases, the threshold which gives the highest sensitivity may be appropriate in medical tests to rule out disease. Another option *is to use all available thresholds per study*. Specifically, Dukic and Gatsonis extended the HSROC model to analyze sensitivity and specificity data reported at more than one threshold.[41] This model represents an extension of the HSROC model discussed above. Further, if each study reports enough data on sensitivity and specificity to construct a ROC curve, Kester and Buntinx[42] proposed a little-used method to combine whole ROC curves.

Both models are theoretically motivated. The Dukic and Gatsonis model is more elaborate and more technical in its implementation than the Kester and Buntinx variant. There is no empirical evidence on the performance of either model in a large number of applied examples. Therefore, we refrain from providing a strong recommendation to always perform such analyses. Systematic reviewers are mildly encouraged to perform explorations, including analyses with these models. Should they opt to do so, they should provide adequate description of the

employed models and their assumptions, as well as a clear intuitive interpretation of the parameters of interest in the models. At a minimum, we suggest that systematic reviewers perform explorations in a qualitative, graphical depiction of the data in the ROC space (see the Algorithm section). This will provide a qualitative summary and will highlight similarities and differences among the studies. An example of such a graph is Figure 8–3, which illustrates the diagnostic performance of early measurements of total serum bilirubin (TSB) to identify post-discharge TSB above the 95th hour-specific percentile in newborns.[43]

Figure 8–3. Graphical presentation of studies reporting data at multiple thresholds



Ability of early total serum bilirubin measurements to identify postdischarge total serum bilirubin above the 95th hour-specific percentile. Sensitivity and 100 percent minus specificity pairs from the same study (obtained with different cut-offs for the early total serum bilirubin measurement) are connected with lines. These lines are reconstructed based on the reported cut-offs, and are not perfect representations of the actual ROC curves in each study (they show only a few thresholds that could be extracted from the study). Studies listed on the left lightly shaded area have a positive likelihood ratio of at least 10. Studies listed on the top lightly shaded area have a negative likelihood ratio of at most 0.1. Studies listed at the intersection of the gray areas (darker gray polygon) have both a positive likelihood ratio of at least 10 and a negative likelihood ratio of 0.1 or less.[43]

A Workable Algorithm

We propose using the following three-step algorithm for meta-analyzing studies of medical test performance when there is a "gold standard." This algorithm should assist meta-analysts in deciding whether a summary point, a summary line, or both are helpful syntheses of the data. When reviewing the three step algorithm, keep these points in mind:

- A summary point may be less helpful or interpretable when the studies have different explicit thresholds for positive tests, and when the estimates of sensitivity vary widely along different specificities. In such cases, a summary line may be more informative.
- A summary line may not be well estimated when the sensitivities and specificities of the various studies show little variability or when their estimated correlation across studies is small. Further, if there is evidence that the estimated correlation of sensitivity and specificity across studies is positive (rather than negative, which would be more typical), a "threshold effect" is not a plausible explanation for the observed pattern across studies. Rather, it is likely that an important covariate has not been taken into account.
- In many applications, a reasonable case can be made for summarizing studies with both a summary point and a summary line, as these provide alternative perspectives.

Step 1: Start by considering sensitivity and specificity independently.

This step is probably self-explanatory; it encourages reviewers to familiarize themselves with the pattern of study-level sensitivities and specificities. It is very instructive to create side-by-side forest plots of sensitivity and specificity in which studies are ordered by either sensitivity or specificity. The point of this graphical assessment is to obtain a visual impression of the variability of sensitivity and specificity across studies, as well as an impression of any relationship between sensitivity and specificity across studies, particularly if such a relationship is prominent (Figure 8–1 and illustrative examples).

If a summary point is deemed a helpful summary of the data, it is reasonable first to perform separate meta-analyses of sensitivity and specificity. The differences in the point estimates of summary sensitivity and specificity with univariate (separate) versus bivariate (joint) meta-analyses is often small. In an empirical exploration of 308 meta-analyses, differences in the estimates of summary sensitivity and specificity were rarely larger than 5 percent.[40] The width of the confidence intervals for the summary sensitivity and specificity is also similar between univariate and bivariate analyses. This suggests that, practically, univariate and multivariate analyses may yield comparable results. However, our recommendation is to prefer reporting the results from the hierarchical (multivariate) meta-analysis methods because of their better theoretical motivation and because of their natural symmetry with the multivariate methods that yield summary lines.

Step 2: Perform multivariate meta-analysis (when each study reports a single threshold).

To obtain a summary point, meta-analysts should perform bivariate meta-analyses (preferably using the exact binomial likelihood).

Meta-analysts should obtain summary lines based on multivariate meta-analysis models. The interpretation of the summary line should not automatically be that there are "threshold effects." This is most obvious when performing meta-analyses with evidence of a positive correlation between sensitivity and specificity, which cannot be attributed to a "threshold effect," as mentioned above.

If more than one threshold is reported per study and there is no strong *a priori* rationale to review only results for a specific threshold, meta-analysts should consider incorporating alternative thresholds into the appropriate analyses discussed previously. Tentatively, we

encourage both qualitative analysis via graphs and quantitative analyses via one of the multivariate methods mentioned above.

Step 3. Explore between-study heterogeneity.

Other than accounting for the presence of a "threshold effect," the HSROC and bivariate models provide flexible ways to test and explore between-study heterogeneity. The HSROC model allows one to examine whether any covariates (study characteristics) explain the observed heterogeneity in the accuracy and threshold parameters. One can use the same set of covariates for both parameters, but this is not mandatory, and should be judged for the application at hand. On the other hand, bivariate models allow one to use covariates to explain heterogeneity in sensitivity or specificity or both; and again, covariates for each measure can be different. Covariates that reduce the unexplained variability across studies (heterogeneity) may represent important characteristics that should be taken into account when summarizing the studies, or they may represent spurious associations.We refer to other texts for a discussion of the premises and pitfalls of metaregressions.[25,44] Factors reflecting differences in patient populations and methods of patient selection, methods of verification and interpretation of results, clinical setting, and disease severity are common sources of heterogeneity. Investigators are encouraged to use multivariate models to explore heterogeneity, especially when they have chosen these methods for combining studies.

Illustrations

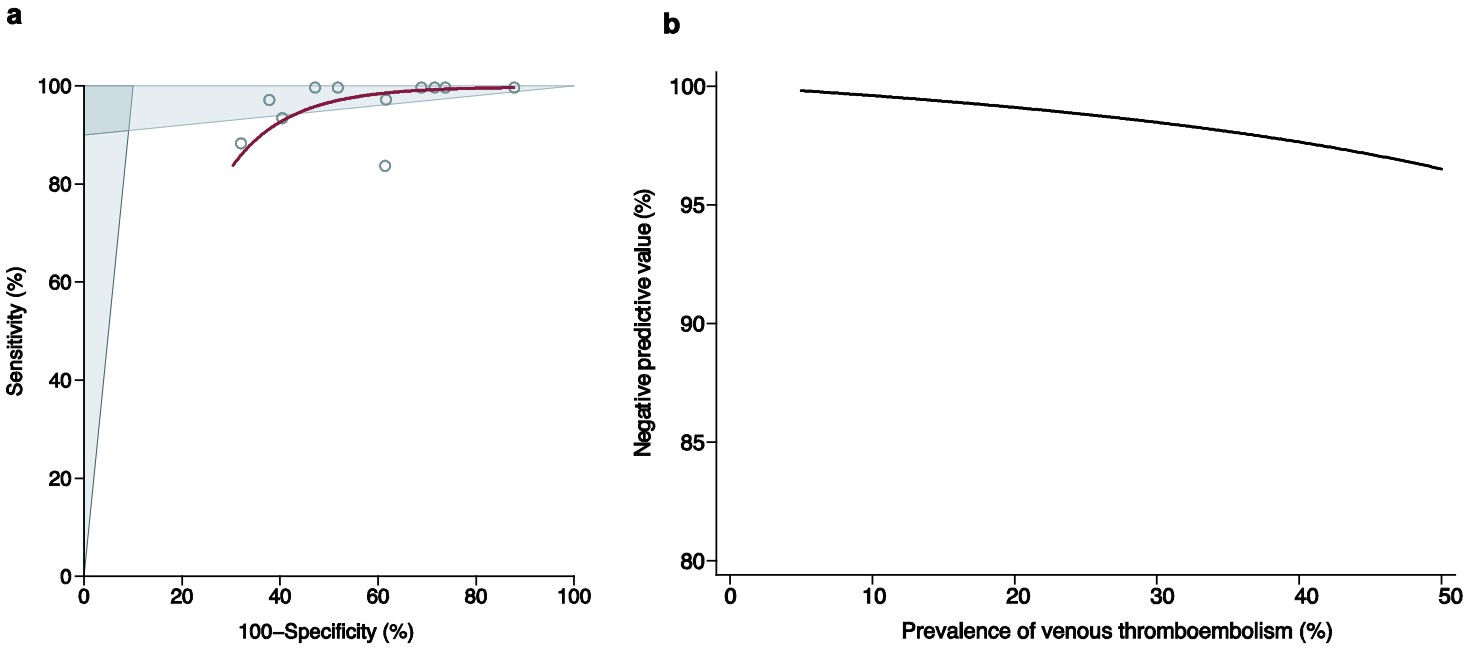
We briefly demonstrate the above with two applied examples. The first example on D-dimer assays for the diagnosis of venous thromboembolism[16] shows heterogeneity which could be attributed to a "threshold effect" as discussed by Lijmer et al.[45] The second example is from an evidence report on the use of serial creatine kinase-MB measurements for the diagnosis of acute cardiac ischemia,[17;18] and shows heterogeneity for another reason.

First Example: D-dimers for Diagnosis of Venous Thromboembolism

D-dimers are fragments specific for fibrin degradation in plasma, and can be used to diagnose venous thromboembolism. Figure 8–1 presents forest plots of the sensitivity and specificity and the likelihood ratios for the D-dimer example.[45] Sensitivity and specificity appear more heterogeneous than the likelihood ratios. (This is verified by formal testing for heterogeneity). This may be due to threshold variation in these studies (from 120 to 550 ng/mL, when stated; Figure 8–1), or due to other reasons.[45]

Because of the explicit variation in the thresholds for studies of D-dimers, it is probably more helpful to summarize the performance of the test using a HSROC, rather than to provide summary sensitivities and specificities (Figure 8–4a). (For simplicity, we select the highest threshold from two studies that report multiple ELISA [enzyme-linked immunosorbent assay] thresholds.) This test has very good diagnostic ability, and it appropriately focuses on minimizing false negative diagnoses. It is also informative to estimate "summary" negative (or positive) predictive values for this test. As described previously, we can calulate them based on the summary sensitivity and specificity estimates and over a range of plausible values for the prevalence. Figure 8–4b shows such an example using the summary sensitivity and specificity of the 11 studies of Figure 8–4a.

Figure 8–4. HSROC for the ELISA-based D-dimer tests



(**a**) Hierarchical summary receiver-operator curve (HSROC) of the studies plotted in Figure 8–1a.
(**b**) Calculated negative predictive value for the ELISA-based D-dimer test if the sensitivity and specificity are fixed at 80% and 97%, respectively, and prevalence of venous thromboembolism varies from 5 to 50 percent.

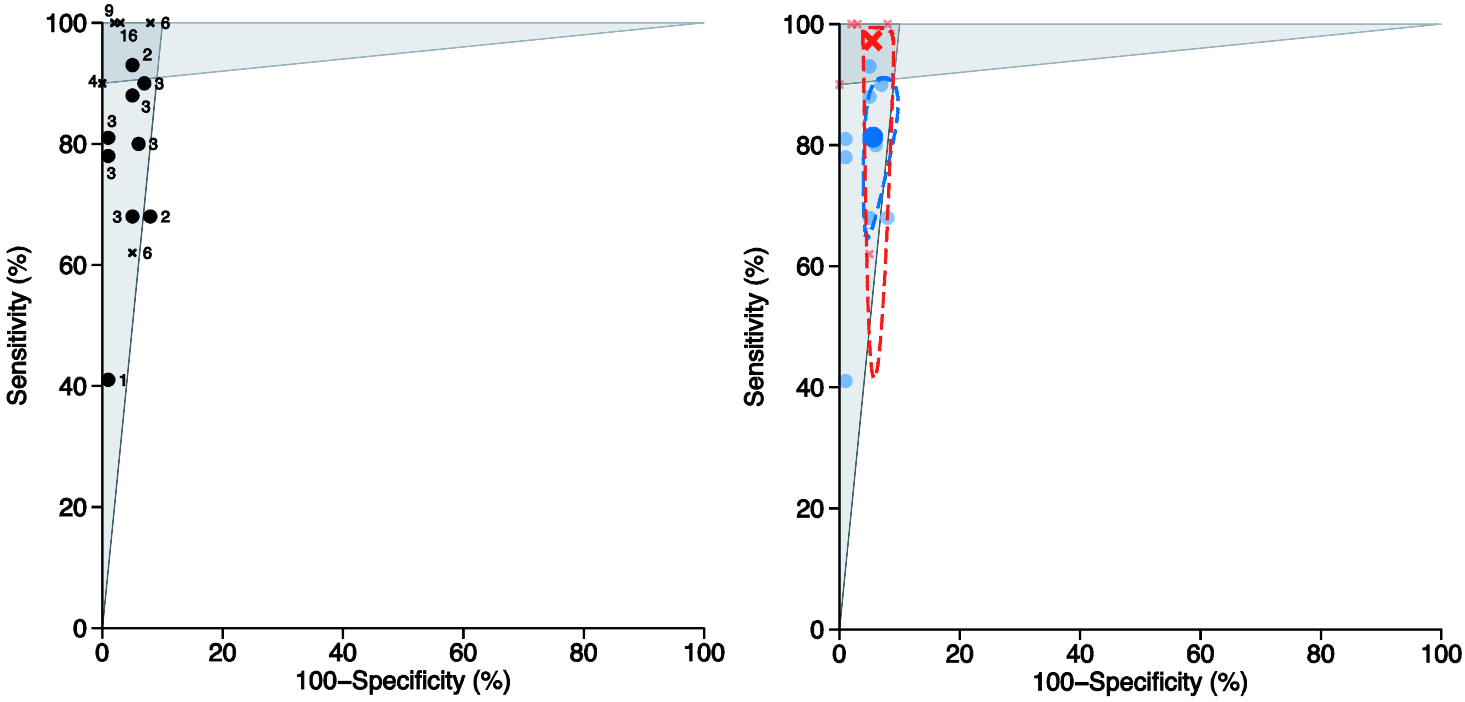
Second Example: Serial Creatine Kinase-MB Measurements for Diagnosing Acute Cardiac Ischemia

An evidence report examined the ability of serial creatine kinase-MB (CK-MB) measurements to diagnose acute cardiac ischemia in the emergency department.[17,18] Figure 8–5 shows the 14 eligible studies along with how many hours after symptom onset the last measurement was taken. It is evident that there is between-study heterogeneity in the sensitivities, and that sensitivity increases with longer time from symptom onset.

For illustrative purposes, we compare the summary sensitivity and specificity of studies where the last measurement was performed within 3 hours of symptom onset versus greater than 3 hours from symptom onset (Table 8–2 and Figure 8–5). We use a bivariate multilevel model with exact binomial likelihood. In the fixed effects part of the model, we include a variable that codes whether the last measurement was earlier than 3 hours from symptom onset or not. We allow this variable to have different effects on the summary sensitivity and on the summary specificity. This is essentially a bivariate meta-regression.

Note that properly specified bivariate meta-regressions (or HSROC-based meta-regressions) can be used to compare two or more medical tests. The specification of the meta-regression models will be different when the comparison is indirect (when different medical tests are examined in independent studies) or direct (when the different medical tests are applied in the same patients in each study).

Figure 8–5. Sensitivity 1–specificity plot for studies of serial CK-MB measurements



The left panel shows the sensitivity and specificity of 14 studies according to the timing of the last serial CK-MB measurement for diagnosis of acute cardiac ischemia. The numbers next to each study point are the actual length of the time interval from symptom onset to last serial CK-MB measurement. Filled circles: at most 3 hours; "x" marks: longer than 3 hours.
The right panel plots the summary points and the 95% confidence regions for the aforementioned subgroups of studies (at most 3 hours: blue filled circles; longer than 3 hours – red "x"s). Estimates are based on a bivariate meta-regression using the time interval as a predictor. The predictor has distinct effects for sensitivity and specificity. This is the same analysis as in Table 8–2.

Table 8–2. Meta-regression-based comparison of diagnostic performance*

Meta-Analysis Metric	≤3 Hours	>3 Hours	p-Value for the Comparison Across Subgroups
Summary sensitivity (percent)	80 (64 to 90)	96 (85 to 99)	0.036
Summary specificity (percent)	97 (94 to 98)	97 (95 to 99)	0.56

*Results based on a bivariate meta-regression that effectively compared the summary sensitivity and summary specificity according to the timing of the last serial CK-MB measurement for diagnosis of acute cardiac ischemia. The meta-regression is on a variable that takes the value 1 if the time from the onset of symptoms to testing was 3 hours or less, and the value 0, when the respective time interval was more than 3 hours. The bivariate meta-regression model allows for different effects of timing on sensitivity and specificity. To facilitate interpretation, we present the summary sensitivity and specificity in each subgroup, calculated from the parameters of the meta-regression model, which also gave the p-values for the effect of timing on test performance.

Overall Recommendations

We summarize:

- Consider presenting a "summary point" when sensitivity and specificity do not vary widely across studies and studies use the same explicit or "implicit threshold."
    - To obtain a summary sensitivity and specificity use the theoretically motivated bivariate meta-analysis models.

    - Back-calculate overall positive and negative predictive values from summary estimates of sensitivity and specificity, and for a plausible range of prevalence values rather than meta-analyzing them directly.
    - Back-calculate overall positive and negative likelihood ratios from summary estimates of sensitivity and specificity, rather than meta-analyzing them directly.
- If the sensitivity and specificity vary over a large range, it may be more helpful to use a summary line, which best describes the relationship of the average sensitivity and specificity. The summary line approach is also most helpful when different explicit thresholds are used across studies. To obtain a summary line use multivariate meta-analysis methods such as the HSROC model.
    - Several SROC lines can be obtained based on multivariate meta-analysis models, and they can have different shapes.
    - If there is evidence of a positive correlation, the variability in the studies cannot be secondary to a "threshold effect"; explore for missing important covariates. Arguably, the summary line is a valid description of how average sensitivity relates to average specificity.
- If more than one threshold is reported per study, this has to be taken into account in the quantitative analyses. We encourage both qualitative analysis via graphs and quantitative analyses via proper methods.
- One should explore the impact of study characteristics on summary results in the context of the primary methodology used to summarize studies using meta-regression-based analyses or subgroup analyses.

References

1. Methods Guide for Medical Test Reviews. AHRQ Publication No. 12-EC017. Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published as a special supplement to the Journal of General Internal Medicine, July 2012.
2. Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. Lancet 1998; 351(9096):123-127.
3. Tatsioni A, Zarin DA, Aronson N, Samson DJ, Flamm CR, Schmid C et al. Challenges in systematic reviews of diagnostic technologies. Ann Intern Med 2005; 142(12 Pt 2):1048-1055.
4. Lijmer JG, Leeflang M, Bossuyt PM. Proposals for a phased evaluation of medical tests. Med Decis Making 2009; 29(5):E13-E21.
5. Lord SJ, Irwig L, Bossuyt PM. Using the principles of randomized controlled trial design to guide test evaluation. Med Decis Making 2009; 29(5):E1-E12.
6. Trikalinos TA, Kulasingam S, Lawrence WF. Deciding whether to complement a systematic review of medical tests with decision modeling. AHRQ Publication No. EHC082-EF. Chapter 10 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.

7. Matchar DB. Introduction to the methods guide for medical test reviews. AHRQ Publication No. EHC073-EF. Chapter 1 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.

8. Hartmann KE, Matchar DB and Chang S. Assessing applicability of medical test studies in systematic reviews. AHRQ Publication No. 12-EHC078-EF. Chapter 6 of of Methods Guide for Medical Test Reviews. AHRQ Publication No. 12-EHC017. Rockville, MD: Agency for Healthcare Research and Quality; June 2012. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.

9. Trikalinos TA, Ballion CM. Options for summarizing medical test performance in the absence of a "gold standard." AHRQ Publication No. 12-EHC081-EF. Chapter 9 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Also published in a special supplement to the Journal of General Internal Medicine, July 2012.

10. Loong TW. Understanding sensitivity and specificity with the right side of the brain. BMJ 2003; 327(7417):716-719.

11. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003; 138(1):W1-12.

12. Santaguida PL, Riley CM and Matchar DB. Assessing risk of bias as a domain of quality in medical test studies. AHRQ Publication No. 12-EHC077-EF. Chapter 5 of Methods Guide for Medical Test Reviews. AHRQ Publication No. 12-EHC017. Also published in a specual supplement to the Journal of General Internal Medicine, July 2012.

13. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999; 282(11):1061-1066.

14. Rutjes AW, Reitsma JB, Di NM, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006; 174(4):469-476.

15. 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.

16. Becker DM, Philbrick JT, Bachhuber TL, Humphries JE. D-dimer testing and acute venous thromboembolism. A shortcut to accurate diagnosis? Arch Intern Med 1996; 156(9):939-946.

17. Balk EM, Ioannidis JP, Salem D, Chew PW, Lau J. Accuracy of biomarkers to diagnose acute cardiac ischemia in the emergency department: a meta-analysis. Ann Emerg Med 2001; 37(5):478-494.

18. Lau J, Ioannidis JP, Balk E, Milch C, Chew P, Terrin N et al. Evaluation of technologies for identifying acute cardiac ischemia in emergency departments. Evid Rep Technol Assess (Summ) 2000;(26):1-4.

19. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics 2007; 8(2):239-251.

20. Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC et al. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 1994; 120(8):667-676.

21. Leeflang MMG, Bossuyt PM, Irwig L. Diagnostic accuracy may vary with prevalence: Implications for evidence-based diagnosis. J Clin Epidemiol 2008; (accepted).

22. Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. BMJ 2004; 329(7458):168-169.

23. Zwinderman AH, Bossuyt PM. We should not pool diagnostic likelihood ratios in systematic reviews. Stat Med 2008; 27(5):687-697.

24. Simel DL, Bossuyt PM. Differences between univariate and bivariate models for summarizing diagnostic accuracy may not be large. J Clin Epidemiol 2009.

25. Fu R, Gartlehner G, Grant M, Shamliyan T, Sedrakyan A, Wilt TJ et al. Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program. J Clin Epidemiol 2011; 64(11):1187-1197.

26. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol 2003; 56(11):1129-1135.

27. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005; 58(10):982-990.

28. Riley RD, Abrams KR, Sutton AJ, Lambert PC, Thompson JR. Bivariate random-effects meta-analysis and the estimation of between-study correlation. BMC Med Res Methodol 2007; 7:3.

29. Riley RD, Abrams KR, Lambert PC, Sutton AJ, Thompson JR. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. Stat Med 2007; 26(1):78-97.

30. Harbord RM, Whiting P, Sterne JA, Egger M, Deeks JJ, Shang A et al. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. J Clin Epidemiol 2008; 61(11):1095-1103.

31. Rutter CM, Gatsonis CA. Regression methods for meta-analysis of diagnostic test data. Acad Radiol 1995; 2 Suppl 1:S48-S56.

32. Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Ann Intern Med 2002; 137(7):598-602.

33. Kardaun JW, Kardaun OJ. Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation. Methods Inf Med 1990; 29(1):12-22.

34. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Making 1993; 13(4):313-321.

35. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med 1993; 12(14):1293-1316.

36. Oei EH, Nikken JJ, Verstijnen AC, Ginai AZ, Myriam Hunink MG. MR imaging of the menisci and cruciate ligaments: a systematic review. Radiology 2003; 226(3):837-848.

37. Visser K, Hunink MG. Peripheral arterial disease: gadolinium-enhanced MR angiography versus color-guided duplex US--a meta-analysis. Radiology 2000; 216(1):67-77.

38. Arends LR, Hamza TH, van Houwelingen JC, Heijenbrok-Kal MH, Hunink MG, Stijnen T. Bivariate random effects meta-analysis of ROC curves. Med Decis Making 2008; 28(5):621-638.

39. Chappell FM, Raab GM, Wardlaw JM. When are summary ROC curves appropriate for diagnostic meta-analyses? Stat Med 2009; 28(21):2653-2668.

40. Dahabreh IJ, Chung M, Kitsios G, Terasawa T, Raman G, Tatsioni, A, Tobar A, LauJ, Trikalinos TA, Schmid CH. Evaluating Practices and Developing Tools for Comparative Effectiveness Reviews of Diagnostic Test Accuracy. Task 1: Comprehensive Overview of Methods and Reporting of Meta- Analyses of Test Accuracy. AHRQ Methods Research Report. AHRQ Publication No. 12-EHC044-EF. Rockville, MD: Agency for Healthcare Research and Quality; March 2012. http://www.effectivehealthcare.ahrq.gov/ehc/products/288/1018/ComprehensiveOverview_MethodsResearchReport_20120327.pdf

41. Dukic V, Gatsonis C. Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. Biometrics 2003; 59(4):936-946.

42. Kester AD, Buntinx F. Meta-analysis of ROC curves. Med Decis Making 2000; 20(4):430-439.

43. Trikalinos TA, Chung M, Lau J, Ip S. Systematic review of screening for bilirubin encephalopathy in neonates. Pediatrics 2009; 124(4):1162-1171.

44. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. Stat Med 1999; 18(20):2693-2708.

45. Lijmer JG, Bossuyt PM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. Stat Med 2002; 21(11):1525-1537.

46. Hamza TH, van Houwelingen HC, Stijnen T. The binomial distribution of meta-analysis was preferred to model within-study variability. J Clin Epidemiol 2008; 61(1):41-51.

47. Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. J Clin Epidemiol 2006; 59(12):1331-1332.

48. Hamza TH, Reitsma JB, Stijnen T. Meta-analysis of diagnostic studies: a comparison of random intercept, normal-normal, and binomial-normal bivariate summary ROC approaches. Med Decis Making 2008; 28(5):639-649.

49. Cochrane Diagnostic Test Accuracy Working Group. Handbook for diagnostic test accuracy reviews. 2011. The Cochrane Collaboration.

Funding: Funded by the Agency for Health Care Research and Quality (AHRQ) under the Effective Health Care Program.

Conflict of interest: None of the authors has affiliations or financial involvements that conflict with the information presented in this chapter.

Corresponding author: TA Trikalinos, Tufts Medical Center, 800 Washington St, Box#63, Boston, MA 02111, US. | Telephone: +1 617 636 0734 | Fax: +1 617 636 8628. Email: Thomas.Trikalinos@tufts.edu.

Suggested citation: Trikalinos TA, Balion CM, Coleman CI, et al. Meta-analysis of test performance when there is a "gold standard." AHRQ Publication No. 12-EHC080-EF. Chapter 8 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm . Also published in a special supplement to the Journal of General Internal Medicine, July 2012.

Chapter 9

Options for Summarizing Medical Test Performance in the Absence of a "Gold Standard"

Thomas A. Trikalinos, M.D., Tufts Evidence-based Practice Center, Boston, MA
Cynthia M. Balion, Ph.D., Hamilton General Hospital and McMaster Evidence-based Practice Center, Hamilton, Ontario, CA

Abstract

The classical paradigm for evaluating test performance compares the results of an index test with a reference test. When the reference test does not mirror the "truth" adequately well (e.g. "imperfect" reference standard), the typical ("naïve") estimates of sensitivity and specificity are biased. One has at least four options when performing a systematic review of test performance when the reference standard is imperfect: (a) to forgo the classical paradigm and assess the index test's ability to predict patient-relevant outcomes instead of test accuracy (i.e., treat the index test as a predictive instrument); (b) to assess whether the results of the two tests (index and reference) agree or disagree (i.e., treat them as two alternative measurement methods); (c) to calculate naïve estimates of the index test's sensitivity and specificity from each study included in the review, and discuss the direction in which they are biased; (d) mathematically adjust the naïve estimates of sensitivity and specificity of the index test to account for the imperfect reference standard. We discuss these options and illustrate some of them through examples.

Introduction

In the classical paradigm for evaluating the "accuracy" or performance of a medical test (index test), the results of the test are compared with the "true" status of every tested individual or every tested specimen. Sometimes, this "true" status is directly observable (e.g., for tests predicting short-term mortality after a procedure). However, in many cases the "true" status of the tested subject is judged based upon another test as a reference method. Problems can arise when the reference test does not mirror the "truth" adequately well: one will be measuring the performance of the index test against a faulty standard, and is bound to err. The worse the deviation of the reference test from the unobserved "truth," the poorer the estimate of the index test's performance will be. This phenomenon is known as "reference standard bias."[1–4]

In this chapter we discuss how researchers engaged in the Effective Health Care Program of the United States Agency for Healthcare Research and Quality (AHRQ) think about synthesizing data on the performance of medical tests when the reference standard is imperfect. Because this challenge is a general one and not specific to AHRQ's program, we anticipate that the current discussion is of interest to the wider group of those who perform or use systematic reviews of

medical tests. Of the many challenges that pertain to issues with reference standards, we will discuss only one, namely, the case of a reference standard test that itself misclassifies the test subjects at a rate we are not willing to ignore (imperfect reference standard). We will not discuss verification bias, where the use of the reference standard is guided by the results of the index test and is not universal.

Imperfect Reference Standards

What is meant by "imperfect reference standard," and why is it important for meta-analysis and synthesis in general?

Perhaps the simplest case of test performance evaluation includes an "index test" and a reference test ("reference standard") whose results are dichotomous in nature (or are made dichotomous). Both tests are used to provide information on the presence or absense of the condition of interest, or predict the occurrence of a future event. For the vast majority of medical tests, both the results of the index test and the reference test can be different from the true status of the condition of interest. Figure 9–1 shows the correspondence between the true 2 X 2 table probabilities (proportions) and the eight strata defined by the combinations of index and reference test results and the presense or absense of the condition of interest. These eight probabilities (α_1, β_1, γ_1, δ_1, α_2, β_2, γ_2 and δ_2) are not known, and have to be estimated from the data (from studies of diagnostic or prognostic accuracy). More accurately, a study of diagnostic accuracy tries to estimate quantities that are functions of the eight probabilities.

Diagnostic Accuracy, or the Case of the "Perfect" Reference Standard

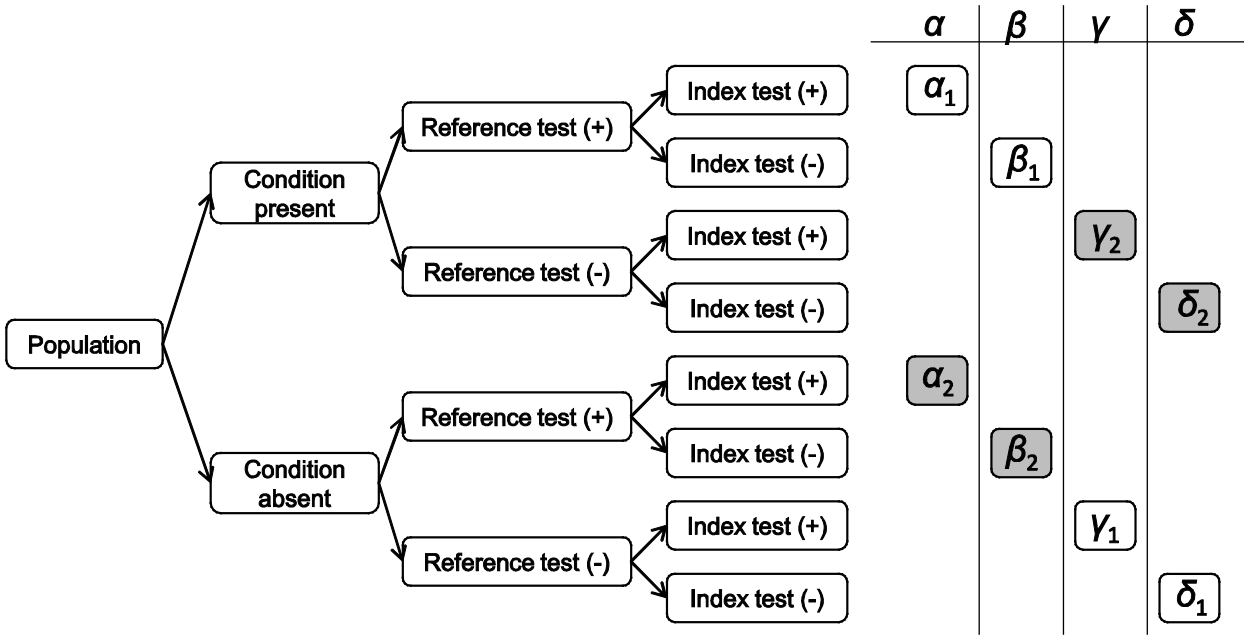
A "perfect" reference standard would be infallible. It would always match the condition of interest, and, thus, in Figure 9–1 the proportions in the grey boxes (α_2, β_2, γ_2 and δ_2) would be zero. The data in the 2 X 2 table are then sufficient to estimate the four remaining probabilities (α_1, β_1, γ_1, and δ_1). Because the four probabilities necessarily sum to 1, it is sufficient to estimate any three. In practice, one estimates three other parameters, which are functions of the probabilities in the cells, namely, the sensitivity and specificity of the index test and the prevalence of the condition of interest (Table 9–1). If the counts in the 2 X 2 table are available (e.g., from a cohort study assessing the index test's performance), one can estimate the sensitivity and the specificity of the index test in a straightforward manner:

$$Se_{index} = \frac{\text{true positives}}{\text{true positives+false negatives}}, \text{ and } Sp_{index} = \frac{\text{true positives}}{\text{true positives+false negatives}}$$

respectively.

Figure 9–1. Correspondence of test results and true proportions in the 2 X 2 table

	Reference standard (+)	Reference standard (-)
Index test (+)	α	γ
Index test (-)	β	δ



The cells in the 2 X 2 table, α, β, γ, *δ*, are the true population proportions corresponding to combinations of test results. The diagram depicts how these proportions break down according to the (unknown) true status of the condition of interest. For example, the proportion when both the index test and the reference standard are positive is $\alpha = \alpha_1 + \alpha_2$ (i.e., the sum of the proportion of positive index and reference test results when the condition is present (α_1) and absent (α_2)), and similarly for the other groups. A white colored box and the subscript 1 is used when the reference standard result matches the true status of the condition of interest; a grey colored box and the subscript 2 is used when it does not.

Table 9–1. Parameterization when the reference standard is assumed "perfect" ("gold standard")

	Reference Standard (+)	Reference Standard (-)
Index Test (+)	$\underbrace{p \times Se_{index}}_{\alpha_1} + \underbrace{0}_{\alpha_2}$	$\underbrace{(1-p) \times (1-Sp_{index})}_{\gamma_1} + \underbrace{0}_{\gamma_2}$
Index Test (-)	$\underbrace{p \times (1-Se_{index})}_{\beta_1} + \underbrace{0}_{\beta_2}$	$\underbrace{(1-p) \times Sp_{index}}_{\delta_1} + \underbrace{0}_{\delta_2}$

Only three unknowns exist: the sensitivity and specificity of the index test (Se_{index} and Sp_{index}, respectively) and the disease prevalence (p). The under-braces refer to the probabilities of the 8 strata in Figure 9–1. These can be estimated from test results in a study, as discussed in the Appendix.

Diagnostic Accuracy – The Case of the "Imperfect" Reference Standard

Only rarely are we sure that the reference standard is a perfect reflection of the truth. Most often in our assessments we accept some degree of misclassification by the reference standard, implicitly accepting it as being "as good as it gets." Table 9–2 lists some situations where we

might question the validity of the reference standard. Unfortunately, there are no hard and fast rules for judging the adequacy of the reference standard; systematic reviewers should consult content experts in making such judgments.

Table 9–2. Situations where one can question the validity of the reference standard

Situation	Example
The reference method yields different measurements over time or across settings.	Briefly consider the diagnosis of obstructive sleep apnea, which typically requires a high Apnea-Hypopnea Index (AHI, an objective measurement), and the presence of suggestive symptoms and signs. However, there is large night-to-night variability in the measured AHI, and there is also substantial variability between raters and between labs.
The condition of interest is variably defined.	This can be applicable to diseases that are defined in complex ways or qualitatively (e.g., based both on symptom intensity and on objective measurements). Such an example could be a complex disease such as psoriatic arthritis. There is no single symptom, sign, or measurement that suffices to make the diagnosis of the disease with certainty. Instead a set of criteria including symptoms, signs, imaging, and laboratory measurements are used to identify it. Unavoidably, diagnostic criteria will be differentially applied across studies, and this is a potential explanation for the varying prevalence of the disease across geographic locations[34] and over time.
The new method is an improved version of a usually applied test.	Older methodologies for the measurement of parathyroid hormone (PTH) are being replaced by newer, more specific ones. PTH measurements with different methodologies do not agree very well.[35] Here, it would be wrong to assume that the older version of the test is the reference standard for distinguishing patients with high PTH from those without.

Table 9–3 shows the relationship between the sensitivity and specificity of the index and reference tests and the prevalence of the condition of interest when the results of the index and reference tests are independent among those with and without the condition of interest (i.e, where there is"conditional independence," one of several possibilities).

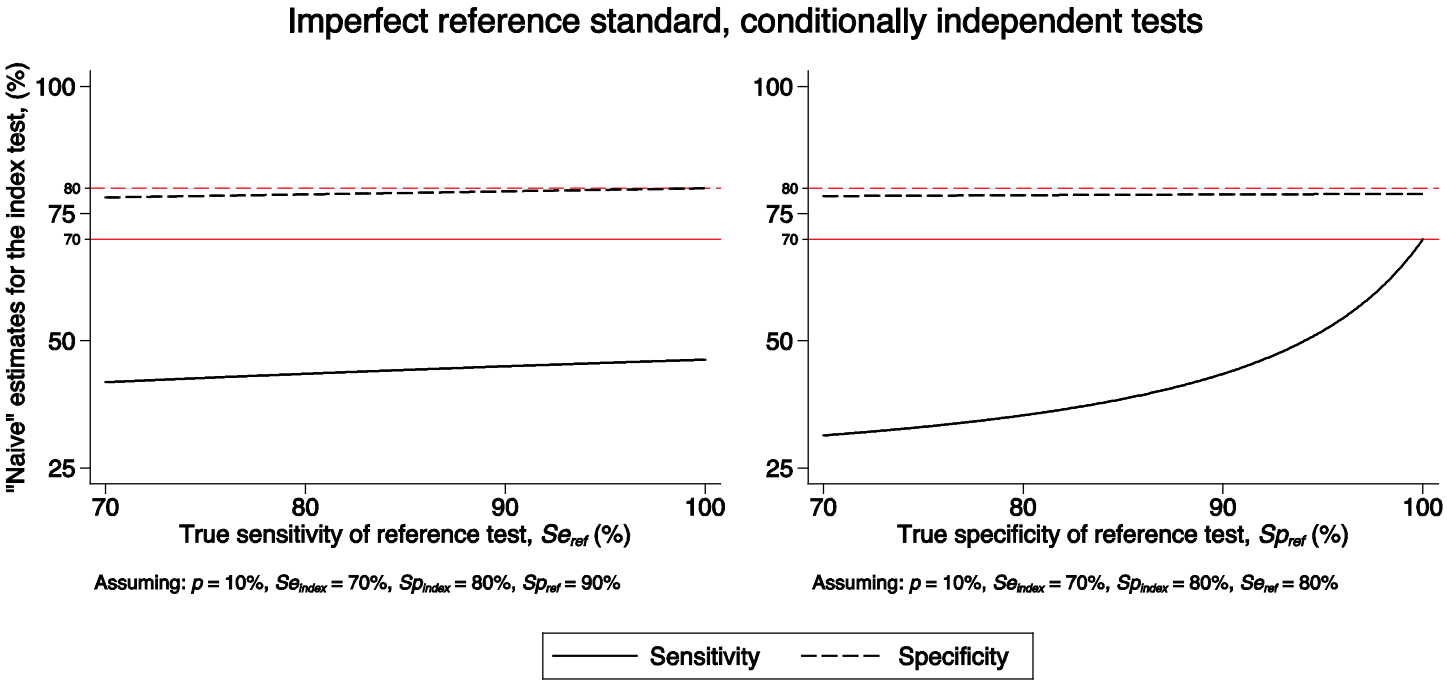
Table 9–3. Parameterization when the reference test is assumed to be imperfect, and the index and reference test results are assumed independent within the strata of the condition of interest

	Reference Test (+)	Reference Test (-)
Index Test (+)	$\underbrace{p \times Se_{ref} \times Se_{index}}_{\alpha_1} + \underbrace{(1-p) \times (1-Sp_{ref}) \times (1-Sp_{index})}_{\alpha_2}$	$\underbrace{(1-p) \times Sp_{ref} \times (1-Sp_{index})}_{\gamma_1} + \underbrace{p \times (1-Se_{ref}) \times Se_{index}}_{\gamma_2}$
Index Test (-)	$\underbrace{p \times Se_{ref} \times (1-Se_{index})}_{\beta_1} + \underbrace{(1-p) \times (1-Sp_{ref}) \times Sp_{index}}_{\beta_2}$	$\underbrace{(1-p) \times Sp_{ref} \times Sp_{index}}_{\delta_1} + \underbrace{p \times (1-Se_{ref}) \times (1-Se_{index})}_{\delta_2}$

We now have five unknowns: the sensitivity and specificity of the index test (Se_{index} and Sp_{index}, respectively) and of the reference test (Se_{ref} and Sp_{ref}, respectively), and the disease prevalence (p). The under-braces refer to the probabilities of the 8 strata in Figure 9–1. Note that sensitivity and specificity *always refer to the (unknown) true status of the condition of interest.* Further, the results of the tests are assumed to be independent given the true status of the condition of interest. The cross-tabulation of the results of the index and reference tests is not sufficient to specify the problem, and additional information is necessary. It is easy to see that if the reference test is "perfect" ($Se_{ref} = 1$, $Sp_{ref} = 1$), one obtains the parameterization in Table 9–1. If the results of the index and reference tests are not independent among units with or without the condition of interest, the formulas in Table 9–1 change; in fact, several parameterizations are possible.[18,25–31]

For conditionally independent tests, estimates of sensitivity and specificity from the standard formulas (naïve estimates) are always smaller than the true values (see an example in Figure 9–2, and later for a more detailed discussion).

Figure 9–2. Naïve estimates versus true values for the performance of the index test with an imperfect reference standard



Se_{index} and Sp_{index} = sensitivity and specificity of the index test, respectively; Se_{ref} and Sp_{ref}: sensitivity and specificity of the reference test, respectively; p: disease prevalence.
If the results of the index and reference tests are independent conditional on disease status, the naïve estimates for the performance of the index test are underestimates. The red (lighter) reference lines are the true sensitivity (solid) and specificity (dashed) of the index test. Note that the naïve estimate for the sensitivity and specificity of the index test approach the true values as the sensitivity and specificity of the reference test approaches 100%. In the left plot the naïve estimate of sensitivity does not reach 70% (the true value) when the sensitivity of the reference test, Seref, is 100%, because the specificity of the reference test is not perfect (Sp_{ref}=90%). Similarly, on the plot on the right, the specificity of the index test does not reach the true value of 80% when the specificity of the reference test, Sp_{ref}, is 100%, because the sensitivity of the reference test is not perfect (Se_{ref}=80%). The naïve estimates would be the same as the true values only if both the sensitivity and the specificity of the reference test are 100%.

Options for Systematic Reviewers

So how should one approach the challenge of synthesizing information on diagnostic or prognostic tests when the purported "reference standard" is judged to be inadequate? At least four options exist. The first two change the framing of the problem, and forgo the classical paradigm for evaluating test performance. The third and fourth work within the classical paradigm and rely on qualifying the interpretation of results, or on mathematical adjustments:

1. Forgo the classical paradigm; assess the index test's ability to predict patient-relevant outcomes instead of test accuracy (i.e., treat the index test as a predictive instrument).[5,6] This reframing applies when outcome information (usually on long-term outcomes) exists, and the measured patient outcomes are themselves valid. If so, the approach to such a review is detailed in Chapter 11 in this *Medical Test Methods Guide*.[7]

2. Forgo the classical paradigm; assess simply whether the results of the two tests (index and reference) agree or disagree (i.e., treat them as two alternative measurement methods). Instead of calculating sensitivity and specificity one would calculate statistics on test concordance, as mentioned later.
3. Work within the classical paradigm, and calculate "naïve estimates" of the index test's sensitivity and specificity from each study, but qualify the study findings.
4. Adjust the naïve estimates of sensitivity and specificity of the index test to account for the imperfect reference standard.

Our subjective assessment is that, when possible, the first option is preferable as it recasts the problem into one that is inherently clinically meaningful. The second option may be less clinically meaningful, but is a defensible alternative to treating an inadequate reference standard as if it were effectively perfect. The third option is potentially subject to substantial bias, which is especially difficult to interpret when the results of the test under review and the "reference standard" are not conditionally independent (i.e., when an error in one is more or less likely when there is an error in the other). The fourth option would be ideal if the adjustment methods were successful (i.e., if the adjustments eliminated biased estimates of sensitivity and specificity in the face of an imperfect reference standard). However, the techniques available necessarily require information that is typically not included in the reviewed studies, and require advanced statistical modeling.

Option 1. Assess the index test's ability to predict patient-relevant outcomes instead of test accuracy.

This option is not universally possible. Instead of assessing the diagnostic or screening performance of the test, it quantifies the impact of patient management strategies that include testing on (usually long-term) clinical outcomes. When it is possible and desirable to recast the evaluation question as an assessment of a tests ability to predict health outcomes, there are specific methods to consider when performing the assessment. For a more detailed discussion, the reader is referred to Chapter 11 of this *Medical Test Methods Guide.*[7]

Option 2. Assess the concordance of difference tests instead of test accuracy.

Here, the index and reference tests are treated as two alternative measurement methods. One explores how well one test agrees with the other test(s), and perhaps asks if one test can be used in the place of the other. Assessing concordance may be the only meaningful option if none of the compared tests is an obvious choice for a reference standard (e.g., when both tests are alternative methodologies to measure the same quantity).

In the case of categorical test results, one can summarize the extent of agreement between two tests using Cohen's κ statistic (a measure of categorical agreement which takes into account the probability that some agreement will occur by chance). A meta-analysis of κ statistics may also be considered to supplement a systematic review.[8] Because such a meta-analysis is not common practice in the medical literature, it should be explained and interpreted in some detail.

In the case of continuous test results, one is practically limited by the data available. If individual data points are available or extractable (e.g., in appendix tables or by digitizing plots), one can directly compare measurements with one test versus measurements with the other test. One way to do so is to perform an appropriate regression to obtain an equation for translating the

measurements with one test to the measurements of the other. Because both measurements have random noise, an ordinary least squares regression is not appropriate; it treats the "predictor" as fixed and error-free and thus underestimates the slope of the relationship between the two tests. Instead, one should use a major axis or similar regression,[9–12] or more complex regressions that account for measurement error; consulting a statistician is probably wise. An alternative and well-known approach is to perform difference versus average analyses (Bland-Altman-type of analyses[13–15]). A qualitative synthesis of information from Bland-Altman plots can be quite informative (see example).[16] As of this writing, the authors have not encountered any methods for incorporating difference versus average information from multiple studies.

If individual data points are not available, one has to summarize study-level information of the agreement of individual measurements. Of importance, care is needed when selecting which information to abstract. Summarizing results from major axis regressions or Bland-Altman analyses is probably informative. However, other metrics are not necessarily as informative. For example, Pearson's correlation coefficient, while often used to "compare" measurements with two alternative methods, is not a particularly good metric for two reasons: First, it does not provide information on the slope of the line describing the relationship between the two measurements; it informs on the degree of linearity of the relationship. Further, its value can be high (e.g., >0.90) even when the differences between the two measurements are clinically important. Thus, one should be circumspect in using and interpreting a high Pearson's correlation coefficient for measurement comparisons.

Option 3. Qualify the interpretation of naïve estimates of the index test's performance.

This option is straightforward. One could obtain naïve estimates of index test performance and make qualitative judgments on the direction of the bias of these naïve estimates.

Tests With Independent Results Within the Strata of the Disease

We have seen already in Table 9–3 that, when the results of the index and reference test are independent among those with and without the disease (conditional independence), the naïve sensitivity and specificity of the index test is biased down. The more imperfect the reference standard, the greater the difference between the naïve estimates and true test performance for the index test (Figure 9–2).

Tests With Correlated Results Within the Strata of the Disease

When the two tests are correlated conditional on disease status, the naïve estimates of sensitivity and specificity can be overestimates or underestimates, and the formulas in Table 9–3 do not hold. They can be overestimates when the tests tend to agree more than expected by chance. They can be underestimates when the correlation is relatively small, or when the tests disagree more than expected by chance.

A clinically relevant example is the use of prostate-specific antigen (PSA) to detect prostate cancer. PSA levels have been used to detect the presence of prostate cancer, and over the years, a number of different PSA detection methods have been developed. However, PSA levels are not elevated in as many as 15 percent of individuals with prostate cancer, making PSA testing prone to misclassification error.[17] One explanation for these misclassifications (false-negative results) is that obesity can reduce serum PSA levels. The cause of misclassification (obesity) will likely affect all PSA detection methods—patients who do not have elevated PSA by a new detection

method are also likely to not have elevated PSA by the older test. This "conditional dependence" will likely result in an overestimation of the diagnostic accuracy of the newer (index) test. In contrast, if the newer PSA detection method was compared to a non-PSA based reference standard that would not be prone to error due to obesity, such as prostate biopsy, conditional dependence would not be expected and estimates of diagnostic accuracy of the newer PSA method would likely be underestimated if misclassification occurs.

Because of the above, researchers should not assume conditional independence of test results without justification, particularly when the tests are based upon a common mechanism (e.g., both tests are based upon a particular chemical reaction, so that something which interferes with the reaction for one of the tests will likely interfere with the other test as well).[18]

Option 4. Adjust or correct the naïve estimates of sensitivity and specificity.

Finally, one can mathematically adjust or correct the naïve estimates of sensitivity and specificity of the index test to account for the imperfect reference standard. The 2×2 cross-tabulation of test results is not sufficient to estimate the true sensitivities and specificities of the two tests, the prevalence of the conditions of interest, and correlations between sensitivities and specificities among those with and without the condition of interest. Therefore, additional information is needed. Several options have been explored in the literature. The following is by no means a comprehensive description; it is just an outline of several of the numerous approaches that have been proposed.

The problem is much easier if one can assume conditional independence for the results of the two tests, and further, that some of the parameters are known from prior knowledge. For example, one could assume that the sensitivity and specificity of the reference standard to detect true disease status is known from external sources, such as other studies,[19] or that the specificities for both tests are known (from prior studies) but the sensitivities are unknown.[20] In the same vein one can encode knowledge from external sources with *prior distributions instead of fixed values*, using Bayesian inference.[21–24] Using a whole distribution of values rather than a single fixed value is less restrictive, and probably less arbitrary. The resulting posterior distribution provides information on the specificities and sensitivities of both the index test and the reference standard, and of the prevalence of people with disease in each study.

When conditional independence cannot be assumed, the conditional correlations have to be estimated as well. Many alternative parameterizations for the problem have been proposed.[18,25–31] It is beyond the scope of this chapter to describe them. Again, it is advisable to seek expert statistical help when considering such quantitative analyses, as modeling assumptions can have unanticipated implications[32] and model misspecification can result in biased estimates.[33]

Illustration

As an illustration we use a systematic review on the diagnosis of obstructive sleep apnea (OSA) in the home setting.[16] Briefly, OSA is characterized by sleep disturbances secondary to upper airway obstruction. It is prevalent in two to four percent of middle-aged adults, and has been associated with daytime somnolence, cardiovascular morbidity, diabetes and other metabolic abnormalities, and increased likelihood of accidents and other adverse outcomes. Treatment (e.g., with continuous positive airway pressure) reduces symptoms, and, hopefully, long-term risk for cardiovascular and other events. There is no "perfect" reference standard for

OSA. The diagnosis of OSA is typically established based on suggestive signs (e.g. snoring, thick neck) and symptoms (e.g., somnolence), and in conjunction with an objective assessment of breathing patterns during sleep. The latter assessment is by means of facility-based polysomnography, a comprehensive neurophysiologic study of sleep in the lab setting. Most commonly, polysomnography quantifies one's apnea-hypopnea index (AHI) (i.e., how many episodes of apnea [no airflow] or hypopnea [reduced airflow] a person experiences during sleep). Large AHI is suggestive of OSA. At the same time, portable monitors can be used to measure AHI instead of facility-based polysomnography.

Identifying (Defining) the Reference Standard

One consideration is what reference standard is most common, or otherwise "acceptable," for the main analysis. In all studies included in the systematic review, patients were enrolled only if they had suggestive symptoms and signs (although it is likely that these were differentially ascertained across studies). Therefore, in these studies, the definition of sleep apnea is practically equivalent to whether people have a "high enough" AHI.

Most studies and some guidelines define AHI $\geq$15 events per hour of sleep as suggestive of the disease, and this is the cutoff selected for the main analyses. In addition, identified studies used a wide range of cutoffs in the reference method to define sleep apnea (including 5, 10, 15, 20, 30, and 40 events per hour of sleep). As a sensitivity analysis, the reviewers decided to summarize studies also according to the 10 and the 20 events per hour of sleep cutoffs; the other cutoffs were excluded because data were sparse. It is worth noting that, in this case, the exploration of the alternative cutoffs did not affect the results or conclusions of the systematic review, but did require substantial time and effort.

Deciding How To Summarize the Findings of Individual Studies and How To Present Findings

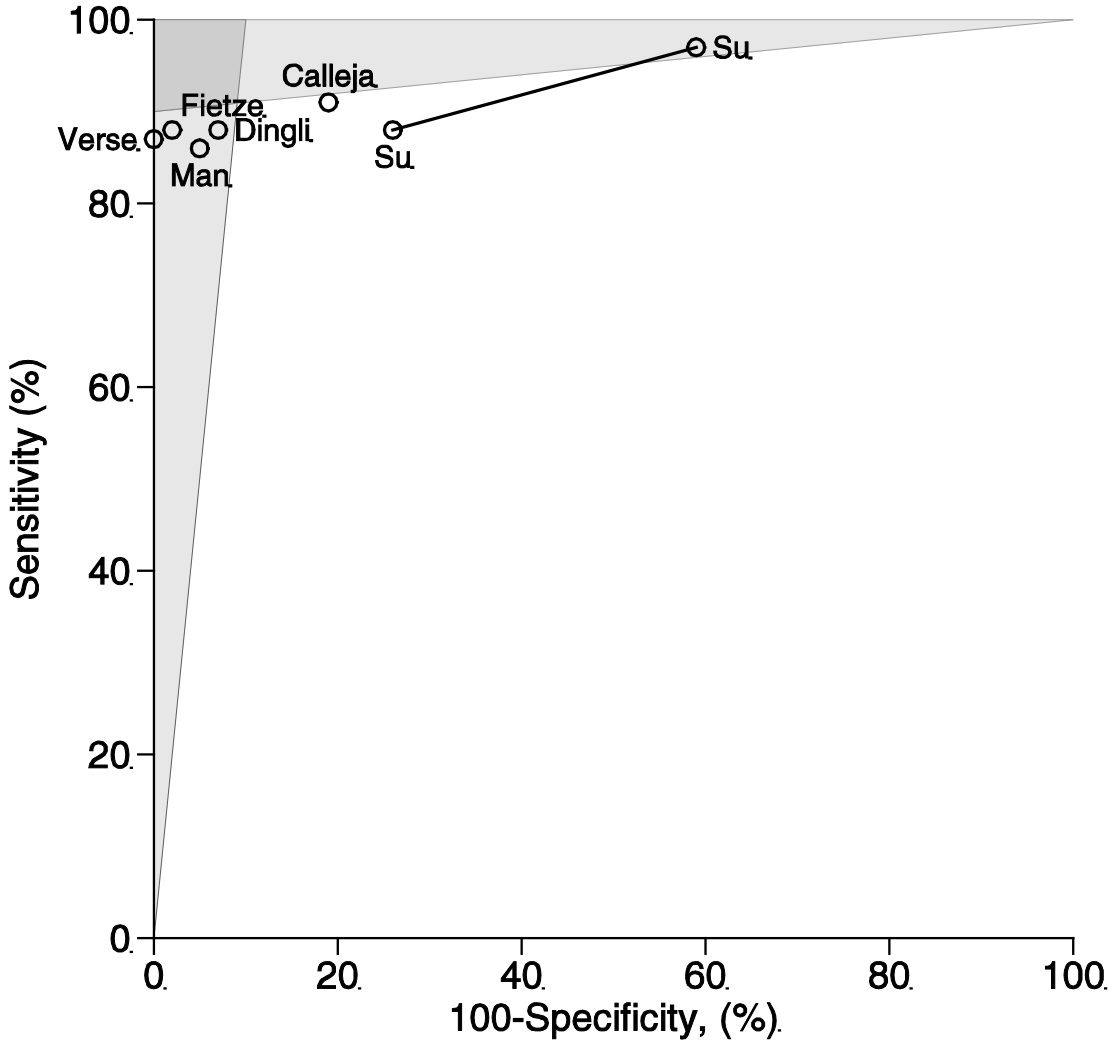
The reviewers calculated naïve estimates of sensitivity and specificity of portable monitors, and qualified their interpretation (option 3). They also performed complementary analyses outside the classical paradigm for evaluating test performance, to describe the concordance of measurements with portable monitors (index test) and facility-based polysomnography (reference test; this is option 2).

Qualitative Analyses of Naïve Sensitivity and Specificity Estimates

The reviewers depicted graphs of the naïve estimates of sensitivity and specificity in the ROC space (see Figure 9–3). These graphs suggest a high sensitivity and specificity of portable monitors to diagnose AHI$\geq$15 events per hour with facility-based polysomnography. However, it is very difficult to interpret these high values. First, there is considerable night-to-night variability in the measured AHI, as well as substantial between-rater and between-lab variability. Second, it is not easy to deduce whether the naïve estimates of sensitivity and specificity are underestimates or overestimates compared to the unknown "true" sensitivity and specificity to identify sleep apnea.

The systematic reviewers suggested that a better answer would be obtained by studies that perform a clinical validation of portable monitors (i.e., their ability to predict patients' history, risk propensity, or clinical profile—this would be option 1) and identified this as a gap in the pertinent literature.

Figure 9–3. "Naïve" estimates of the ability of portable monitors versus laboratory-based polysomnography to detect AHI>15 events/hour



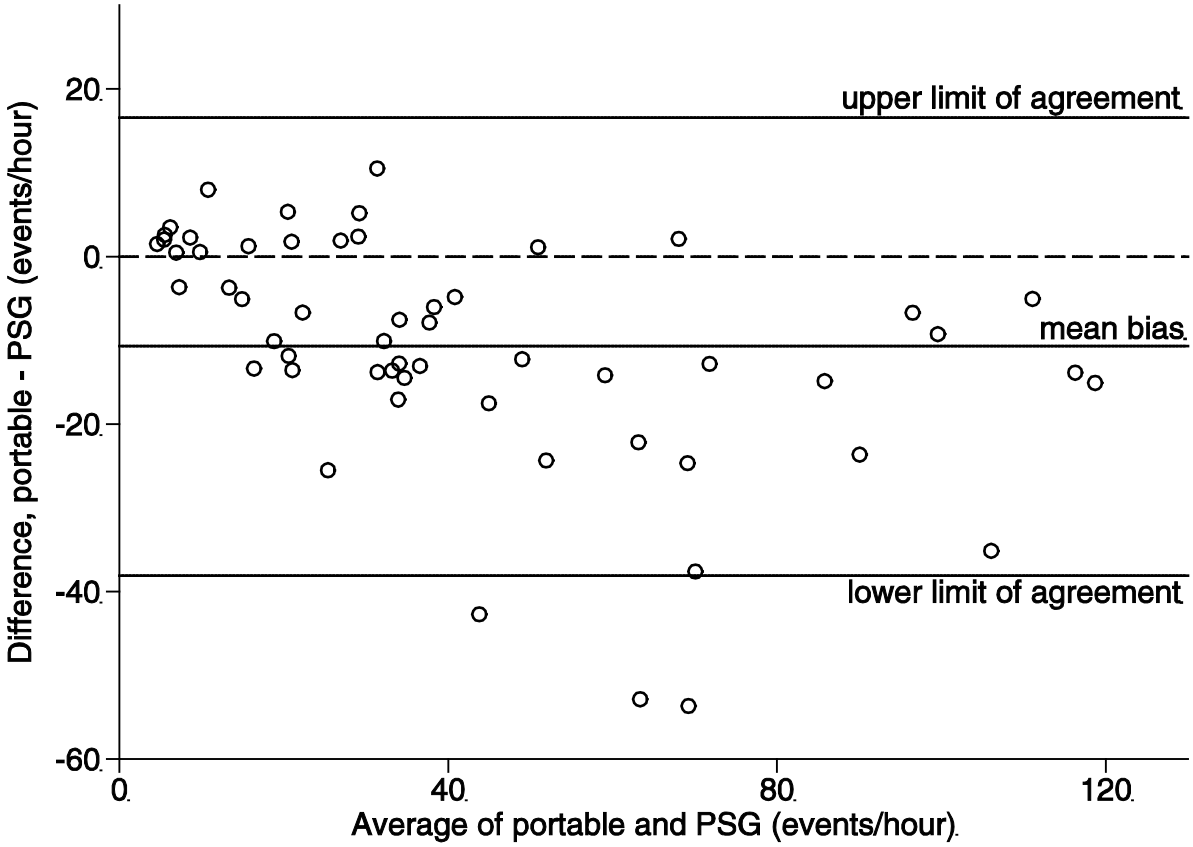
These data are on a subset of studies from the systematic review used in the illustration (studies that used manual scoring or combined manual and automated scoring for a type III portable monitor). Naive sensitivity/specificity pairs from the same study (obtained with different cutoffs for the portable monitor) are connected with lines. Studies lying on the left lightly shaded area have a positive likelihood ratio of 10 or more. Studies lying on the top lightly shaded area have a negative likelihood ratio of 0.1 or less. Studies lying on the intersection of the grey areas (darker grey polygon) have both a positive likelihood ratio more than 10 and a negative likelihood ratio less than 0.1.

Qualitative Assessment of the Concordance Between Measurement Methods

The systematic reviewers decided to summarize Bland-Altman type analyses to obtain information on whether facility-based polysomnography and portable monitors agree well enough to be used interchangeably. For studies that did not report Bland-Altman plots, the systematic reviewers performed these analyses using patient-level data, extracted by digitizing plots, from each study, An example is shown in Figure 9–4. The graph plots the differences between the two measurements against their average (which is the best estimate of the true unobserved value). An important piece of information gained from such analyses is the range of

values defined by the 95 percent limits of agreement (i.e., the region in which 95 percent of the differences are expected to fall). When the 95 percent limits of agreement are very broad, the agreement is suboptimal.

Figure 9–4. Illustrative example of a difference versus average analysis of measurements with facility-based polysomnography and portable monitors

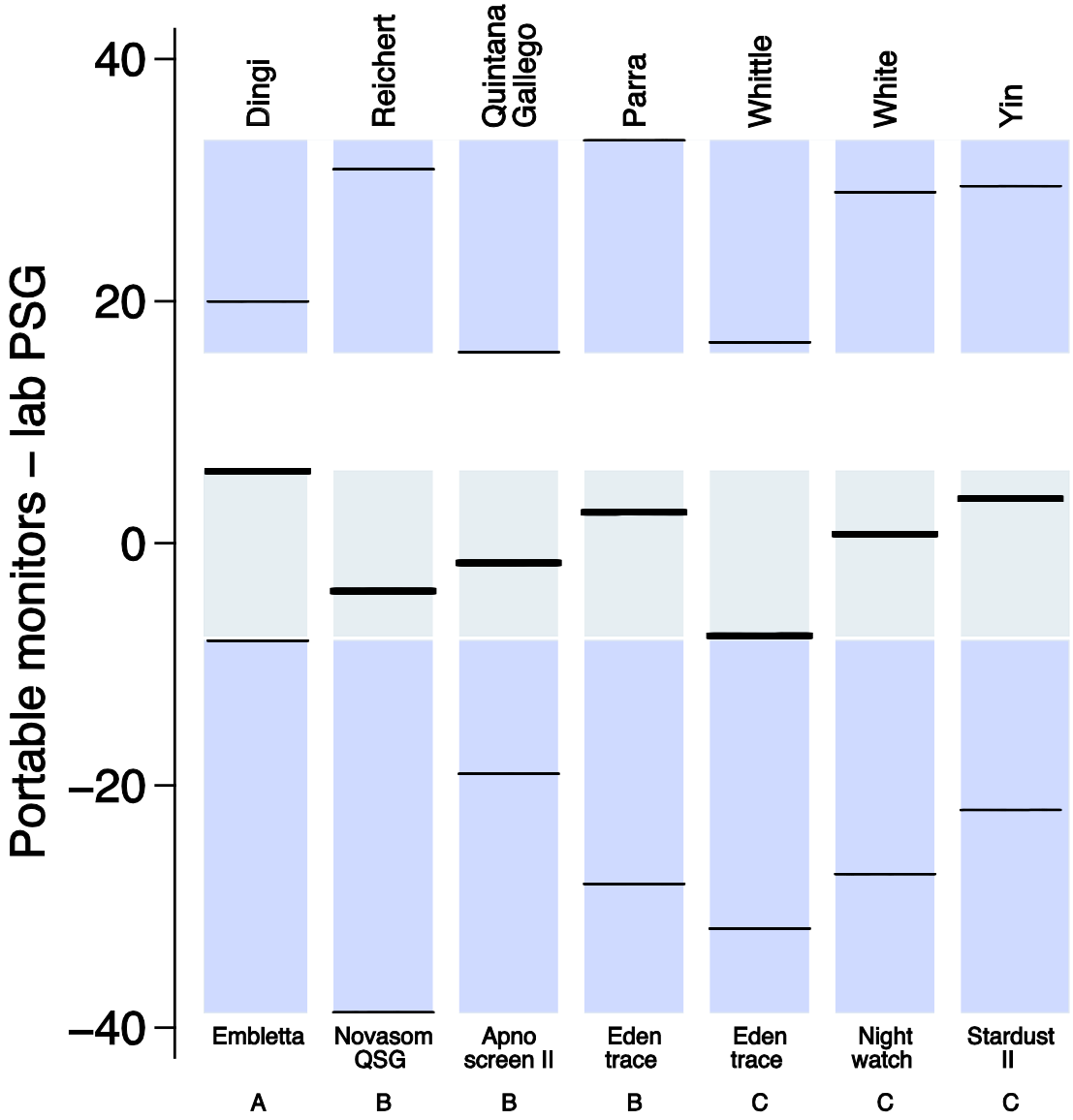


Digitized data from an actual study where portable monitors (Pro-Tech PTAF2 and Compumedics P2) were compared with facility-based polysomnography (PSG).[16] The dashed line at zero difference is the line of perfect agreement. The mean bias stands for the average systematic difference between the two measurements. The 95 percent limits of agreement are the boundaries within which 95 percent of the differences lie. If these are very wide and encompass clinically important differences, one may conclude that the agreement between the measurements is suboptimal. Note that the spread of the differences increases for higher measurement values. This indicates that the mean bias and 95 percent limits of agreement do not describe adequately the differences between the two measurements; differences are smaller for smaller values and larger for larger AHI values. In this example mean bias = -11 events/hour (95 percent limits of agreement: -38, 17), with statistically significant dependence of difference on average (Bradley-Blackwood F test, $p<0.01$).

Figure 9–5 summarizes such plots across several studies. For each study, it shows the mean difference in the two measurements (mean bias) and the 95 percent limits of agreement. The qualitative conclusion is that the 95 percent limits of agreement are very wide in most studies, suggesting great variability in the measurements with the two methods.

Thus, AHI measurements with the two methods generally agree on who has 15 or more events per hour of sleep (which is a low AHI). They disagree on the exact measurement among people who have larger measurements on average: One method may calculate 20 and the other 50 events per hour of sleep for the same person. The two methods are expected to disagree on who has AHI for those with >20, >30, or >40 events per hour.

Figure 9–5. Schematic representation of the mean bias and limits of agreement across several studies



Schematic representation of the agreement between portable monitors and facility-based polysomnography as conveyed by difference versus average analyses across seven studies. (The study of Figure 9–4 is not included.) The study author and the make of the monitor are depicted in the upper and lower part of the graph, respectively. The difference versus average analyses from each study are represented by three horizontal lines: a thicker middle line (denoting the mean bias); and two thinner lines, which represent the 95 percent limits of agreement and are symmetrically positioned above and below the mean bias line. The figure facilitates comparisons of the mean bias and the 95 percent limits of agreement across the studies by means of colored horizontal zones. The middle gray-colored zone shows the range of the mean bias in the seven studies, which is from +6 events per hour of sleep in the study by Dingi et al. (Embletta monitor) to -8 events per hour of sleep in the study by Whittle et al. (Edentrace monitor). The uppermost and lowermost shaded areas show the corresponding range of the upper 95 percent limits of agreement (upper purple zone) and the lower 95 percent limits of agreement (lower purple zone) in the seven studies.

Summary

In approaching a systematic review of the performance of a medical test, one is often faced with a reference standard which itself is subject to error. Four potential approaches are suggested:

1. Option 1. If possible, recast the assessment task in which the index test is used as an instrument to predict clinical outcomes. This reframing is potentially applicable only when measured patient outcomes are themselves valid. If so, the approach to such a review is detailed in Chapter 11 of this *Medical Test Methods Guide.*
2. Option 2. Assess the condordance in the results of the index and reference tests (i.e., treat them as two alternative measurement methods).
3. Option 3. Calculate naïve estimates of the index test's sensitivity and specificity from each study, but qualify study findings.
4. Option 4. Adjust the naïve estimates of sensitivity and specificity of the index test to account for the imperfect reference standard.

Systematic reviewers should decide which of the four options is more suitable for evaluating the performance of an index test versus an imperfect reference standard. To this end, the following considerations should be taken into account in the planning stages of the review: First, it is possible that multiple (imperfect) reference standard tests, or multiple cutoffs for the same reference test, are available. If an optimal choice is not obvious, the systematic reviewer should consider assessing more than one reference standard, or more than one cutoff for the reference test (as separate analyses). Whatever the choice, the implications of using the reference standard(s) should be described explicitly. Second, the reviewers should decide which option(s) for synthesizing test performance is (are) appropriate. The four options need not be mutually exclusive, and in some cases can be complementary (e.g., a naïve and "adjusted" analyses would reinforce assessments of a test if they both lead to similar clinical implications.) Finally, most of the analyses alluded to in option 4 would require expert statistical help; further, we have virtually no empirical data on the merits and pitfalls of methods that mathematically adjust for an imperfect reference standard. In our opinion, in most cases options 1 through 3 would provide an informative summary of the data.

References

1. Bossuyt PM. Interpreting diagnostic test accuracy studies. Semin Hematol 2008; 45(3):189-95.
2. Reitsma JB, Rutjes AW, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. J Clin Epidemiol 2009; 62(8):797-806.
3. Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess 2007; 11(50):iii, ix-51.
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. Trikalinos TA, Balion CM, Coleman CI, et al. Meta-analysis of test performance when there is a "gold standard." AHRQ Publication No. 12-EHC080-EF. Chapter 8 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, Maryland: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published in a spedial supplement to the Journal of General Internal Medicine, July 2012.
6. Reitsma JB, Rutjes AW, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. J Clin Epidemiol 2009; 62(8):797-806.

7. Jonas DE, Wilt TJ, Taylor BC, Wilkins TM, Matchar DB. Challenges in and principles for conducting systematic reviews of genetic tests used as predictive indicators. AHRQ Publication No. 12-EHC083-EF. Chapter 11 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.

8. Sun S. Meta-analysis of Cohen's kappa. Health Serv Outcomes Res Method 2011; 11:145-63.

9. Sokal RR, Rohlf EF. Biometry. New York: Freeman; 1981.

10. Bablok W, Passing H, Bender R, Schneider B. A general regression procedure for method transformation. Application of linear regression procedures for method comparison studies in clinical chemistry, Part III. J Clin Chem Clin Biochem 1988; 26(11):783-90.

11. Linnet K. Estimation of the linear relationship between the measurements of two methods with proportional errors. Stat Med 1990; 9(12):1463-73.

12. Linnet K. Performance of Deming regression analysis in case of misspecified analytical error ratio in method comparison studies. Clin Chem 1998; 44(5):1024-31.

13. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ 1995; 311(7003):485.

14. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999; 8(2):135-60.

15. Bland JM, Altman DG. Applying the right statistics: analyses of measurement studies. Ultrasound Obstet Gynecol 2003; 22(1):85-93.

16. Trikalinos TA, Ip S, Raman G, Cepeda MS, Balk EM, D'Ambrosio C et al. Home diagnosis of obstructive sleep apnea-hypopnea syndrome. Evidence Report/Technology Assessment. 127 pages. Rockville, MD: Agency for Healthcare Research and Quality; August 8, 2007. In Medicare's Technology Assessment series, at https://www.cms.gov/Medicare/Coverage/DeterminationProcess/downloads//id48TA.pdf. Accessed April 5, 2012.

17. Thompson IM, Pauler DK, Goodman PJ, Tangen CM, Lucia MS, Parnes HL et al. Prevalence of prostate cancer among men with a prostate-specific antigen level < or =4.0 ng per milliliter. N Engl J Med 2004; 350(22):2239-46.

18. Vacek PM. The effect of conditional dependence on the evaluation of diagnostic tests. Biometrics 1985; 41(4):959-68.

19. Gart JJ, Buck AA. Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. Am J Epidemiol 1966; 83(3):593-602.

20. Goldberg JD, Wittes JT. The estimation of false negatives in medical screening. Biometrics 1978; 34(1):77-86.

21. Gyorkos TW, Genta RM, Viens P, MacLean JD. Seroepidemiology of Strongyloides infection in the Southeast Asian refugee population in Canada. Am J Epidemiol 1990; 132(2):257-64.

22. Joseph L, Gyorkos TW. Inferences for likelihood ratios in the absence of a "gold standard". Med Decis Making 1996; 16(4):412-7.

23. Walter SD, Irwig L, Glasziou PP. Meta-analysis of diagnostic tests with imperfect reference standards. J Clin Epidemiol 1999; 52(10):943-51.

24. Walter SD, Irwig LM. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. J Clin Epidemiol 1988; 41(9):923-37.

25. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. Biometrics 2001; 57(1):158-67.

26. Black MA, Craig BA. Estimating disease prevalence in the absence of a gold standard. Stat Med 2002; 21(18):2653-69.

27. Dendukuri N, Hadgu A, Wang L. Modeling conditional dependence between diagnostic tests: a multiple latent variable model. Stat Med 2009; 28(3):441-61.

28. Garrett ES, Eaton WW, Zeger S. Methods for evaluating the performance of diagnostic tests in the absence of a gold standard: a latent class model approach. Stat Med 2002; 21(9):1289-307.

29. Hui SL, Zhou XH. Evaluation of diagnostic tests without gold standards. Stat Methods Med Res 1998; 7(4):354-70.

30. Qu Y, Tan M, Kutner MH. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. Biometrics 1996; 52(3):797-810.

31. Torrance-Rynard VL, Walter SD. Effects of dependent errors in the assessment of diagnostic test performance. Stat Med 1997; 16(19):2157-75.

32. Toft N, Jorgensen E, Hojsgaard S. Diagnosing diagnostic tests: evaluating the assumptions underlying the estimation of sensitivity and specificity in the absence of a gold standard. Prev Vet Med 2005; 68(1):19-33.

33. Albert PS, Dodd LE. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. Biometrics 2004; 60(2):427-35.

34. Alamanos Y, Voulgari PV, Drosos AA. Incidence and prevalence of psoriatic arthritis: a systematic review. J Rheumatol 2008; 35(7):1354-8.

35. Cantor T, Yang Z, Caraiani N, Ilamathi E. Lack of comparability of intact parathyroid hormone measurements among commercial assays for end-stage renal disease patients: implication for treatment decisions. Clin Chem 2006; 52(9):1771-6.

Funding: Funded by the Agency for Health Care Research and Quality (AHRQ) under the Effective Health Care Program.

Accessibility: Persons using assistive technology may not be able to fully access information in this report. For assistance contact EffectiveHealthCare@ahrq.hhs.gov.

Conflict of interest: None of the authors has affiliations or financial involvements that conflict with the information presented in this chapter.

Corresponding author: TA Trikalinos, Tufts Medical Center, 800 Washington St, Box#63, Boston, MA 02111, US. | Telephone: +1 617 636 0734 | Fax: +1 617 636 8628. Email: Thomas.Trikalinos@tufts.edu.

Suggested citation: Trikalinos TA, Balion CM. Options for summarizing medical test performance in the absence of a "gold standard." AHRQ Publication No. 12-EHC081-EF. Chapter 9 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, Maryland: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.

Chapter 10

Deciding Whether To Complement a Systematic Review of Medical Tests With Decision Modeling

Thomas A. Trikalinos, M.D., Tufts Evidence-based Practice Center, Boston, MA
Shalini Kulasingam, Ph.D., University of Minnesota School of Public Health, Minneapolis, MN
William F. Lawrence, M.D., M.S., Center for Outcomes and Evidence, Agency for Healthcare Research and Quality, Rockville, MD

Abstract

Limited by what is reported in the literature, most systematic reviews of medical tests focus on "test accuracy" (or better, test performance) rather than on the impact of testing on patient outcomes. The links between testing, test results, and patient outcomes are typically complex: even when testing has high accuracy, there is no guarantee that physicians will act according to tests results, that patients will follow their orders, or that the intervention will yield a beneficial endpoint. Therefore, test performance is typically not sufficient for assessing the usefulness of medical tests. Modeling (in the form of decision or economic analysis) is a natural framework for linking test performance data to clinical outcomes. We propose that (some) modeling should be considered to facilitate the interpretation of summary test performance measures by connecting testing and patient outcomes. We discuss a simple algorithm for helping systematic reviewers think through this possibility, and illustrate it by means of an example.

Introduction

In this chapter of the *Methods Guide to Medical Test Reviews* (also referred to as the Medical Test Methods Guide) we focus on modeling as an aid to understanding and interpreting the results of systematic reviews of medical tests.[1] Limited by what is reported in the literature, most systematic reviews focus on "test accuracy" (or better, test performance) rather than on the impact of testing on patient outcomes.[2,3] The links between testing, test results, and patient outcomes are typically complex: even when testing has high accuracy, there is no guarantee that physicians will act according to tests results, that patients will follow their orders, or that the intervention will yield a beneficial endpoint.[3] Therefore, test performance is typically not sufficient for assessing the usefulness of medical tests. Instead, one should compare complete test-and-treat strategies (for which test performance is but a surrogate), but such studies are very rare. Most often, evidence on diagnostic performance, effectiveness and safety of interventions and testing, patient adherence, and costs is available from different studies. Much like the pieces of a puzzle, these pieces of evidence should be put together to better interpret and contextualize the results of a systematic review of medical tests.[2,3] Modeling (in the form of decision or

economic analysis) is a natural framework for performing such calculations for test-and-treat strategies. It can link together evidence from different sources; explore the impact of uncertainty; make implicit assumptions clear; evaluate tradeoffs in benefits, harms and costs; compare multiple test-and-treat strategies that have never been compared head-to-head; and explore hypothetical scenarios (e.g., assume hypothetical interventions for incurable diseases).

This chapter focuses on modeling for enhancing the interpretation of systematic reviews of medical test accuracy, and does not deal with the much more general use of modeling as a framework for exploring complex decision problems. Specifically, modeling that informs broader decisionmaking may not fall within the purview of a systematic review. Whether or not to perform modeling for informing decisionmaking is often up to the decisionmakers themselves (e.g., policymakers, clinicians, or guideline developers), who would actually have to be receptive and appreciative of its usefulness.[4] Here we are primarily concerned with a narrower use of modeling, namely to facilitate the interpretation of summary test performance measures by connecting the link between testing and patient outcomes. This decision is within the purview of those planning and performing the systematic review. In all likelihood, it would be impractical to develop elaborate simulation models from scratch merely to enhance the interpretation of a systematic review of medical tests, but simpler models (be they decision trees or even Markov process-based simulations) are feasible even in a short time span and with limited resources.[4–6] Finally, how to evaluate models is discussed in guidelines for good modeling practices,[7–14] but not here.

Undertaking a modeling exercise requires technical expertise, good appreciation of clinical issues, and (sometimes extensive) resources, and should be pursued only when it is likely to be informative. So when is it reasonable to perform decision or cost effectiveness analyses to complement a systematic review of medical tests? We provide practical suggestions in the form of a stepwise algorithm.

A Workable Algorithm

Table 10–1 outlines a practical five-step approach that systematic reviewers could use to decide whether modeling could be used for interpreting and contextualizing the findings of a systematic review of test performance, within time and resource constraints. We outline these steps in an illustrative example at the end of the paper.

Table 10–1: Proposed algorithm to decide if modeling should be a part of the systematic review

Step	Description
1	Define how the test will be used.
2	Use a framework to identify consequences of testing as well as management strategies for each test result.
3	Assess if modeling is useful.
4	Evaluate prior modeling studies.
5	Consider whether modeling is practically feasible in the time frame given.

Step 1. Define how the test will be used.

The PICOTS typology (Population, Intervention, Comparators, Outcomes, Timing, Study design) is a widely adopted formalism for establishing the context of a systematic review.[15] It clarifies the setting of interest (whether the test will be used for screening, diagnosis, treatment guidance, patient monitoring, or prognosis) and the intended role of the medical test (whether it is the only test, an add-on to previously applied tests, or a tool for deciding on further diagnostic

workups). The information conveyed by the PICOTS items is crucial not only for the systematic review, but for planning a meaningful decision analysis as well.

Step 2. Use a framework to identify consequences of testing as well as management strategies for each test result.

Medical tests exert most of their effects in an indirect way. Notwithstanding the emotional, cognitive, and behavioral changes induced by testing and its results,[16] an accurate diagnosis in itself is not expected to affect patient-relevant outcomes. Nor do changes in test performance automatically result in changes in any patient-relevant outcome. From this point of view, test performance (as conveyed by sensitivity, specificity, positive and negative likelihood ratios, or other metrics) is only a surrogate end point. For example, testing for human immunodeficiency virus has both direct and indirect effects. The direct effects could include, but are not limited to, potential emotional distress attributable to the mere process of testing (irrespective of results); the cognitive and emotional benefits of knowing one's carrier status (for accurate results); perhaps the (very rare) unnecessary stress caused by a false positive diagnosis; or possible behavioral changes secondary to testing or its results. Indirect effects include all the downstream effects of treatment choices guided by the test results, such as benefits and harms of treatment in true positive diagnoses, avoidance of harms of treatment in true negative diagnoses, and cognitive and behavioral changes.

Identifying the consequences of testing and its results is a *sine qua non* for contextualizing and interpreting a medical test's (summary) sensitivity, specificity, and other measures of performance. A reasonable start is the analytic framework that was used to perform the systematic review (see the Introduction to this *Medical Test Methods Guide*).[15] This framework can be used to develop a basic tree illustrating test consequences and management options that depend on test results. Going through this exercise helps the reviewers make explicit the clinical scenarios of interest, the alternate (comparator) strategies, and the assumptions made by the reviewers regarding the test-and-treat strategies at hand.

Step 3. Assess whether modeling may be useful.

In most cases of evaluating medical testing, some type of formal modeling will be useful. This is because of the indirectness of the link between testing and health outcomes, and the multitude of test-and-treat strategies that can be reasonably contrasted. Therefore, it may be easier to examine the opposite question (i.e., when formal modeling may not be necessary or useful). We briefly explore two general cases. In the first, one of the test-and-treat strategies is clearly superior to all alternate strategies. In the second, information is too scarce regarding which modeling assumptions are reasonable, what the downstream effects of testing are, or what are plausible values of multiple central (influential) parameters.

The Case Where a Test-and-Treat Strategy Is a "Clear Winner"

A comprehensive discussion of this case is provided by Lord et al.[17,18] For some medical testing evaluations, one can identify a clearly superior test-and-treat strategy without any need for modeling. The most straightforward case is when there is direct comparative evidence for all the test-and-treat strategies of interest. Such evidence could be obtained from well designed, conducted and analyzed randomized trials, or even nonrandomized studies. Insofar as these studies are applicable to the clinical context of interest in the patient population of interest,

evaluate all important test-and-treat strategies, and identify a dominant strategy with respect to both benefits and harms and with adequate power, modeling may be superfluous. In all fairness, direct comparative evidence for all test-and-treat strategies of interest is exceedingly rare.

In the absence of direct comparisons of complete test-and-treat strategies, one can rely on test accuracy only, as long as it is known that the patients who are selected for treatment using different tests will have the same response to downstream treatments. Although the downstream treatments may be the same in all test-and-treat strategies of interest, one *cannot automatically deduce* that patients selected with different tests will exhibit similar treatment response.[3,15,17,18] Estimates of treatment effectiveness on patients selected with one test do not necessarily generalize to patients selected with another test. For example, the effectiveness of treatment for women with early-stage breast cancer is primarily based on cases diagnosed with mammography. Magnetic resonance imaging (MRI) can diagnose additional cases, but it is at best unclear whether these additional cases have the same treatment response.[19] We will return to this point soon.

If it were known that patient groups identified with different tests respond to treatment in the same way, one could select the most preferable test (test-and-treat strategy) based on considerations of test characteristics alone. Essentially, one would evaluate three categories of attributes: the cost and safety of testing; the sensitivity of the tests (ability to correctly identify those with the disease, and thus to proceed to hopefully beneficial interventions); and the specificity of the tests (ability to correctly identify those without disease, and thus avoid the harms and costs of unnecessary treatment). A test-and-treat strategy would be universally dominant if it were preferable versus all alternative strategies and over all three categories of attributes. In case of tradeoffs, i.e., one test has better specificity but another one is safer (with all other attributes being equal), one would have to explore these tradeoffs using modeling.

So how does one infer whether patient groups identified with different tests have (or should have) the same response to treatment? Several situations may be described. Randomized trials may exist suggesting that the treatment effects are similar in patients identified with different tests. For example, the effect of stenting versus angioplasty on reinfarctions in patients with acute myocardial infarction does not appear to differ by the test combinations used to identify the included patients.[20] Thus, when comparing various tests for diagnosing acute coronary events in the emergency department setting, test performance alone is probably a good surrogate for the clinical outcomes of the complete test-and-treat strategies. Alternatively, in the absence of direct empirical information from trials, one could use judgment to infer whether the cases detected from different tests would have a similar response to treatment:

1. Lord et al. propose that when the sensitivity of two tests is very similar, it is often reasonable to expect that the "case mix" of the patients who will be selected for treatment based on test results will be similar, and thus patients would repond to treatment in a similar way.[17,18] For example, Doppler ultrasonography and venography have similar sensitivity and specificity to detect the treatable condition of symptomatic distal deep venous thrombosis.[21] Because Doppler is easier, faster, and non-invasive, it is the preferable test.
2. When the sensitivities of the compared tests are different, it is more likely that the additional cases detected by the more sensitive tests may not have the same treatment response. In most cases this will not be known, and thus modeling would be useful to explore the impact of potential differential treatment response on outcomes. Sometimes we can reasonably extrapolate that treatment effectiveness will be unaltered in the

additional identified cases. This is when the tests operate on the same principle, and the clinical and biological characteristics of the additional identified cases are expected to remain unaltered. An example is computed tomography (CT) colonography for detection of large polyps, with positive cases subjected to colonoscopy as a confirmatory test. Dual positioning (prone and supine) of patients during the CT is more sensitive than supine-only positioning, without differences in specificity.[22] It is very reasonable to expect that the additional cases detected by dual positioning in CT will respond to treatment in the same way as the cases detected by supine-only positioning, especially since colonoscopy is a universal confirmatory test.

The Case of Very Scarce Information

There are times when we lack an understanding of the underlying disease processes to such an extent that we are unable to develop a credible model to estimate outcomes. In such circumstances, modeling is not expected to enhance the interpretation of a systematic review of test accuracy, and thus should not be performed with this goal in mind. This is a distinction between the narrow use of modeling we explore here (to contextualize the findings of a systematic review) and its more general use for decisionmaking purposes. Arguably, in the general decisionmaking case, modeling is especially helpful, because it is a disciplined and theoretically motivated way to explore alternative choices. In addition, it can help identify the major factors that contribute to the uncertainty, as is done in "value of information" analyses.[23,24]

Step 4. Evaluate prior modeling studies.

Before developing a model *de novo* or adapting an existing model, reviewers should consider searching the literature to ensure that the modeling has not already been done. There are several considerations when evaluating previous modeling studies.

First, reviewers need to judge the quality of the models. Several groups have made recommendations on evaluating the quality of modeling studies, especially in the context of cost-effectiveness analyses.[7,9–14] Evaluating the quality of a model is a very challenging task. More advanced modeling can be less transparent and difficult to describe in full technical detail. Increased flexibility often has its toll: Essential quantities may be completely unknown ("deep" parameters), and must be set through assumptions or by calibrating model predictions, versus real empirical data.[25] MISCAN-COLON[26,27] and SimCRC[28] are two microsimulation models that describe the natural history of colorectal cancer. Both assume an adenoma-carcinoma sequence for cancer development but differ in their assumptions on adenoma growth rates. Tumor dwell time (an unknown deep parameter in both models) was set to approximately 10 years in MISCAN-COLON;[27,29] and to approximately 30 years in SimCRC. Because of such differences, models can reach different conclusions.[30] Ideally, simulation models should be validated against independent datasets that are comparable to the datasets on which the models were developed.[25] External validation is particularly important for simulation models in which the unobserved deep parameters are set without calibration, based on assumptions and analytical calculations.[25,26]

Second, once the systematic reviewers deem that good quality models exist, they need to examine whether the models are applicable to the interventions and populations of the current evaluation; i.e., if they match the PICOTS items of the systematic review. In addition, the reviewers need to judge whether methodological and epidemiological challenges have been adequately addressed by the model developers.[3]

Third, the reviewers need to explore the applicability of the underlying parameters of the models. Most importantly, preexisting models will not have had the benefit of the current systematic review to estimate diagnostic accuracy, and they may have used estimates that differ from the ones obtained by the systematic review. Also, consideration should be given to whether our knowledge of the natural history of disease has changed since publication of the modeling study (thus potentially affecting parameters in the underlying disease model).

If other modeling papers meet these three challenges, then synthesizing the existing modeling literature may suffice. Alternatively, developing a new model may be considered, or reviewers could explore the possibility of cooperating with developers of existing high quality models to address the key questions of interest. The U.S. Preventive Services Task Force (USPSTF) and the Technology Assessment program of the Agency for Healthcare Research and Quality (AHRQ) have followed this practice for specific topics. For example, the USPSTF recommendations for colonoscopy screening[31] were informed by simulations based on the aforementioned MISCAN-COLON and SimCRC microsimulation models,[28,32] which were developed outside the EPC program.[26,27]

Step 5. Consider whether modeling is practically feasible in the given time frame.

Even if modeling is determined to be useful, it may still not be feasible to develop an adequately robust model within the context of a systematic review. Time and budgetary constraints, lack of experienced personnel, and other needs may all play a role in limiting the feasibility of developing or adapting a model to answer the relevant questions. Even if a robust and relevant model has been published, it may not necessarily be accessible. Models are often considered intellectual property of their developers or institutions, and they may not be unconditionally available for a variety of reasons. Further, even if a preexisting model is available, it may not be sufficient to address the key questions without extensive modifications by experienced and technically adept researchers. Additional data may be necessary, but they may not be available. Of importance, the literature required for developing or adapting a model does not necessarily overlap with that used for an evidence report.

Further, it may also be the case that the direction of the modeling project changes based on insights gained during the conduct of the systematic review or during the development of the model. Although this challenge can be mitigated by careful planning, it is not entirely avoidable.

If the systematic reviewers determine that a model would be useful but not feasible within the context of the systematic review, consideration should be given to whether these efforts could be done sequentially as related but distinct projects. The systematic review could synthesize available evidence, identify gaps, and estimate many necessary parameters for a model. The systematic review can also call for the development of a model in the future research recommendations section. A subsequent report that uses modeling could provide information on long-term outcomes.

Illustration

Here, we illustrate how the aformentioned algorithm could be applied, using an example of a systematic review of medical tests in which modeling was deemed important to contextualize findings on test performance.[33] Specifically, we discuss how the algorithm could be used to determine if a model is necessary for an evidence report on the ability of positron emission

tomography (PET) to guide the management of suspected Alzheimer's disease (AD), a progressive neurodegenerative disease for which current treatment options are at best modestly effective.[33] The report addressed three key questions, expressed as three clinical scenarios:

1. Scenario A: In patients with dementia, can PET be used to determine the type of dementia that would facilitate early treatment of AD and perhaps other dementia subtypes?
2. Scenario B: For patients with mild cognitive impairment, could PET be used to identify a group of patients with a high probability of AD so that they could start early treatment?
3. Scenario C: Is the available evidence enough to justify the use of PET to identify a group of patients with a family history of AD so that they could start early treatment?

The systematic review of the literature provides summaries of the diagnostic performance of PET to identify AD, but does not include longitudinal studies or randomized trials on the effects of PET testing on disease progression, mortality, or other clinical outcomes. In the absence of direct comparative data for the complete test-and-treat strategies of interest, decision modeling may be needed to link test results to long term patient-relevant outcomes.

Step 1: Define how PET will be used.

The complete PICOTS specification for the PET example is described in the evidence report[33] and is not reviewed here in detail. In brief, the report focuses on the *diagnosis* of the disease (AD) in the three scenarios of patients with suggestive symptoms. AD is typically diagnosed with a clinical examination that includes complete history, physical and neuropsychiatric evaluation, and screening laboratory testing.[34] In all three scenarios, we are only interested in PET as a "confirmatory" test (i.e., we are only interested in PET added to the usual diagnostic workup). Specifically, we assume that PET (1) is used for *diagnosing* patients with different severities or types of AD (mild or moderate AD, mild cognitive impairment, family history of AD), (2) it is an *add-on* to a clinical exam, and (3) it should be compared against the clinical examination (i.e. *no PET* as an add-on test). We are explicitly not evaluating patient management strategies where PET is the only test (i.e., PET "replaces" the typical examination) or where it triages who will receive the clinical examination (an unrealistic scenario). Table 10–2 classifies the results of PET testing.

Table 10–2. Cross-tabulation of PET results and actual clinical status among patients with initial clinical examination suggestive of Alzheimer's

	AD in Long-Term Clinical Evaluation	No AD in Long-Term Clinical Evaluation
PET Suggestive of AD	"True positive"	"False positive"
PET not Suggestive of AD	"False negative"	"True negative"

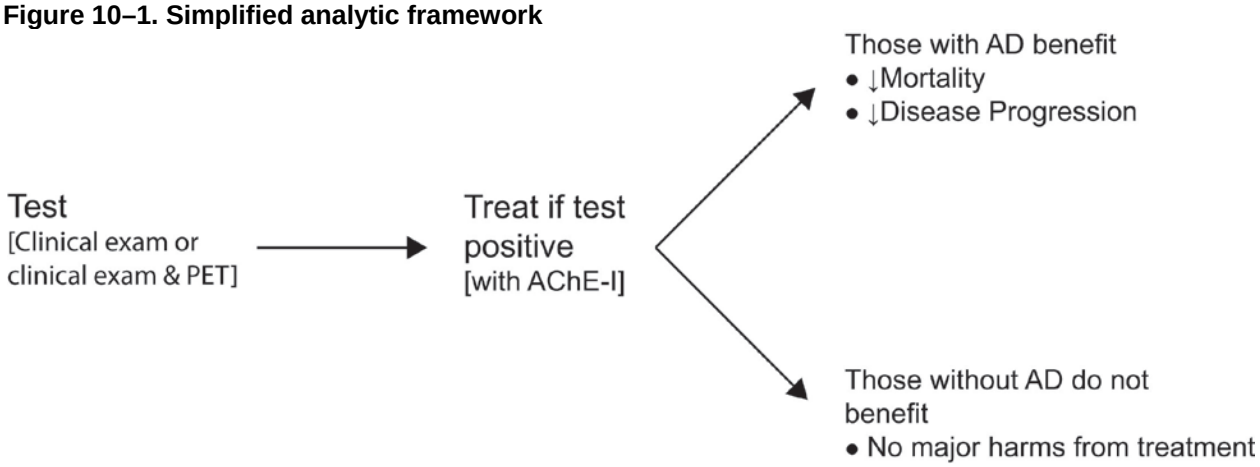
AD = Alzheimer's disease; PET = positron emission tomography
Counts in this table correspond to patients with an initial clinical examination suggestive of AD (as defined in the three clinical scenarios). Patients without suggestive clinical examination are not candidates for PET testing.

Step 2: Create a simplified analytic framework and outline how patient management will be affected by test results.

The PET evidence report does not document any appreciable direct effects or complications of testing with or without PET. Thus, it would be reasonable to consider all direct effects of testing as negligible when interpreting the results of the systematic review of test performance. A

simplified analytic framework is depicted in Figure 10–1, and represents the systematic reviewers' understanding of the setting of the test, and its role in the test-and-treat strategies of interest. The analytic framework also outlines the reviewers' understanding regarding the anticipated effects of PET testing on mortality and disease progression: any effects are only indirect, and conferred exclusively through the downstream clinical decision whether to treat patients. In the clinical scenarios of interest, patients with a positive test result (either by clinical examination or by the clinical examination–PET combination) will receive treatment. However, only those with AD (true positives) would benefit from treatment. Those who are falsely positive would receive no benefit but will still be exposed to the risk of treatment-related adverse effects, and the accompanying polypharmacy. (By design, the evidence report on which this illustration is based did not address costs, and thus we make no mention of costs here.)

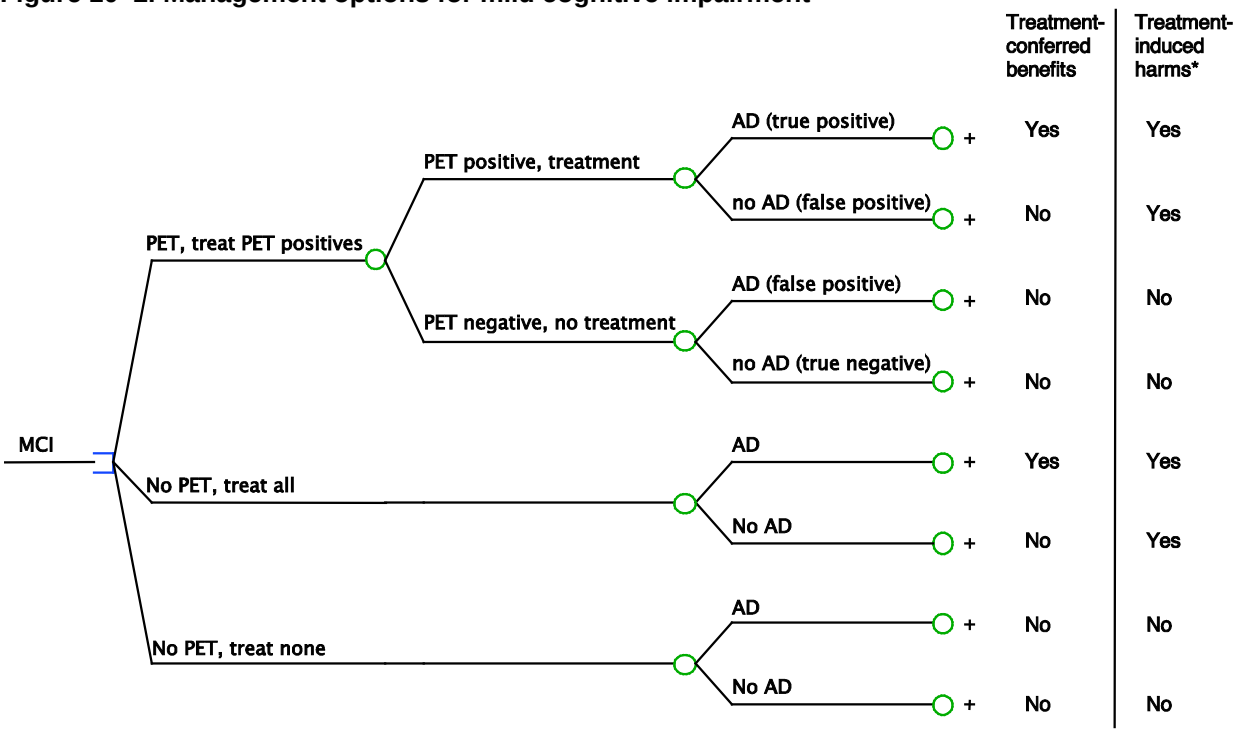
Figure 10–1. Simplified analytic framework



AD = Alzheimer's disease; AChE-I = acetylcholinesterase inhibitors (the treatment available at the time of the evidence report)
The framework assumes no major adverse effects from the treatment.

Figure 10–2 shows an outline of the management options in the form of a simple tree, for the clinical scenario of people with mild cognitive impairment (MCI) in the initial clinical exam (scenario B above). Similar basic trees can be constructed for the other clinical scenarios. The aim of this figure is to outline the management options for positive and negative tests (here they are simple: receive treatment or not) and the important consequences of being classified as a true positive, true negative, false positive or false negative, as well as to make explicit the compared test-and-treat strategies. This simplified outline is an overview of a decision tree for the specific clinical test.

Figure 10–2. Management options for mild cognitive impairment



AD = Alzheimer's disease; MCI = mild cognitive impairment; PET = positron emission tomography
*When applicable. As per the evidence report, the then-available treatment options (acetylcholinesterase inhibitors) do not have important adverse effects. However, in other cases, harms can be induced both by the treatment and the test (e.g., if the test is invasive). The evidence report also modeled hypothetical treatments with various effectiveness and safety profiles to gain insight on how sensitive their conclusions were to treatment characteristics. Note that at the time the evidence report was performed, other testing options for Alzheimer's were not in consideration.

Step 3: Assess whether modeling could be useful in the PET and AD evidence report.

In the example, no test-and-treat strategies have been compared head-to-head in clinical studies. Evidence exists to estimate the benefits and harms of pharmacologic therapy in those with and without AD. Specifically, the treatments for MCI in AD are at best only marginally effective,[33] and it is unknown whether subgroups of patients identified by PET may have differential responses to treatment. Hence, we cannot identify a "clear winner" based on test performance data alone. Thus, modeling was deemed useful here.

Step 4: Assess whether prior modeling studies could be utilized.

In this particular example, the systematic reviewers performed decision modeling. In addition to using the model to better contextualize their findings, they also explored whether their conclusions would differ if the treatment options were more effective than the options currently available. The exploration of such "what if" scenarios can inform the robustness of the conclusions of the systematic review, and can also be a useful aid in communicating conclusions to decisionmakers. It is not stated whether the systematic reviewers searched for prior modeling studies in the actual example. Although we do not know of specialized hedges to identify

modeling studies, we suspect that even simple searches using terms such as “model(s),” “modeling,” “simulat*”, or terms for decision or economic analysis would suffice.

Step 5. Consider whether modeling is practically feasible in the time frame given.

Obviously modeling was deemed feasible in the example at hand.

Overall Suggestions

Many systematic reviews of medical tests focus on test performance rather than the clinical utility of a test. Systematic reviewers should explore whether modeling may be helpful in enhancing the interpretation of test performance data, and in offering insight into the dynamic interplay of various factors on decision-relevant effects.

The five-step algorithm of Table 10–1 can help evaluate whether modeling is appropriate for the interpretation of a systematic review of medical tests.

References

1. Methods Guide for Medical Test Reviews. AHRQ Publication No. 12-EC017. Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published as a special supplement to the Journal of General Internal Medicine, July 2012.
2. Tatsioni A, Zarin DA, Aronson N, Samson DJ, Flamm CR, Schmid C et al. Challenges in systematic reviews of diagnostic technologies. Ann Intern Med 2005; 142(12 Pt 2):1048-1055.
3. Trikalinos TA, Siebert U, Lau J. Decision-analytic modeling to evaluate benefits and harms of medical tests: uses and limitations. Med Decis Making 2009; 29(5):E22-E29.
4. Claxton K, Ginnelly L, Sculpher M, Philips Z, Palmer S. A pilot study on the use of decision theory and value of information analysis as part of the NHS Health Technology Assessment programme. Health Technol Assess 2004; 8(31):1-103, iii.
5. Meltzer DO, Hoomans T, Chung JW, Basu A. Minimal Modeling Approaches to Value of Information Analysis for Health Research. AHRQ Publication No. 11-EHC062-EF. Rockville, MD: Agency for Healthcare Research and Quality; June 2011. http://ncbi.nlm.nih.gov/books/ NBK62146. Accessed April 10, 2012.
6. Trikalinos TA, Dahabreh IJ, Wong J, Rao M. Future Research Needs for the Comparison of Percutaneous Coronary Interventions with Bypass Graft Surgery in Nonacute Coronary Artery Disease: Identification of Future Research Needs. Future Research Needs Papers No. 1. AHRQ Publication No. 10-EHC068-EF. Rockville, MD: Agency for Healthcare Research and Quality; June 2010. http://ncbi.nlm.nih.gov/ books/NBK51079. Accessed April 10, 2012.
7. Weinstein MC, O'Brien B, Hornberger J, Jackson J, Johannesson M, McCabe C et al. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices--Modeling Studies. Value Health 2003; 6(1):9-17.
8. Trikalinos TA, Balion CM, Colemlan CI, et al. Meta-analysis of test performance when there is a "gold standard." AHRQ Publication No. 12-EHC080-EF. Chapter 8 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.

9. Sculpher M, Fenwick E, Claxton K. Assessing quality in decision analytic cost-effectiveness models. A suggested framework and example of application. Pharmacoeconomics 2000; 17(5):461-477.

10. Richardson WS, Detsky AS. Users' guides to the medical literature. VII. How to use a clinical decision analysis. B. What are the results and will they help me in caring for my patients? Evidence Based Medicine Working Group. JAMA 1995; 273(20):1610-1613.

11. Richardson WS, Detsky AS. Users' guides to the medical literature. VII. How to use a clinical decision analysis. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA 1995; 273(16):1292-1295.

12. Philips Z, Ginnelly L, Sculpher M, Claxton K, Golder S, Riemsma R et al. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. Health Technol Assess 2004; 8(36):iii-xi, 1.

13. Philips Z, Bojke L, Sculpher M, Claxton K, Golder S. Good practice guidelines for decision-analytic modelling in health technology assessment: a review and consolidation of quality assessment. Pharmacoeconomics 2006; 24(4):355-371.

14. Decision analytic modelling in the economic evaluation of health technologies. A consensus statement. Consensus Conference on Guidelines on Economic Modelling in Health Technology Assessment. Pharmacoeconomics 2000; 17(5):443-444.

15. Matchar DB. Introduction to the Methods Guide for Medical Test Reviews. AHRQ Publication No. 12-EHC073-EF. Chapter 1 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.

16. Bossuyt PM, McCaffery K. Additional patient outcomes and pathways in evaluations of testing. Med Decis Making 2009; 29(5):E30-E38.

17. Lord SJ, Irwig L, Bossuyt PM. Using the principles of randomized controlled trial design to guide test evaluation. Med Decis Making 2009; 29(5):E1-E12.

18. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med 2006; 144(11):850-855.

19. Irwig L, Houssami N, Armstrong B, Glasziou P. Evaluating new screening tests for breast cancer. BMJ 2006; 332(7543):678-679.

20. Nordmann AJ, Bucher H, Hengstler P, Harr T, Young J. Primary stenting versus primary balloon angioplasty for treating acute myocardial infarction. Cochrane Database Syst Rev 2005;(2):CD005313.

21. Gottlieb RH, Widjaja J, Tian L, Rubens DJ, Voci SL. Calf sonography for detecting deep venous thrombosis in symptomatic patients: experience and review of the literature. J Clin Ultrasound 1999; 27(8):415-420.

22. Fletcher JG, Johnson CD, Welch TJ, MacCarty RL, Ahlquist DA, Reed JE et al. Optimization of CT colonography technique: prospective trial in 180 patients. Radiology 2000; 216(3):704-711.

23. Janssen MP, Koffijberg H. Enhancing Value of Information Analyses. Value Health 2009.

24. Oostenbrink JB, Al MJ, Oppe M, Rutten-van Molken MP. Expected value of perfect information: an empirical example of reducing decision uncertainty by conducting additional research. Value Health 2008; 11(7):1070-1080.

25. Karnon J, Goyder E, Tappenden P, McPhie S, Towers I, Brazier J et al. A review and critique of modelling in prioritising and designing screening programmes. Health Technol Assess 2007; 11(52):iii-xi, 1.

26. Habbema JD, van Oortmarssen GJ, Lubbe JT, van der Maas PJ. The MISCAN simulation program for the evaluation of screening for disease. Comput Methods Programs Biomed 1985; 20(1):79-93.

27. Loeve F, Boer R, van Oortmarssen GJ, van BM, Habbema JD. The MISCAN-COLON simulation model for the evaluation of colorectal cancer screening. Comput Biomed Res 1999; 32(1):13-33.

28. National Cancer Institute, Cancer Intervention and Surveillance Modeling Network. http://cisnet.cancer.gov/colorectal/comparative.html. Accessed April 12, 2012.

29. Loeve F, Brown ML, Boer R, van BM, van Oortmarssen GJ, Habbema JD. Endoscopic colorectal cancer screening: a cost-saving analysis. J Natl Cancer Inst 2000; 92(7):557-563.

30. Zauber AG, Vogelaar I, Wilschut J, Knudsen AB, van Ballegooijen M, Kuntz KM. Decision analysis of coorectal cancer screening tests by age to begin, age to end and screening intervals: Report to the United States Preventive Services Task Force from the Cancer Intervention and Surveillance Modelling Network (CISNET) for July 2007. 2007.

31. Screening for colorectal cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 2008; 149(9):627-637.

32. Zauber AG, Lansdorp-Vogelaar I, Knudsen AB, Wilschut J, van BM, Kuntz KM. Evaluating test strategies for colorectal cancer screening: a decision analysis for the U.S. Preventive Services Task Force. Ann Intern Med 2008; 149(9):659-669.

33. Matchar DB, Kulasingam SL, McCrory DC, Patwardhan MB, Rutschmann OT, Samsa GP et al. Use of positron emission tomography and other neuroimaging techniques in the diagnosis and management of Alzheimer's disease and dementia. AHRQ Technology Assessment, Rockville, MD 2001; http://www.cms.gov/determinationprocess/downloads/id9TA.pdf. Accessed February 6, 2012.

34. Knopman DS, DeKosky ST, Cummings JL, Chui H, Corey-Bloom J, Relkin N et al. Practice parameter: diagnosis of dementia (an evidence-based review). Report of the Quality Standards Subcommittee of the American Academy of Neurology. Neurology 2001; 56(9):1143-1153.

Funding: Funded by the Agency for Health Care Research and Quality (AHRQ) under the Effective Health Care Program.

Conflict of interest: None of the authors has any affiliations or financial involvement that conflicts with the information presented in this chapter.

Corresponding author: TA Trikalinos, Tufts Medical Center, 800 Washington St, Box#63, Boston, MA 02111, | Telephone: +1 617 636 0734 | Fax: +1 617 636 8628. Email: Thomas.Trikalinos@tufts.edu.

Suggested citation: Trikalinos TA, Kulasingam S, Lawrence WF. Meta-analysis of test performance when there is a “gold standard.” AHRQ Publication No. 12-EHC082-EF. Chapter 10 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.

Chapter 11

Challenges in and Principles for Conducting Systematic Reviews of Genetic Tests Used as Predictive Indicators

Daniel E. Jonas, M.D., M.P.H., University of North Carolina
Timothy J. Wilt, M.D., M.P.H., Department of Veterans Affairs Health Care System, Minneapolis; University of Minnesota
Brent C. Taylor, Ph.D., M.P.H., Department of Veterans Affairs Health Care System, Minneapolis; University of Minnesota
Tania M. Wilkins, M.S., University of North Carolina
David B. Matchar, M.D., Duke-NUS Graduate Medical School Singapore; Duke University Medical Center, Durham, NC

Abstract

In this chapter we discuss common challenges in and principles for conducting systematic reviews of genetic tests. The types of genetic tests discussed are those used to (1) determine risk or susceptibility in asymptomatic individuals; (2) reveal prognostic information to guide clinical management in those with a condition; or (3) predict response to treatments or environmental factors. This chapter is not intended to provide comprehensive guidance on evaluating all genetic tests. Rather, it focuses on issues that have been of particular concern to analysts and stakeholders and on areas that are of particular relevance for the evaluation of studies of genetic tests.

Introduction

With recent advances in genotyping, it is expected that whole genome sequencing will soon be available for less than $1,000. Consequently, the number of studies of genetic tests will likely increase substantially, as will the need to evaluate studies of genetic tests. The general principles for evaluating genetic tests are similar to those for interpreting other prognostic or predictive tests, but there are differences in how the principles need to be applied and the degree to which certain issues are relevant, particularly when considering genetic test results that provide predictive rather than diagnostic information.

This chapter focuses on issues of particular concern to analysts and stakeholders and areas of particular relevance for the evaluation of studies of genetic tests. It is not intended to provide comprehensive guidance on evaluating all genetic tests. We reflect on genetic tests used to (1) determine risk or susceptibility in asymptomatic individuals (e.g., to identify individuals at risk for future health conditions, such as BRCA1 and BRCA2 for breast and ovarian cancer); (2) reveal prognostic information to guide clinical management and treatment in those with a condition (e.g., Oncotype Dx® for breast cancer recurrence, a test to evaluate the tumor genome

of surgically excised tumors from patients with breast cancer); or (3) predict response to treatments or environmental factors including diet (nutrigenomics), drugs (pharmacogenomics, such as CYP2C9 and VKORC1 tests to inform warfarin dosing), infectious agents, chemicals, physical agents, and behavioral factors. We do not address genetic tests used for diagnostic purposes. We address issues related to both heritable mutations and somatic mutations (e.g., genetic tests for tumors).

Clinicians, geneticists, analysts, policymakers, and other stakeholders may have varying definitions of what is considered a "genetic test." We have chosen to use a broad definition in agreement with that of the Centers for Disease Control and Prevention (CDC)–sponsored Evaluation of Genomic Applications in Practice and Prevention (EGAPP) and the Secretary's Advisory Committee on Genetics, Health, and Society,[1] namely: "A genetic test involves the analysis of chromosomes, deoxyribonucleic acid (DNA), ribonucleic acid (RNA), genes, or gene products (e.g., enzymes and other proteins) to detect heritable or somatic variations related to disease or health. Whether a laboratory method is considered a genetic test also depends on the intended use, claim, or purpose of a test."[1] The same technologies are used for diagnostic and predictive genetic tests; it is the intended use of the test result that determines whether it is a diagnostic or predictive test.

In this chapter we discuss principles for addressing challenges related to developing the topic and structuring a genetic test review (context and scoping), as well as performing the review. This chapter is meant to complement the Methods Guide for Comparative Effectiveness Reviews.[2] We do not attempt to reiterate the challenges and principles described in earlier sections of this *Medical Test Methods Guide*, but focus instead on issues of particular relevance for evaluating studies of genetic tests. Although we have written this chapter to serve as guidance for the Agency for Healthcare Research and Quality (AHRQ) Evidence-based Practice Centers (EPCs), we also intend it as a useful resource for other investigators interested in conducting systematic reviews on genetic tests.

Common Challenges

Genetic tests are different from other medical tests in their relationship to the outcomes measured. Reviewers need to take into account the penetrance of the disease, time lag to outcomes, variable expressivity, and pleiotropy (as defined below). These particular aspects of genetic tests result in specific actions at various stages of planning and performing the review. Both single-gene and polygenic disorders are known. Single gene disorders are the result of a single mutated gene and may be passed on to subsequent generations in various well described ways (e.g., autosomal dominant, autosomal recessive, X-linked). Polygenic disorders are the result of the combined action of more than one gene and are not inherited according to simple Mendelian patterns. Some examples include heart disease and diabetes. Some of the terms described below (penetrance, variable expressivity, and pleiotropy) are generally used to describe single-gene disorders.

Penetrance

Evaluations of predictive genetic tests should always consider penetrance, defined as "the proportion of people with a particular genetic change who exhibit signs and symptoms of a disorder."[3] Penetrance is a key factor in determining the future risk of developing disease and assessing the overall clinical utility of predictive genetic tests. Sufficient data to determine precise estimates of penetrance are sometimes lacking.[4–5] This can be due to a lack of reliable prevalence data or of long-term outcomes data. In such cases, determining the overall clinical

utility of a genetic test is difficult. In some cases, modeling with sensitivity analyses can help develop estimates.[4]

Time Lag

The time lag between genetic testing and clinically important events should be assessed in critical appraisal of studies of such tests. Whether the duration of studies is sufficient to characterize the relationship between positive tests and clinical outcomes is an important consideration. In addition, it should be determined whether or not subjects have reached the age beyond which clinical expression is likely.

Variable Expressivity

Variable expressivity refers to the range of severity of the signs and symptoms that can occur in different people with the same condition.[3] For example, the features of hemochromatosis vary widely. Some individuals have mild symptoms, while others experience life-threatening complications such as liver failure. The degree of expressivity should be considered in the evaluation of genetic tests.

Pleiotropy

Pleiotropy occurs when a single gene influences multiple phenotypic traits. For example, the genetic mutation causing Marfan syndrome results in cardiovascular, skeletal, and ophthalmologic abnormalities. Similarly, BRCA mutations can increase the risk of a number of cancers, including breast, ovarian, prostate, and melanoma.

Other Common Challenges

Another common challenge for the evaluation of predictive genetic tests is that direct evidence is often lacking on the impact of the test results on health outcomes. The evidence base is often too limited in scope to evaluate the clinical utility of the test. In addition, it is often difficult to find published information on various aspects of genetic tests, especially data related to analytic validity. For example, laboratory-developed tests (LDT) are regulated by the Centers for Medicare & Medicaid Services (CMS) under Clinical Laboratory Improvement Act (CLIA) regulations for clinical laboratories. CLIA does not require clinical validation and many LDTs have had no clinical validation or clinical utility studies.

Genetic tests also raise a number of technical issues particularly relevant to the assessment of their analytic validity. These technical issues may differ according to the type of genetic test and may influence the interpretation of a genetic test result. Technical issues may also differ depending on the specimen being tested. For example, there are different considerations when assessing tumor genomes as opposed to human genomes.

Several common challenges arise in using genetic tests to determine susceptibility or risk in asymptomatic individuals. The utility of such tests may depend on the ability of respondents, such as the patient or their relative, to report and identify certain clinical factors. For instance, if patients cannot accurately recall the family history of a heritable disease, it can be difficult to assess their risk of developing the disease.

Finally, statistical issues must be taken into account when evaluating studies of genetic tests. For example, genetic test results are often derived from analytically complex studies that have undergone a very large number of statistical tests, creating a high risk of Type I error (i.e., when a spurious association is deemed significant).

Principles for Addressing the Challenges

The eight principles described in this section can be used to address the challenges related to developing, structuring, and performing a genetic test review (Table 11–1).

Table 11–1. Principles for addressing common challenges when evaluating genetic tests used as predictive indicators

Principle 1: Use an organizing framework appropriate for genetic tests.
Principle 2: Develop analytic frameworks that reflect the predictive nature of genetic tests and incorporate appropriate outcomes.
Principle 3: Search databases appropriate for genetic tests.
Principle 4: Consult with experts to determine which technical issues are important to address in assessing genetic tests.
Principle 5: Distinguish between functional assays and DNA-based assays to determine important technical issues.
Principle 6: Evaluate case-control studies carefully for potential selection bias.
Principle 7: Determine the added value of the genetic test over existing risk assessment approaches.
Principle 8: Understand statistical issues of particular relevance to genetic tests.

Principle 1: Use an organizing framework appropriate for genetic tests.

Organizing frameworks for evaluating genetic tests have been developed by the United States Preventive Services Task Force (USPSTF), the CDC, and EGAPP.[1,6–7] The model endorsed by the EGAPP initiative[1] was based on a previous report of the NIH Task Force on Genetic Testing[8] and developed through a CDC-sponsored project, which piloted an evidence evaluation framework that applied the following three criteria: (1) analytic validity (technical accuracy and reliability); (2) clinical validity (ability to detect or predict an outcome, disorder, or phenotype); and (3) clinical utility (whether use of the test to direct clinical management improves patient outcomes). A fourth criterion was added: (4) ethical, legal, and social implications.[6] The ACCE model (Analytic validity, Clinical validity, Clinical utility, and Ethical, legal and social implications) includes a series of 44 questions that are useful for analysts in defining the scope of a review, as well as for critically appraising studies of genetic tests (Table 11-2). The initial seven questions help to guide an understanding of the disorder, the setting, and the type of testing. A detailed description of the methods of the EGAPP Working Group is published elsewhere.[1]

Principle 2: Develop analytic frameworks that reflect the predictive nature of genetic tests and incorporate appropriate outcomes.

It is important to have a clear definition of the clinical scenario and analytic framework when evaluating any test, including a predictive genetic test. Prior to performing a review, analysts should develop clearly defined key questions and understand the needs of decisionmakers and the context in which the tests are used. They should consider whether this is a test used for determining future risk of disease in asymptomatic individuals, establishing prognostic information that will influence treatment decisions, or predicting response to treatments (either effectiveness or harms)—or whether it is used for some other purpose. They should clarify the type of specimens used for the genetic test under evaluation (i.e., patient genome or tumor genome). The PICOTS typology (Patient population, Intervention, Comparator, Outcomes, Timing, Setting) should be clearly described, as it will inform the development of the analytic framework and vice versa.

Table 11–2. ACCE model questions for reviews of genetic tests[6]

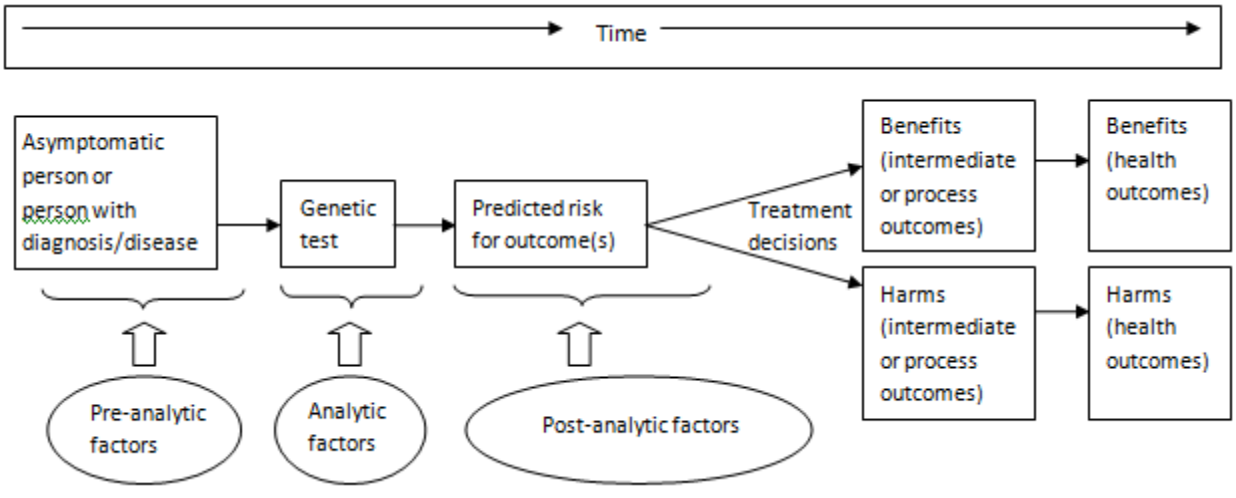
Element	Questions
Disorder/setting	1. What is the specific clinical disorder to be studied? 2. What are the clinical findings defining this disorder? 3. What is the clinical setting in which the test is to be performed? 4. What DNA test(s) are associated with this disorder?* 5. Are preliminary screening questions employed? 6. Is it a stand-alone test or is it one of a series of tests? 7. If it is part of a series of screening tests, are all tests performed in all instances (parallel) or are only some tests performed on the basis of other results (series)?
Analytic validity	8. Is the test qualitative or quantitative? 9. How often is the test positive when a mutation is present?* 10. How often is the test negative when a mutation is not present?* 11. Is an internal quality control program defined and externally monitored? 12. Have repeated measurements been made on specimens? 13. What is the within-and between-laboratory precision? 14. If appropriate, how is confirmatory testing performed to resolve false positive results in a timely manner? 15. What range of patient specimens have been tested? 16. How often does the test fail to give a usable result? 17. How similar are results obtained in multiple laboratories using the same, or different technology?
Clinical validity	18. How often is the test positive when the disorder is present? 19. How often is the test negative when a disorder is not present? 20. Are there methods to resolve clinical false positive results in a timely manner? 21. What is the prevalence of the disorder in this setting? 22. Has the test been adequately validated on all populations to which it may be offered? 23. What are the positive and negative predictive values? 24. What are the genotype/phenotype relationships?* 25. What are the genetic, environmental or other modifiers?*
Clinical utility	26. What is the natural history of the disorder? 27. What is the impact of a positive (or negative) test on patient care? 28. If applicable, are medical tests available? 29. Is there an effective remedy, acceptable action, or other measurable benefit? 30. Is there general access to that remedy or action? 31. Is the test being offered to a socially vulnerable population? 32. What quality assurance measures are in place? 33. What are the results of pilot trials? 34. What health risks can be identified for follow-up testing and/or intervention? 35. What are the financial costs associated with testing? 36. What are the economic benefits associated with actions resulting from testing? 37. What facilities/personnel are available or easily put in place? 38. What educational materials have been developed and validated and which of these are available?* 39. Are there informed consent requirements?* 40. What methods exist for long term monitoring? 41. What guidelines have been developed for evaluating program performance?
Ethical, legal, and social implications	42. What is known about stigmatization, discrimination, privacy/confidentiality and personal/family social issues?* 43. Are there legal issues regarding consent, ownership of data and/or samples, patents, licensing, proprietary testing, obligation to disclose, or reporting requirements?* 44. What safeguards have been described and are these safeguards in place and effective?*

ACCE = Analytic validity, Clinical validity, Clinical utility, and Ethical, legal and social implications;
DNA = deoxyribonucleic acid

*Many of the questions in this Table (or variants of the questions) are relevant for evaluating most medical tests, not just genetic tests. Those with an asterisk (questions 4, 9, 10, 24, 25, 38, 39, 42, 43, and 44) are relevant only for evaluating genetic tests or may require extra scrutiny when evaluating genetic tests.

In constructing an analytic framework, it may be useful for analysts to consider preanalytic, analytic, and postanalytic factors particularly applicable to genetic tests (described later in this chapter), as well as the key outcomes of interest. Analytic frameworks should incorporate the factors and outcomes of greatest interest to decision makers. Figure 11–1 illustrates a generic analytic framework for evaluating predictive genetic tests that can be modified as necessary for various situations.

Figure 11–1. Generic analytic framework for evaluating predictive genetic tests

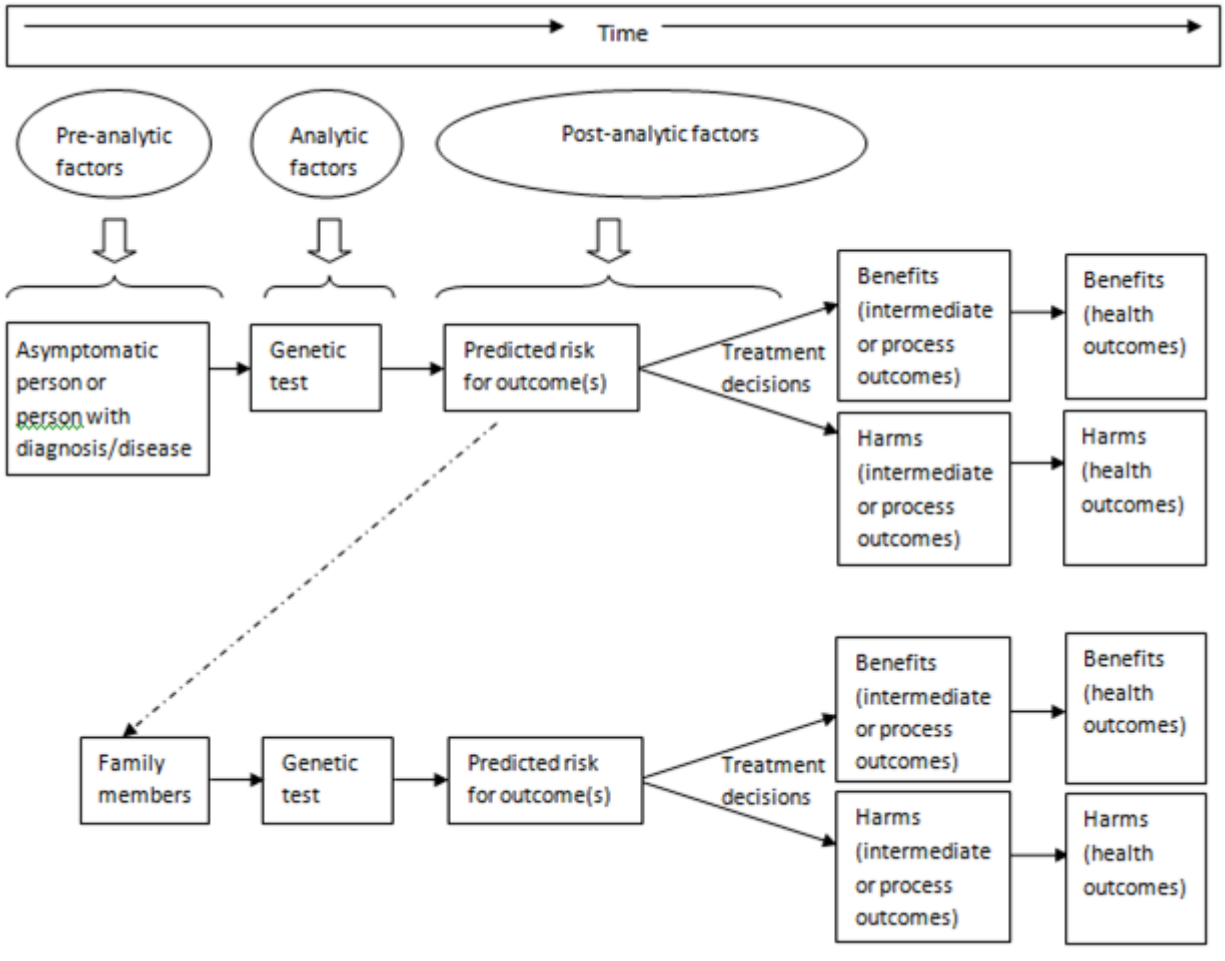


In addition to effects on family members, psychological distress and possible stigmatization or discrimination are potential harms that may result from predictive genetic tests, particularly if the test results predict probability of disease occurring with a high likelihood, especially if no proven preventive or ameliorative measures are available. For these potential harms, analysts should take into account whether the testing is for inherited or acquired genetic mutations, since these factors influence the potential for harms. In addition, whether the condition related to the test is multifactorial or follows classic Mendelian inheritance will affect the potential for these harms.

Other important outcomes to consider when evaluating genetic tests include, but are not limited to, cost, quality of life, long-term morbidity, and indirect impact. Genetic tests may have an impact that is difficult to measure, such as impact on important decisions regarding pregnancy.

Depending on the context, the impact of genetic testing on family members may be important, particularly in cases that involve testing for heritable conditions. One approach to including family members in the analytic framework is illustrated in Figure 11–2.

Figure 11–2. Generic analytic framework for evaluating predictive genetic tests when the impact on family members is important



Principle 3: Search databases appropriate for genetic tests.

The Human Genome Epidemiology Network (HuGE Net) Web site can provide a helpful supplement to searches, as it includes many meta-analyses of genetic association studies as well as a source called the HuGE Navigator that can identify all types of available studies related to a genetic test.[9]

The U.S. Food and Drug Administration (FDA)–approved test package inserts for genetic tests contain summaries of the analytic validity data. These summaries can be retrieved through searches of the gray literature. Package inserts are available on the FDA and manufacturer Web sites. Laboratory-developed tests do not require FDA clearance, and there is no requirement for publicly available data on analytic validity. When there are no published data on analytic validity of a genetic test, the external proficiency testing program carried out jointly by the American College of Medical Genetics (ACMG) and the College of American Pathologists (CAP) can be useful in establishing the degree of laboratory-to-laboratory variability, as well as some sense of reproducibility.[10–12] Other potentially useful sources of unpublished data include conference publications from professional societies (e.g., the College of American Pathologists), the GeneTests Web site (www.genetests.org), the Association for Molecular Pathology Web site (www.amp.org), CDC programs (e.g., the Genetic Testing Reference Materials Coordination

Program and the Newborn Screening Quality Assurance Program), and international proficiency testing programs.[13]

An AHRQ "horizon scan" found two databases—LexisNexis® (www.lexisnexis.com) and Cambridge Healthtech Institute (CHI) (www.healthtech.com/)—that had high utility in identifying genetic tests in development for clinical cancer care. A number of others had low-to-moderate utility, and some were not useful.[14]

Principle 4: Consult with experts to determine which technical issues are important to address in assessing genetic tests.

There are a number of technical issues related to analytic validity that can influence the interpretation of a genetic test result, including preanalytic, analytic, and postanalytic factors.[15–16] In general, preanalytic steps are those involved in obtaining, fixing or preserving, and storing samples prior to staining and analysis. Important analytic variables include the type of assay chosen and its reliability, types of samples, the specific analyte investigated, specific genotyping methods, timing of sample analysis, and complexity of performing the assay. Postanalytic variables relate to the complexity of interpreting the test result, variability from laboratory to laboratory, and quality control.[15–16] To determine which of these technical issues are pertinent for a given review, comparative effectiveness review teams should include or consult with molecular pathologists, geneticists, or others familiar with the issues related to the process of performing and reporting genetic tests. Table 11–3 summarizes some of the preanalytic, analytic, and postanalytic questions that should be addressed.

For genetic testing of tumor specimens, it is important to understand that the tumor genome may be in a dynamic state, with mutations emerging over time (e.g., due to drug exposure or disruption of cellular repair). Tumor specimens will often contain normal cells from the patient as well as tumor cells. To accurately assess for somatic mutations using tumor specimens, particular strategies may be needed, such as enriching samples for tumor cells (e.g., by microscopic evaluation and dissection of the cells).

Table 11–3. Questions for assessing preanalytic, analytic, and postanalytic factors for evaluating predictive genetic tests*

Element	Questions
Preanalytic	What patient characteristics are relevant to the analytic validity of the test (e.g., age, sex, ethnicity, race, ancestry, parental history of consanguinity, family health history)? What types of samples were used? How were samples obtained? How were samples handled and stored prior to analysis?
Analytic	What type of assay was used? What is the reliability of the assay? What specific analyte was investigated (e.g., specification of which alleles, genes, or biochemical analytes were evaluated)? For DNA-based tests, what is the definition of the genotype investigated? Did the study test for all potentially relevant alleles? For DNA-based tests, what genotyping methods were used? When were samples analyzed (compared to when they were collected)? Was the timing of analysis equal for both study groups (if applicable)? How often does the test give a usable result? (i.e., What is the "call rate"?)
Postanalytic	How are the test results interpreted and applied? How complex is interpretation and application? What quality control measures were used? Were repeated measurements made on specimens? How reproducible is the test over time? How reproducible is the test when repeated in the same patient multiple times? How reproducible is the test from laboratory to laboratory?

*Portions of this table were adapted from Burke et al., 2002[15] and Little et al., 2002.[16]

Principle 5: Distinguish between functional assays and DNA-based assays to determine important technical issues.

Some studies may utilize DNA-based assays, whereas others may utilize functional assays with different sensitivities and specificities. Functional assays, in which a substrate or product of a metabolic process affected by a particular genetic polymorphism is measured, may have the advantage of showing potentially more important information than the presence of the genetic polymorphism itself. However, they may be affected by a number of factors and do not necessarily reflect the polymorphism alone. Unmeasured environmental factors, other genetic polymorphisms, and various disease states may influence the results of functional assays. In addition, functional assays that measure enzyme activity are taken at a single point in time. Depending on the enzyme and polymorphism being evaluated, the variation in enzyme activity over time should be considered in critical appraisal. Inconsistent results have been reported between studies using DNA-based molecular methods and those using phenotypic assays.[16–18]

For DNA-based tests, a variety of sample sources are available (e.g., blood, cheek swab, hair) that should hypothetically result in identical genotype results.[16,19–23] However, DNA may be more difficult to obtain and purify from some tissues than from blood, particularly if the tissues have been fixed in paraffin versus fresh samples. (DNA extraction from formalin-fixed tissue is difficult, but sometimes possible).[16] Some studies utilize different sources of DNA for cases and controls, introducing potential measurement bias from differences in ease of technique and test accuracy. Extraction of DNA from tumors in oncology studies may raise additional issues that influence analytic validity, including the quantity of tissue, admixture of normal and cancerous tissue, amount of necrosis, timing of collection, and storage technique (e.g., fresh, frozen, paraffin, formalin).[16]

When evaluating DNA-based molecular tests, complexity of the test method, laboratory-to-laboratory variability, and quality control should be assessed. A number of methods are available for genotyping single nucleotide polymorphisms that vary in complexity and potential for polymorphism misclassification.[16,24–26] Considering laboratory reporting of internal controls and repetitive experiments can be useful in assessment of overall analytic validity. The method of interpreting test results may influence complexity as well. For example, some tests require visual inspection of electrophoresis gels. Inter-observer variability should be considered for such tests.[16,27]

Principle 6: Evaluate case-control studies carefully for potential selection bias.

In critical appraisal of any case-control study, it is important to determine whether cases and controls were selected from the same source population. In the case of genetic studies, the geographic location of the population does not suffice. Rather, having cases and controls matched for ethnicity/race or ancestry (i.e., population stratification) is important, since the frequencies of DNA polymorphisms vary from population to population. It has been noted that many case-control studies of gene-disease associations have selected controls from a population that does not represent the population from which the cases arose.[16–17,28–30] In general, only nested case-control studies could have low enough potential for selection bias to provide reliable information.

Principle 7: Determine the added value of the genetic test over existing risk assessment approaches.

For some scenarios, a number of clinical factors associated with risk assessment or susceptibility may already be well characterized. In such cases, comparative effectiveness reviews should determine the added value of using genetic testing along with known factors, compared with using the known factors alone. For example, age, sex, smoking, hypertension, diabetes, and cholesterol are all well established risk factors for cardiovascular disease. Risk stratification of individuals to determine cholesterol-lowering targets is based on these factors.[31] Assessment of newly identified polymorphisms—such as those described on chromosome 9p21[32]—that may confer increased risk of cardiovascular disease and have potential implications for medical interventions should be evaluated in the context of these known risk factors. In this scenario, investigators should determine the added value of testing for polymorphisms of chromosome 9p21 in addition to known clinical risk factors.

Multiple polymorphisms may be associated with risk of disease, prognosis, or prediction of drug response. In such cases, the effect of multiple polymorphisms can be explored using a multiple regression model. Once this is done, prospective studies would usually be needed to determine whether the model, including the genetic tests, has clinical utility. For example, VKORC1 and CYP2C9 genotypes have been associated with warfarin dose requirements in multiple regression models. In order to determine whether tests for VKORC1 and CYP2C9 have clinical utility, studies would need to compare the use of a prediction model that contains the genetic tests in combination with known clinical factors that affect warfarin dose (e.g., age, BMI) with the use of clinical factors alone.[33–35]

Principle 8: Understand statistical issues of particular relevance to genetic tests.

Hardy-Weinberg Equilibrium

In population genetics, most allele distributions follow a usual distribution, known as the Hardy-Weinberg equilibrium (HWE). Genetic association studies should generally report whether the frequencies of the alleles being evaluated follow HWE. There are a number of reasons that distributions may deviate from HWE, including new mutations, selection, migration, genetic drift, and inbreeding.[36] In addition, when numerous polymorphisms are tested for associations with diseases or outcomes, as in many genome-wide association studies, many of them (5 percent) will deviate from HWE based on chance alone (related to multiple testing).[37] Deviation from HWE may be a clue to bias and genotyping error, but it is not specific and possibly not sensitive.[37] Analysts should consider whether studies have tested for and reported HWE. A more detailed discussion of this topic as it relates to genetic association studies has been published elsewhere.[36–37]

Sample Size Calculations

When assessing internal validity of studies, it is important to assess whether sample size calculations appropriately accounted for the number of variant alleles and the prevalence of variants in the population of interest. This is particularly relevant for pharmacogenomic studies evaluating the functional relevance of genetic polymorphisms.[38] Such studies often enroll an insufficient number of subjects to account for the number of variant alleles and the prevalence of variants in the population.[38]

Genetic Association Studies and Multiple Comparisons

Genetic test results are sometimes derived from analytically complex studies that have undergone a very large number of statistical tests. These may be in the form of genome-wide association studies searching for associations between a huge number of genetic polymorphisms and health conditions. Such association studies may enhance understanding of the importance of genetics in relation to a variety of health conditions, but should generally be used to generate hypotheses rather than to test hypotheses or to confirm cause-effect relationships.[16] Close scrutiny should be applied to ensure that the evidence for the association has been validated in multiple studies to minimize both potential confounding and potential publication bias issues. In addition, reviewers should note whether appropriate adjustments for multiple comparisons were used. Many investigators recommend using a P value of less than 5×10^{-8} for the threshold of significance in large genome-wide studies.[37,39–40] Other approaches include assessing the false positive report probability and controlling the false discovery rate.[41–43]

When a genetic mutation associated with increased risk is present, evaluating potential causality can be difficult, as many factors other than the mutation may influence associations. These include environmental exposures, behaviors, and other genes. Many genetic variants identified that are thought to influence susceptibility to diseases are associated with low relative and absolute risk.[16,44] Thus, exclusion of non-causal explanations for associations and consideration of potential confounders are central to critical appraisal of such associations. It may also be important to explore biologic plausibility (e.g., from *in vitro* studies) to help support or oppose theories of causation.[16]

Overlapping Data Sets

Be cautious of publications that report prevalence estimates for genetic variants that have actually arisen from overlapping data sets.[16] For example, genome-wide association studies or other large collaborative efforts, such as the International Warfarin Pharmacogenomics Consortium, may pool samples of patients that were previously included in other published studies.[3] To the degree possible, investigators should identify overlapping data sets and avoid double-counting. It may be useful to organize evidence tables by study time period and geographic area to identify potential overlapping data sets.[16]

Assessing Tumor Genetics

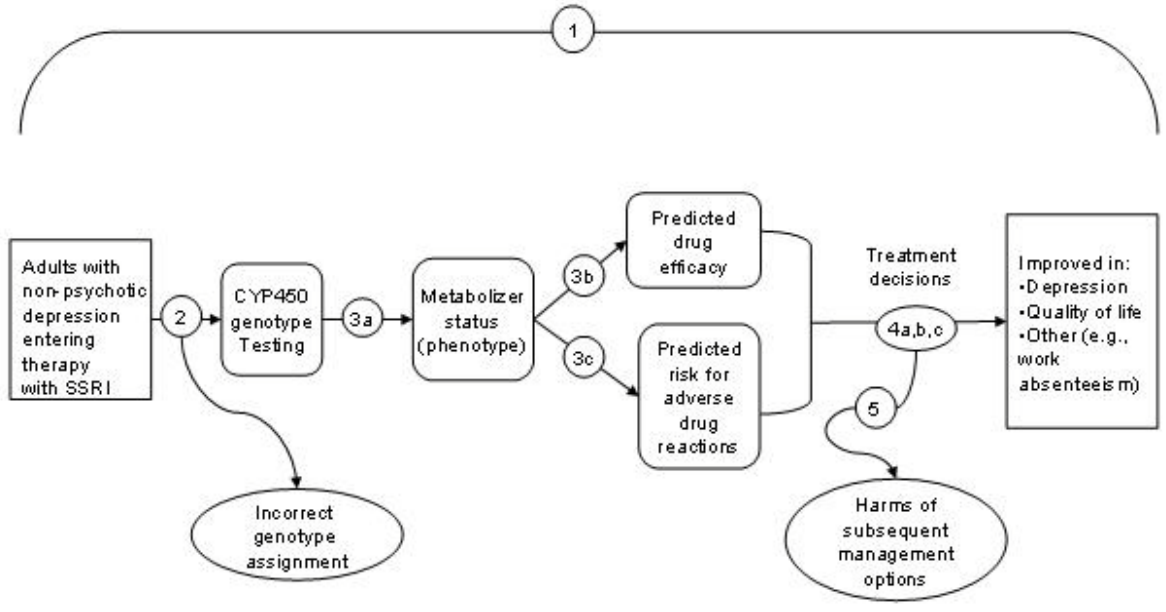
As mentioned under Principle 4, it is important to understand that a tumor genome may be in a dynamic state. In addition, tumor specimens will often contain normal cells from the patient. The characteristics of the specimen will influence the sensitivity and operating characteristics of the test. Tests with greater sensitivity may be required when specimens contain both normal cells and tumor cells.

Illustrations

Since the completion of the Human Genome Project, the Hap Map project, and related works, there have been a great number of publications describing the clinical validity of genetic test results (e.g., gene-disease associations), but far fewer studies of their clinical utility. A review of genetic testing for cytochrome P450 polymorphisms in adults with depression treated with selective serotonin reuptake inhibitors (SSRIs) developed an analytic framework and five corresponding key questions which, taken together, provide an example of a well defined predictive genetic test scenario that explores a potential chain of evidence relating to intermediate outcomes **(**Figure 11–3).[45] The authors found no prospective studies with clinical outcomes that used genotyping to guide treatment. They constructed a chain of questions to

assess whether sufficient indirect evidence could answer the overarching question by evaluating the links between genotype and metabolism of SSRIs (phenotype), metabolism and SSRI efficacy, and metabolism and adverse drug reactions to SSRIs.

Figure 11–3. Analytic framework for evidence gathering on CYP450 genotype testing for SSRI treatment of depression.



CYP450 = cytochrome p450; SSRI = selective serotonin reuptake inhibitor
Numbers in this figure represent the research questions addressed in the systematic review:[45]
1 (overarching question): Does testing for cytochrome P450 (CYP450) polymorphisms in adults entering selective serotonin reuptake inhibitor (SSRI) treatment for non-psychotic depression lead to improvement in outcomes, or are testing results useful in medical, personal, or public health decisionmaking?
2: What is the analytic validity of tests that identify key CYP450 polymorphisms?
3a: How well do particular CYP450 genotypes predict metabolism of particular SSRIs? Do factors such as race/ethnicity, diet, or other medications, affect this association?
3b: How well does CYP450 testing predict drug efficacy? Do factors such as race/ethnicity, diet, or other medications, affect this association?
3c: How well does CYP450 testing predict adverse drug reactions? Do factors such as race/ethnicity, diet, or other medications, affect this association?
4a: Does CYP450 testing influence depression management decisions by patients and providers in ways that could improve or worsen outcomes?
4b: Does the identification of the CYP450 genotypes in adults entering SSRI treatment for non-psychotic depression lead to improved clinical outcomes compared to not testing?
4c: Are the testing results useful in medical, personal or public health decisionmaking?
5: What are the harms associated with testing for CYP450 polymorphisms and subsequent management options?

An EPC report on HER2 testing to manage patients with breast cancer and other solid tumors provides a detailed assessment of challenges in conducting a definitive evaluation of preanalytic, analytic, and postanalytic factors when there is substantial heterogeneity or lack of available information related to the methods of testing.[46] The authors noted that it had been only very recently that many aspects of HER2 assays were standardized, and that the effects of widely varying testing methods could not be isolated. Thus, they approached this challenge by providing a narrative review for their first key question (What is the evidence on concordance and discrepancy rates for methods [e.g., FISH, IHC, etc.] used to analyze HER2 status in breast tumor tissue?).

Additional considerations arise when evaluating genetic test results used to determine susceptibility or risk in asymptomatic individuals. The utility of such tests may depend on the ability of patients and providers to report and identify certain clinical factors. For example, a review of genetic risk assessment and BRCA mutation testing underscores the importance of accurately determining family history.[4,47] The analytic framework begins by classifying asymptomatic women into high, moderate, or average risk categories. This is a good example of incorporating a key preanalytic factor (family history), that has an important influence on analytic validity. Tests for BRCA mutations may be used to predict the risk for breast and ovarian cancer in high-risk women (i.e., those with a family history suggesting increased risk). However, because we do not know all of the genes that contribute to hereditary breast and ovarian cancer and because analytic methods to detect mutations in the known genes are not perfect, population-based testing for hereditary susceptibility to breast and ovarian cancer is currently not an appropriate strategy. Rather, family history-based testing is the paradigm that is recommended to guide the use of BRCA testing.[4,47]

Thus, family history is a genetic/genomics tool that is used to (1) identify people with possible inherited disease susceptibilities, (2) guide genetic testing strategies, (3) help interpret genetic test results, and (4) assess disease risk. The ability of providers to accurately determine a family history that confers increased risk is a key prerequisite to the utility of BRCA mutation and other predictive genetic testing. It is sometimes difficult for people to accurately recall the presence of a condition in their relatives. Sensitivity and specificity of self-reported family history are important in determining overall usefulness of predictive genetic testing.[4]

Conclusions

Analysts should understand common challenges, and apply the principles for addressing those challenges, when conducting systematic reviews of genetic tests used as predictive indicators. Key points include:

1. The general principles that apply in evaluating genetic tests are similar to those for other prognostic or predictive tests, but there are differences in how the principles need to be applied or the degree to which certain issues are relevant.
2. A clear definition of the clinical scenario and an analytic framework is important when evaluating *any* test, including genetic tests.
3. Organizing frameworks and analytic frameworks are useful constructs for approaching the evaluation of genetic tests.
4. In constructing an analytic framework for evaluating a genetic test, analysts should consider preanalytic, analytic, and postanalytic factors; such factors are useful when assessing analytic validity.
5. Predictive genetic tests are generally characterized by a delayed time between testing and clinically important events.
6. Published information on the analytic validity of some genetic tests may be difficult to find. Web sites (FDA or diagnostic companies) and gray literature may be important sources.
7. In situations where clinical factors associated with risk are well characterized, comparative effectiveness reviews should assess the added value of using genetic testing along with known factors, compared with using the known factors alone.
8. For genome-wide association studies, reviewers should determine whether the association has been validated in multiple studies to minimize both potential confounding and publication bias. In addition, reviewers should note whether appropriate adjustments for multiple comparisons were used.

References

1. Teutsch SM, Bradley LA, Palomaki GE, et al. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) initiative: methods of the EGAPP Working Group. Genet Med. 2008.
2. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. AHRQ Publication No. 10(11)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality; March 2011; Available at: www.effectivehealthcare.ahrq.gov. Accessed August 22, 2011.
3. Lister Hill National Center for Biomedical Communications: Collections of the National Library of Medicine. What are reduced penetrance and variable expressivity? [electronic resource]; 2008; Available at: http://ghr.nlm.nih.gov/handbook/inheritance/penetranceexpressivity. Accessed August 22, 2011.
4. Nelson HD, Huffman LH, Fu R, Harris EL. Genetic risk assessment and BRCA mutation testing for breast and ovarian cancer susceptibility: systematic evidence review for the U.S. Preventive Services Task Force. Annals of internal medicine. 2005;143(5):362-79.
5. Whitlock EP, Garlitz BA, Harris EL, Beil TL, Smith PR. Screening for hereditary hemochromatosis: a systematic review for the U.S. Preventive Services Task Force. Annals of internal medicine. 2006;145(3):209-23.
6. National Office of Public Health Genomics C. ACCE Model Process for Evaluating Genetic Tests. 2007; Available at: http://www.cdc.gov/genomics/gtesting/ACCE/index.htm. Accessed August 22, 2011.
7. Harris RP, Helfand M, Woolf SH, et al. Current methods of the US Preventive Services Task Force: a review of the process. American journal of preventive medicine. 2001;20(3 Suppl):21-35.
8. Task Force on Genetic Testing (NIH). Promoting Safe and Effective Genetic Testing in the United States. Final Report of the Task Force on Genetic Testing. 1997; Available at: http://www.genome.gov/10001733. Accessed August 22, 2011.
9. Khoury MJ, Dorman JS. The Human Genome Epidemiology Network. American journal of epidemiology. 1998;148(1):1-3.
10. Palomaki GE, Bradley LA, Richards CS, Haddow JE. Analytic validity of cystic fibrosis testing: a preliminary estimate. Genet Med. 2003;5(1):15-20.
11. Palomaki GE, Haddow JE, Bradley LA, FitzSimmons SC. Updated assessment of cystic fibrosis mutation frequencies in non-Hispanic Caucasians. Genet Med. 2002;4(2):90-4.
12. Palomaki GE, Haddow JE, Bradley LA, Richards CS, Stenzel TT, Grody WW. Estimated analytic validity of HFE C282Y mutation testing in population screening: the potential value of confirmatory testing. Genet Med. 2003;5(6):440-3.
13. Sun F, Bruening W, Erinoff E**,** Schoelles KM. Addressing Challenges in Genetic Test Evaluation. Evaluation Frameworks and Assessment of Analytic Validity. Methods Research Report (Prepared by the ECRI Institute Evidence-based Practice Center under Contract No. 290-2007-10063-I.) AHRQ Publication No. 11-EHC048-EF. Rockville, MD: Agency for Healthcare Research and Quality; June 2011. Available at: www.effectivehealthcare.ahrq.gov/reports/final.cfm.
14. Agency for Healthcare Research and Quality. *Genetic Tests for Cancer*. Technology Assessment. Rockville, MD: Agency for Healthcare Research and Quality; 2006; Available at: http://archive.ahrq.gov/clinic/ta/gentests/. Accessed August 22, 2011.
15. Burke W, Atkins D, Gwinn M, et al. Genetic test evaluation: information needs of clinicians, policy makers, and the public. American journal of epidemiology. 2002;156(4):311-8.
16. Little J, Bradley L, Bray MS, et al. Reporting, appraising, and integrating data on genotype prevalence and gene-disease associations. American journal of epidemiology. 2002;156(4):300-10.
17. Brockton N, Little J, Sharp L, Cotton SC. N-acetyltransferase polymorphisms and colorectal cancer: a HuGE review. American journal of epidemiology. 2000;151(9):846-61.
18. d'Errico A, Malats N, Vineis P, Boffetta P. Review of studies of selected metabolic polymorphisms and cancer. IARC scientific publications. 1999(148):323-93.

19. Yang M, Hendrie HC, Hall KS, Oluwole OS, Hodes ME, Sahota A. Improved procedure for eluting DNA from dried blood spots. Clinical chemistry. 1996;42(7):1115-6.

20. Gale KB, Ford AM, Repp R, et al. Backtracking leukemia to birth: identification of clonotypic gene fusion sequences in neonatal blood spots. Proceedings of the National Academy of Sciences of the United States of America. 1997;94(25):13950-4.

21. Walker AH, Najarian D, White DL, Jaffe JF, Kanetsky PA, Rebbeck TR. Collection of genomic DNA by buccal swabs for polymerase chain reaction-based biomarker assays. Environmental health perspectives. 1999;107(7):517-20.

22. Harty LC, Shields PG, Winn DM, Caporaso NE, Hayes RB. Self-collection of oral epithelial cell DNA under instruction from epidemiologic interviewers. American journal of epidemiology. 2000;151(2):199-205.

23. Garcia-Closas M, Egan KM, Abruzzo J, et al. Collection of genomic DNA from adults in epidemiological studies by buccal cytobrush and mouthwash. Cancer Epidemiol Biomarkers Prev. 2001;10(6):687-96.

24. Hixson JE, Vernier DT. Restriction isotyping of human apolipoprotein E by gene amplification and cleavage with HhaI. Journal of lipid research. 1990;31(3):545-8.

25. Tobe VO, Taylor SL, Nickerson DA. Single-well genotyping of diallelic sequence variations by a two-color ELISA-based oligonucleotide ligation assay. Nucleic acids research. 1996;24(19):3728-32.

26. Lee LG, Connell CR, Bloch W. Allelic discrimination by nick-translation PCR with fluorogenic probes. Nucleic acids research. 1993;21(16):3761-6.

27. Bogardus ST, Jr., Concato J, Feinstein AR. Clinical epidemiological quality in molecular genetic research: the need for methodological standards. JAMA. 1999;281(20):1919-26.

28. Botto LD, Yang Q. 5,10-Methylenetetrahydrofolate reductase gene variants and congenital anomalies: a HuGE review. American journal of epidemiology. 2000;151(9):862-77.

29. Dorman JS, Bunker CH. HLA-DQ locus of the human leukocyte antigen complex and type 1 diabetes mellitus: a HuGE review. Epidemiologic reviews. 2000;22(2):218-27.

30. Cotton SC, Sharp L, Little J, Brockton N. Glutathione S-transferase polymorphisms and colorectal cancer: a HuGE review. American journal of epidemiology. 2000;151(1):7-32.

31. National Cholesterol Education Program. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report; 2002 Dec 17. Report No.: 1524-4539 (Electronic).

32. Schunkert H, Gotz A, Braund P, et al. Repeated replication and a prospective meta-analysis of the association between chromosome 9p21.3 and coronary artery disease. Circulation. 2008;117(13):1675-84.

33. Gage BF, Eby C, Milligan PE, Banet GA, Duncan JR, McLeod HL. Use of pharmacogenetics and clinical factors to predict the maintenance dose of warfarin. Thromb Haemost. 2004;91(1):87-94.

34. Gage BF, Lesko LJ. Pharmacogenetics of warfarin: regulatory, scientific, and clinical issues. J Thromb Thrombolysis. 2008;25(1):45-51.

35. Jonas DE, McLeod HL. Genetic and clinical factors relating to warfarin dosing. Trends Pharmacol Sci. 2009;30(7):375-86.

36. Attia J, Ioannidis JP, Thakkinstian A, et al. How to use an article about genetic association: A: Background concepts. JAMA. 2009;301(1):74-81.

37. Attia J, Ioannidis JP, Thakkinstian A, et al. How to use an article about genetic association: B: Are the results of the study valid? JAMA. 2009;301(2):191-7.

38. Williams JA, Johnson K, Paulauskis J, Cook J. So many studies, too few subjects: establishing functional relevance of genetic polymorphisms on pharmacokinetics. Journal of clinical pharmacology. 2006;46(3):258-64.

39. Hoggart CJ, Clark TG, De Iorio M, Whittaker JC, Balding DJ. Genome-wide significance for dense SNP and resequencing data. Genetic epidemiology. 2008;32(2):179-85.

40. McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature reviews. 2008;9(5):356-69.

41. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. Behav Brain Res. 2001;125(1-2):279-84.

42. Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J Natl Cancer Inst. 2004;96(6):434-42.

43. Ziegler A, Konig IR, Thompson JR. Biostatistical aspects of genome-wide association studies. Biom J. 2008;50(1):8-28.

44. Caporaso N. Selection of candidate genes for population studies. In: Vineis P, Malats N, Lang M, et al, editors. Metabolic polymorphisms and susceptibility to cancer. Lyon, France: IARC Monogr Eval Carcinog Risks Hum; 1999. p. 23-36.

45. Matchar DB, Thakur ME, Grossman I, et al. Testing for cytochrome P450 polymorphisms in adults with non-psychotic depression treated with selective serotonin reuptake inhibitors (SSRIs). Evid Rep Technol Assess (Full Rep). 2007(146):1-77.

46. Seidenfeld J, Samson DJ, Rothenberg BM, Bonnell CJ, Ziegler KM, Aronson N. HER2 Testing to Manage Patients With Breast Cancer or Other Solid Tumors/Technology Assessment No. 172. Rockville, MD: (Prepared by Blue Cross and Blue Shield Association Technology Evaluation Center Evidence-based Practice Center, under Contract No. 290-02-0026.) 2008.

47. Genetic risk assessment: recommendation statement. Genetic risk assessment and BRCA mutation testing for breast and ovarian cancer susceptibility: recommendation statement. Annals of internal medicine. 2005;143(5):355-61.

Acknowledgements: We would like to thank Halle R. Amick (University of North Carolina, Cecil G. Sheps Center for Health Services Research) and Crystal M. Riley (Duke-NUS Graduate Medical School Singapore) for their assistance with preparation of this manuscript, insightful editing, and outstanding attention to detail. We deeply appreciate the considerable support, commitment, and contributions of Stephanie Chang, MD, MPH, the AHRQ Task Order Officer for this project and the Evidence-based Practice Center Director.

Funding: Funded by the Agency for Health Care Research and Quality (AHRQ) under the Effective Health Care Program.

Conflict of interest: None of the authors has any affiliations or financial involvement that conflicts with the information presented in this chapter.

Corresponding author: Daniel E. Jonas, M.D., M.P.H., 5034 Old Clinic Building, CB #7110, Chapel Hill, NC 27599. Phone (919) 966-7102; Fax (919) 966-2274; email daniel_jonas@med.unc.edu

Suggested citation: Jonas DE, Wilt TJ, Taylor BC, Wilkins TM, Matchar DB. Challenges in and principles for conducting systematic reviews of genetic tests used as predictive indicators. AHRQ Publication No. 12-EHC083-EF. Chapter 11 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.

Chapter 12

Systematic Review of Prognostic Tests

Thomas S. Rector, Ph.D.; Brent C. Taylor, Ph.D.; Timothy J. Wilt, M.D., M.P.H.
Minneapolis Veterans Affairs Health Care System;
School of Medicine, University of Minnesota,
Minneapolis, MN

Abstract

A number of new biological markers are being studied as predictors of disease or adverse medical events among those who already have a disease. Systematic reviews of this growing literature can help determine whether the available evidence supports use of a new biomarker as a prognostic test that can more accurately place patients into different prognostic groups to improve treatment decisions and the accuracy of outcome predictions. Exemplary reviews of prognostic tests are not widely available, and the methods used to review diagnostic tests do not necessarily address the most important questions about prognostic tests that are used to predict the time-dependent likelihood of future patient outcomes. We provide suggestions for those interested in conducting systematic reviews of a prognostic test.

The proposed use of a prognostic test should serve as the framework for a systematic review and help define the key questions. The outcome probabilities or level of risk and other characteristics of prognostic groups are the most salient statistics for review and perhaps meta-analysis. Reclassification tables can help determine how a prognostic test affects the classification of patients into different prognostic groups, and hence helps determine their treatment. However, review of studies of the association between a potential prognostic test and patient outcomes would have little impact other than to determine whether further development as a prognostic test might be warranted.

Introduction

With increasing frequency, multiple objective measures of normal or pathologic biological processes as well as measures of social, psychological, behavioral, and demographic features are being associated with important patient outcomes. Some of these measures, singly or in combination as a prediction model, can be clinically useful. The plethora of potential new prognostic tests and prediction models, like treatments and diagnostic tests, are appropriate topics for systematic reviews. Such reviews can serve to summarize available evidence, as well as guide further research regarding the usefulness of the tests. The questions that are most salient for clinical practice, and hence a systematic review, concern the accuracy of predictions derived from a test or prediction model, and how the results affect patient management and outcomes.

This paper is meant to complement the Evidence-based Practice Centers' *Methods Guide for Comparative Effectiveness Reviews*, and is not a comprehensive or detailed review of methods that could be used to conduct a systematic review of a prognostic test. Generally speaking, the steps for reviewing evidence for prognostic tests are similar to those used in the review of a diagnostic test and are discussed in other papers in this *Medical Test Methods Guide.* These steps include: (1) using the population, intervention, comparator, outcomes, timing and setting (PICOTS) typology and an analytic framework to develop the topic and focus the review on the most important key questions; (2) conducting a thorough literature search; (3) assessing the quality of reported studies; (4) extracting and summarizing various types of statistics from clinical trials and observational studies; and (5) meta-analyzing study results. However, important differences between diagnostic and prognostic tests that should be considered when planning and conducting a review are highlighted here.

Step 1: Developing the Review Topic and Framework

Developing the review topic, which includes the framework for thinking about the relationship between the test and patient outcomes and the key questions, can be fundamentally different for diagnostic and prognostic tests. A diagnostic test is used to help determine whether a patient has a disease at the time the test is performed. Evaluations of diagnostic tests often use a categorical reference test (gold standard) to determine the true presence or absence of the disease. Typically patients are classified as diagnostic test positive or negative to estimate the test's accuracy as sensitivity (true postive fraction) and specificity (true negative fraction). In contrast, a prognostic test is used to predict a patient's likelihood of developing a disease or experiencing a medical event. Therefore, the "reference test" for a prognostic test is the observed proportion of the population who develop what is being predicted.

For practical purposes, it is often useful to group the results of a prognostic test into parsimonious categories corresponding to the implications for decision making. For example, if the actions that might follow a prognostic test are no further evaluation or treatment of "low" risk cases, initiation of treatment or prevention in "high" risk cases, or further tests or monitoring for "intermediate" risk cases, then it would be useful to structure the review according to these prognostic test categories (low, intermediate, and high risk) and clearly define each group including its outcome probabilities. If a decision model is used as the framework for a systematic review and meta-analysis of a prognostic test, the precision and accuracy of estimates of outcome probabilites within these different prognostic groups may be the primary focus. These considerations, among others, are summarized in Table 12–1, which provides a general PICOTS framework for systematically reviewing prognostic tests.

In some contexts, it may be informative to categorize subjects as those who did or did not experience the predicted outcome within a specified time interval, and then look back to categorize the results of the prognostic test. Much as for a diagnostic test, a systematic review of a prognostic test could then assess the accuracy of the prognostic test by calculating the sensitivity and specificity and predictive values for that point in time. An essential factor to consider in a review is what followup times are especially informative to patients, clinicians, or policymakers.

Table 12–1. General PICOTS typology for review of prognostic tests

Population	Clinical spectrum and other characteristics of the prognostic groups, including the observed probabilities of the outcome being predicted.
Intervention	The prognostic test or assessment including all components, exactly what it measures, how it is done, how clinical specimens are obtained, processed, and stored for testing, exactly what is being predicted, and how the test results are to be interpreted and used by test operators.
Comparator	Standard prognostic tests or assessments for predicting the same outcome.
Outcomes	Time-dependent probabilities (time-to-event curves) of what is being predicted, changes or differences in predicted outcome probabilities or reclassification of patients into different prognostic groups, changes in patient care, the net effect of using the prognostic test on patient outcomes, and cost effectiveness.
Timing	At what stage in the natural history of outcome development is the prognostic test to be used? How much follow-up time does the prognostic test cover? The percentage of patients who experience the outcome usually increases with time thereby changing the performance charactersitics of prognostic tests.
Setting	Who will use prognostic test? How? What is the applicable testing scenario?

A somewhat unique category of prognostic tests are those that can be used to predict beneficial or adverse responses to a treatment. These are commonly known as *predictive tests*. Evidence about the value of a predictive test typically is presented as separate estimates of the treatment effect in subgroups defined by the predictve test, along with a statistical test for interaction between the treatment groups and subgroups defined by a predictive test. Systematic reviews of predictive test/treatment interactions are not specifically discussed in this paper. Interested readers are referred to publications on this topic.[1]

Step 2: Searching for Studies

When developing the literature search strategy, it is important to recognize that studies can relate to one or more of the following categories:[2]

1. Proof of concept: Is the test result associated with a clinically important outcome?
2. Prospective clinical validation: How accurately does the test predict outcomes in different cohorts of patients, clinical practices, and prognostic groups?
3. Incremental predictive value: How much does the new prognostic test change predicted probabilities and increase our ability to discriminate between patients who did or did not experience the outcome of interest within a specific time period?
4. Clinical utility: Does the new prognostic assessment change predicted probabilities enough to reclassify many patients into different prognostic groups that would be managed differently?
5. Clinical outcomes: Would use of the prognostic test improve patient outcomes?
6. Cost effectiveness: Do the improvements in patient outcomes justify the additional costs of testing and subsequent medical care?

Each phase of development is focused on different types of questions, research designs, and statistical methods; however, a single study might address several of these questions. Large cohort studies and secondary analyses of clinical trials may contain the most readily available evidence to answer the first four types of questions. For the latter two types of questions, randomized controlled trials of prognostic tests are preferred. However, they can be costly and time consuming, and thus are rarely done by stakeholders.[3] Before embarking on a review focused on the last two types of key questions, reviewers need to think about what they would do, if anything, in the absence of randomized controlled studies of the effect of a prognostic test

on patient outcomes. One option is to use a decision model to frame the review and to focus on providing the best estimates of outcome probabilities.

Reliable and validated methods to exhaustively search the literature for information about prognostic tests have not been established, and the best bibliographic indexes and search strategies have yet to be determined. Some search strategies have been based on variations of key words in titles or abstracts and index terms that appear in publications meeting the study selection criteria.[4] Others have used search terms such as "cohort," "incidence," "mortality," "followup studies," "course," or the word roots "prognos-" and "predict-" to identify relevant studies.[5] Obviously, the range of terms used to describe the prognostic test(s) and the clinical condition or medical event to be predicted should be used as well. The "find similar" or "related article" functions available in some indexes may be helpful. A manual search of reference lists will need to be done. If a prognostic test has been submitted for review by regulatory agencies such as the Food and Drug Administration, the records that are available for public review should be searched. The Web site of the test producer could provide useful information too.

In contrast to diagnostic tests, many prognostic tests are incorporated into multivariable regression models or algorithms for prediction. Many reports in the literature only provide support for an independent association of a particular variable with the patient outcome to suggest the variable might be useful as a prognostic test.[6–7] The converse finding—that a test variable did not add significantly to a multivariable regression model—is difficult to retrieve, particularly via an electronic search or via reviews of abstracts, where the focus is often on positive findings.[8] Given the potential bias introduced by failing to uncover evidence of lack of a strong association, hence predictive value, if a review is going to focus on proof-of-concept questions, all studies that included the test variable should be sought out, reviewed, and discussed even when the study merely mentions that the outcome was not independently related to the potential prognostic test or a component of a multivariable prediction model.[9]

Whenever a systematic review focuses on key questions about prognostic groups that are defined by predicted outcome probabilities, reviewers should search for decision analyses, guidelines, or expert opinions that help support the outcome probability thresholds used to define clinically meaningful prognostic groups, that is, groups that would be treated differently in practice because of their predicted outcome. Ideally, randomized controlled clinical trials of medical interventions in patients selected based on the prognostic test would help establish the rationale for using the prognostic test to classify patients into the prognostic groups—although this is not always sufficient to evaluate this use of a prognostic test.[1,3]

Step 3: Selecting Studies and Assessing Quality

Previous reviews of prognostic indicators have demonstrated substantial variation in study design, subject inclusion criteria, methods of measuring key variables, followup time, methods of analysis (including definition of prognostic groups), adjustments for covariates, and presentation of results.[10–12] Some of these difficulties could be overcome if reviewers were given access to the individual patient-level data from studies, which would allow them to conduct their own analyses in a more uniform manner. Lacking such data, several suggestions have been made for assessing studies to make judgments about the quality of reports and whether to include or exclude them from a review.[5,13,14] Table 12–2 lists questions that should be considered. At this time, reviewers will need to decide which of these general criteria or others are appropriate for judging studies for their particular review. As always, reviewers should be explicit about any criteria that were

used to exclude or include studies from a review. Validated methods to use criteria to score the quality of studies of prognostic tests need to be developed.

Table 12–2. Outline of questions for judging the quality of individual studies of prognostic tests

1. Was the study designed to evaluate the new prognostic test, or was it a secondary analysis of data collected for other purposes?
2. Were the subjects somehow referred or selected for testing? What was the testing scenario?
3. Was the clinical population clearly described including the sampling plan, inclusion and exclusion criteria, subject participation, and the spectrum of test results? Did the sample represent patients that would be tested in clinical practice?
4. Did everyone in the samples have a common starting point for followup with respect to the outcome of interest including any treatments that could affect the outcome being predicted?
5. Were the prognostic tests clearly described and conducted using a standardized, reliable, and valid method?
   a. Was the test used and interpreted the same way by all sites/studies including any interdeterminate test results?
   b. Were the test results ascertained without knowledge of the outcome?
   c. Were investigators blinded to the test results?
   d. How were previously established prognostic indicators or other prognostic assessments included in the study and analyses?
6. Was the outcome being predicted clearly defined and ascertained using a standardized, reliable, and valid method?
   a. How complete was the followup of subjects, and were losses to followup related to the test results or the outcome being predicted?
   b. Was the duration of followup adequate?
7. Were the data used to develop the prognostic test?
   a. Were the prognostic groups pre-defined based on clinically meaningful decision thresholds for predicted outcome probabilities?
   b. Were the results externally validated using an independent sample or internally validated via boot strap or cross-validation methods?
   c. Were any previously established prognostic indicators or prediction models being used as comparators fit to the sample data in the same manner as the potential new prognostic test?
   d. Were outcome predictions adjusted for any other factors? Which ones? How?

Comparisons of prognostic tests should use data from the same cohort of subjects to minimize confounding the comparison. Within a study, the prognostic tests being compared should be conducted at the same time to ensure a common starting point with respect to the patient outcome being predicted. Reviewers should also note the starting point of each study reviewed. All of the prognostic test results and interpretations should be ascertained without knowledge of the outcome to avoid ascertainment bias. Investigators should be blinded to the results of the prognostic test to avoid selective changes in treatment that could affect the outcome being predicted. Reviewers need to be aware of any previously established prognostic indicators that should be included in a comparative analysis of potential new prognostic tests, and need to pay close attention to that with which a new prognostic test is compared. Any adjustments for covariates that could make studies more or less comparable also need to be noted.[15]

If the investigators fit a new prognostic test or prediction equation to the sample data (test development sample) by using the data to define cutoff levels or model its relationships to the outcome and estimate regression coefficient(s), the estimated predictive performance can be overly optimistic. Fitting the new test to the data might bias a comparison to an established prognostic method that was not fit to the same sample.

Step 4: Extracting Statistics to Evaluate Test Performance

The summary statistics reported in the selected articles need to be appropriate for the key question(s) the review is trying to address. For example, investigators commonly report estimated hazard ratios from Cox regression analyses or odds ratios from logistic regression analyses to test for associations between a potential prognostic test and the patient outcome. These measures of association address only early phases in the development of a potential prognostic test—proof of concept and perhaps validation of a potentially predictive relationship to an outcome in different patient cohorts, and to a very limited extent the potential to provide incremental predictive value. Potential predictors that exhibit statistically significant associations with an outcome often do not substantially discriminate between subjects who eventually do or do not experience the outcome event, because the distributions of the test result in the two outcome groups often overlap substantially even when the means are highly significantly different.[16,17] Statistically significant associations (hazard, relative risk, or odds ratios) merely indicate that more definitive evaluation of a new predictor is warranted.[18–19] Nevertheless, for reviewers who are interested in these associations, there are well established methods for summarizing estimates of hazard, relative risk, or odds ratios.[20–23] However, the questions a systematic review could answer about the use of a prognostic test by summarizing its association with an outcome are quite limited and not likely to impact practice. More relevant are the estimates of absolute risk in different groups defined by the prognostic test.

Discrimination Statistics

The predictive performance of prognostic tests is often reported in a manner similar to diagnostic tests, using estimates of sensitivity, specificity, and the area under the receiver operating characteristic (ROC) curve at one particular followup time. These indices of discrimination can be calculated retrospectively and compared when a new prognostic indicator is added to a predictive model, or a prognostic test is compared to predictions made by other methods, including the judgments of experienced clinicians.[24–27] However, these backward-looking measures of discrimination do not summarize the predicted outcome probabilities and do not directly address questions about the predictions based on a new prognostic test or its impact on patient outcomes.[28–30] The next section on reclassification tables describes other measures of test discrimination that can help reviewers assess, in part, the clinical impact of prognostic tests.

If reviewers elect to use the more familiar and often reported discrimination statistics, then they must be cognizant of the fact that they change over time as more patients develop the outcome being predicted. Time-dependent measures of sensitivity, specificity, and the ROC curve have been developed.[31] Harrell's C-statistic is conceptually similar to an area under an ROC curve and can be derived from time-to-event analyses.[32–33] Examples of systematic reviews and meta-analyses of prognostic tests that used these time-dependent measures of discrimination were not found.

Reclassification Tables

The clinical usefulness of a prognostic test depends largely on its ability to sort patients into different prognostic groups and provide accurate predictions about their future health. For example, expert guidelines use prognostic groups defined by the estimated 10-year risk of developing cardiovascular disease (<10%, 10 to 20% and >20%) based on the Framingham cardiovascular risk score to help determine whether to recommend interventions to prevent

future cardiovascular events.[34] Analyses of reclassification tables are now being reported to determine how adding a prognostic test reclassifies patients into the prognostic groups.[35–38] Table 12–3 shows a hypothetical example of a reclassification table. Ideally, the classification of outcome probabilities into prognostic groups (arbitrarily set at an individual predicted probability >0.10 in the example) should be based on outcome probabilities that will generally lead to different courses of action. If not, the reviewer needs to take note, because the observed reclassifications could be clinically meaningless in the sense that they might not be of sufficient magnitude to alter the course of action; that is to say, some reclassification of patients by a prognostic test might not make any difference in patient care. In the example, adding the new prognostic test reclassified 10 percent of the 1,000 people originally in the lower risk group and 25 percent of the 400 people in the higher risk group.

Table 12–3. Example of a reclassification table based on predicted outcome probabilities

Grouped Mortality Probabilities Estimated by the First Prognostic Test	Grouped Mortality Probabilities Estimated by the First Prognostic Test + a New Prognostic Test		
	0 to 0.10	> 0.10	Total
0 to 0.10			
Patients in prognostic group	900	100 (10%)	1000
Mortality predictions using 1st test	4.0%	8.0%	4.40%
Mortality prediction using both tests	3.8%	11.0%	-
Observed mortality	3.9%	12.0%	4.7%
> 0.10			
Patients in prognostic group	100 (25%)	300	400
Mortality predictions using 1st test	15.0%	17.0%	16.5%
Mortality prediction using both tests	9.0%	19.0%	-
Observed mortality	10.0%	19.0%	16.8%
Total			
Patients in prognostic group	1000	400	1400
Mortality prediction using both tests	4.3%	17.0%	-
Observed mortality	4.5%	17.2%	8.2%

Reclassification tables typically provide information about the observed outcome probabilities in each prognostic group (summarized as percentages in the example) and the predicted probabilites. However, this information is often limited to a single followup time, and the precision of the estimates might not be reported. The differences between the estimated probabilities and observed outcomes for each prognostic group might be analyzed by a chi-square goodness-of-fit test.[39] However, these results will not help the reviewer determine if the differences in predicted and observed probabilities are substantially better when the new prognostic test is added. In the example depicted in Table 12–3, the differences between predicted and observed values for each prognostic test shown in the column and row totals are small, as expected whenever prognostic groups have a narrow range of individual predicted probabilities and the prediction models are fit to the data rather than applied to a new sample.

Reviewers might also encounter articles that report separate reclassification tables for patients who did or did not experience the outcome event within a specific period of time, along with a summary statistic known as the net reclassification improvement (NRI).[40] In the group that developed the outcome event within the specified period of time, the net improvement is the proportion of patients who were reclassified by a prognostic test into a higher probability subgroup minus the proportion who were reclassified into a lower probability subgroup. In a two-by-two reclassification table of only subjects who experienced the outcome event (e.g., those who died), this net difference is the estimated change in test sensitivity. In the group who

did not experience the outcome event, the net improvement is the proportion of patients who were reclassified into a lower probability subgroup minus the proportion who were reclassified into a higher probability subgroup. In a two-by-two reclassification table of only subjects who did not experience the event within the followup period (e.g., those who survived), this net difference is the estimated change in specificity. The NRI is the simple sum of net improvement in classifcation of patients that did or did not experience the outcome.

If these calculations use the mean changes in individual predicted probabilities in the patients that did or did not experience the outcome, the result is known as the integrated discrimination index (IDI). Another formulation of the NRI calculates the probabilities of the predicted event among those that have an increase in their predicted probability given the results of a new prognostic test, the probabilities of the predicted event among those that have a decrease in their predicted probability, and the event probability in the overall sample.[41] These three probabilities can be estimated by time-to-event analysis but still only represent a single point of followup. This so-called continuous formulation of the NRI doesn't require one to define clinically meaningful prognostic categories. Rather, it focuses on subjects that have, to any degree, a higher or lower predicted outcome probability when a new prognostic test is employed. Not all increases or decreases in predicted probabilities would be clinically meaningful in the sense that they would prompt a change in patient management.

Estimates of the NRI or IDI from different studies could be gleaned from the literature comparing prognostic tests. Several issues need to be examined before trying to pool estimates from different studies. Reviewers should make sure the charactersitcs of prognostic groups, definition of the outcome event, overall probability of the event, and the followup time did not vary substantially between studies.

Predictive Values

Treatment decisions based on outcome probabilities are often dichotomous—for example, "treat those at high risk" and "don't treat those at low risk." If patients are treated because a prognostic test indicates they are "high risk," then the observed time-dependent percentages of patients developing the outcome without treatment are essentially positive predictive values (i.e., the proportion of those with a "positive" prognostic test that have the event). If clinicians do not treat patients in the lower risk group, then one minus the observed time-dependent outcome probabilities are the negative predictive values (i.e., the proportion of those with a "negative" prognostic test that don't have the event). For a single point of followup, these positive and negative predictive values can be compared using methods devised for comparing predictive values of diagnostic tests. Most likely the ratios of positive and negative predictive values of two prognostic tests will be summarized in a report, along with a confidence interval.[42] The regression model proposed by Leisenring and colleagues might be used to determine how patient characteristics relate to the relative predictive values.[43] Methods of comparing predictive values of two prognostic tests that are in the form of time-to-event curves are available if such displays of data are encountered during a review.[44–47]

Step 5: Meta-Analysis of Estimates of Outcome Probabilities

The most definitive level of evidence to answer the most important questions about a prognostic test or comparison of prognostic tests would come from randomized controlled trials designed to demonstrate a net improvement in patient outcomes and cost-effectiveness. Many studies of prognostic tests do not provide this ultimate evidence. However, a systematic review

could provide estimates of outcome probabilities for decision models.[48] Estimates could come from either randomized controlled trials or observational studies as long as the prognostic groups they represent are well characterized and similar. A meta-analysis could provide more precise estimates of outcome probabilities. In addtion, meta-analysis of estimated outcome probabilities in a prognostic group extracted from several studies may provide some insights into the stability of the estimates and whether variation in the estimates is related to characteristics of the prognostic groups.

Methods have been developed to combine estimates of outcome probabilities from different studies.[20] Dear's method uses a fixed effects regression model while Arend's method is similar to a DerSimonian-Laird random-effects model when there is only one common followup time for all studies/prognostic groups in the analysis.[49,50] These references should be consulted for guidance on this type of meta-analysis.

Conclusion

There is a large and rapidly growing literature about prognostic tests. A systematic review can determine what is known and what needs to be determined to support use of a prognostic test by decision makers. Hopefully, this guidance will be helpful to reviewers who want to conduct an informative review of a prognostic test, and will spur efforts to establish consensus methods for reporting studies of prognostic tests and conducting reviews of them.

Key Points

- Methods for the conduct of a clincially oriented systematic review of a prognostic test are not well established. Several issues discussed herein will need to be addressed when planning and conducting a review.
- The intended use of the prognostic test under review needs to be specified, and predicted probabilities need to be classified into clinically meaningful prognostic groups; i.e., those that would entail different treatment of patients. The resultant prognostic groups need to be described in detail, including their outcome probabilities.
- A large number of published reports focus on the associations between prognostic indicators and patient outcomes, the first stage of development of prognostic tests. A review of these types of studies would have limited clinical value.
- Criteria to evaluate and score the quality of studies of prognostic tests have not been firmly established. Reviewers can adapt criteria that have been developed for judging studies of diagnostic tests and cohort studies with some modifcations for differences inherent in studies of prognostic tests. Suggestions are listed in Table 12–2.
- Given the fundamental difference between diagnostic tests that determine the current state of disease and prognostic tests that predict a future state of disease, some of the most commonly used statistics for evaluating diagnostic tests, such as point estimates of test sensitivity and specificity and receiver operator characteristic curves, are not as informative for prognostic tests. The most pertinent summary statistics for prognostic tests are: (1) the time-dependent observed outcome probabilities within clearly defined prognostic groups, (2) the closeness of each group's predicted probabilities to the observed outcomes, and (3) the way the use of a new prognostic test reclassifies patients into different prognostic groups and improves predictve accuracy and overall patient outcomes.

- Methods to compare and summarize the predictive performance of prognostic tests need further development and widespread use to facilitate systematic reviews.

References

1. Janes H, Pepe MS, Bossuyt PM, Barlow WE. Measuring the performance of markers for guiding treatment decisions. Ann Intern Med 2011;154:253-259.
2. Hlatky MA, Greenland P, Arnett DK, et al. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. Circulation 2009;119(17):2408-16.
3. Wang TJ. Assessing the role of circulating, genetic and imaging biomarkers in cardiovascular risk prediction. Circulation 2011;123:551-565.
4. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. J Am Med Inform Assoc 2001;8(4):391-7.
5. Hayden JA, Cote P, Bombardier C. Evaluation of the quality of prognosis studies in systematic reviews. Ann Intern Med 2006;144(6):427-37.
6. McShane LM, Altman DG, Sauerbrei W, et al. Reporting recommendations for tumor markerprognostic studies (REMARK). J Natl Cancer Inst 2005;97(16):1180-4.
7. Riley RD, Abrams KR, Sutton AJ, et al. Reporting of prognostic markers: current problems and development of guidelines for evidence-based practice in the future. Br J Cancer 2003;88(8):1191-8.
8. Kyzas PA, Denaxa-Kyza D, Ioannidis JP. Almost all articles on cancer prognostic markers report statistically significant results. Eur J Cancer 2007;43(17):2559-79.
9. Kyzas PA, Loizou KT, Ioannidis JP. Selective reporting biases in cancer prognostic factor studies. J Natl Cancer Inst 2005;97(14):1043-55.
10. Altman DG. Systematic reviews of evaluations of prognostic variables. BMJ 2001;323(7306):224-8.
11. Hall PA, Going JJ. Predicting the future: a critical appraisal of cancer prognosis studies. Histopathology 1999;35(6):489-94.
12. Altman DG, Riley RD. Primer: an evidence-based approach to prognostic markers. Nat Clin Pract Oncol 2005;2(9):466-72.
13. Speight PM. Assessing the methodological quality of prognostic studies. Chapter 3 (p. 7-13) in: Speight, Palmer, Moles, et al. The cost-effectiveness of screening for oral cancer in primary care. Health Technol Assess 2006;10(14):1-144, iii-iv.
14. Pepe MS, Feng Z, Janes H, et al. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. J Natl Cancer Inst 2008;100(20):1432-8.
15. Janes H, Pepe MS. Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: an old concept in a new setting. Am J Epidemiol 2008;168(1):89-97.
16. Ware JH. The limitations of risk factors as prognostic tools. N Engl J Med 2006;355(25):2615-7.
17. Pepe MS, Janes H, Longton G, et al. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol 2004;159(9):882-90.
18. Feng Z, Prentice R, Srivastava S. Research issues and strategies for genomic and proteomic biomarker discovery and validation: a statistical perspective. Pharmacogenomics 2004;5(6):709-19.
19. Riesterer O, Milas L, Ang KK. Use of molecular biomarkers for predicting the response to radiotherapy with or without chemotherapy. J Clin Oncol 2007;25(26):4075-83.
20. Parmar MK, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. Stat Med 1998;17(24):2815-34.
21. The Fibrinogen Studies Collaboration. Measures to assess the prognostic ability of the stratified Cox proportional hazards model. Stat Med 2009;28(3):389-411.

22. Earle CC, Pham B, Wells GA. An assessment of methods to combine published survival curves. Med Decis Making 2000;20(1):104-11.

23. Coplen SE, Antman EM, Berlin JA, et al. Efficacy and safety of quinidine therapy for maintenance of sinus rhythm after cardioversion. A meta-analysis of randomized control trials. Circulation 1990;82(4):1106-16.

24. Sinuff T, Adhikari NK, Cook DJ, et al. Mortality predictions in the intensive care unit: comparing physicians with scoring systems. Crit Care Med 2006;34(3):878-85.

25. Groenveld HF, Januzzi JL, Damman K, et al. Anemia and mortality in heart failure patients: a systematic review and meta-analysis. J Am Coll Cardiol 2008;52(10):818-27.

26. Mackillop WJ, Quirt CF. Measuring the accuracy of prognostic judgments in oncology. J Clin Epidemiol 1997;50(1):21-29.

27. Ingelsson E, Schaefer EJ, Contois JH, et al. Clinical utility of different lipid measures for prediction of coronary heart disease in men and women. JAMA 2007;298(7):776-85.

28. Poses RM, Cebul RD, Collins M, et al. The importance of disease prevalence in transporting clinical prediction rules. The case of streptococcal pharyngitis. Ann Intern Med 1986;105(4):586-91.

29. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007;115(7):928-35.

30. Pepe MS, Janes HE. Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. J Natl Cancer Inst 2008;100(14):978-9.

31. Pepe MS. The statistical evaluation of medical tests for classification and prediction. Oxford, UK: Oxford University Press; 2003. Section 9.2, Incorporating the time dimension; p. 259-67.

32. Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Stat Med 2004;23(13):2109-23.

33. Pepe MS, Zheng Y, Jin Y, et al. Evaluating the ROC performance of markers for future events. Lifetime Data Anal 2008;14(1):86-113.

34. Grundy SM, Cleeman JI, Merz CN, et al. Implications of recent clinical trials for the National Cholesterol Education Program Adult Treatment Panel III guidelines. Circulation 2004;110(2):227-39.

35. Cook NR, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. Ann Intern Med 2009;150(11):795-802.

36. Janes H, Pepe MS, Gu W. Assessing the value of risk predictions by using risk stratification tables. Ann Intern Med 2008;149(10):751-60.

37. Ankle Brachial Index Collaboration, Fowkes FG, Murray GD, et al. Ankle brachial index combined with Framingham Risk Score to predict cardiovascular events and mortality: a meta-analysis. JAMA 2008;300(2):197-208.

38. Meigs JB, Shrader P, Sullivan LM, et al. Genotype score in addition to common risk factors for prediction of type 2 diabetes. N Engl J Med 2008;359(21):2208-19.

39. Pigeon JG, Heyse JF. An improved goodness of fit statistic for probability prediction models. Biom J 1999;41(1):71-82.

40. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008;27(2):157-72; discussion 207-12.

41. Pencina MJ, D'Agostino RB, Sr, Steyerberg EW. Extensions of net reclassifcation improvement calculations to measure usefulness of new biomarkers. Stat Med 2011;30:11-21.

42. Moskowitz CS, Pepe MS. Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs. Clin Trials 2006;3(3):272-9.

43. Leisenring W, Alonzo T, Pepe MS. Comparisons of predictive values of binary medical diagnostic tests for paired designs. Biometrics 2000;56(2):345-51.

44. Graf E, Schmoor C, Sauerbrei W, et al. Assessment and comparison of prognostic classification schemes for survival data. Stat Med 1999;18(17-18):2529-45.

45. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. Stat Med 2004;23(5):723-48.

46. Huang Y, Sullivan Pepe M, Feng Z. Evaluating the predictiveness of a continuous marker. Biometrics 2007;63(4):1181-8.

47. Pepe MS, Feng Z, Huang Y, et al. Integrating the predictiveness of a marker with its performance as a classifier. Am J Epidemiol 2008;167(3):362-8.

48. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26(6):565-74.

49. Dear KB. Iterative generalized least squares for meta-analysis of survival data at multiple times. Biometrics 1994;50(4):989-1002.

50. Arends LR, Hunink MG, Stijnen T. Meta-analysis of summary survival curve data. Stat Med 2008;27(22):4381-96.

Funding: Funded by the Agency for Health Care Research and Quality (AHRQ) under the Effective Health Care Program.

Conflict of interest: None of the authors has any affiliations or financial involvement that conflicts with the information presented in this chapter.

Corresponding author: Thomas S. Rector, Ph.D., Center for Chronic Disease Outcomes Research, Minneapolis VA Medical Center, 152/2E, One Veterans Drive, Minneapolis, MN 55417. Phone: 612-467-2114; Fax: 612-727-5699; Email: thomas.rector@va.gov

Suggested citation: Rector TS, Taylor BC, Wilt TJ. Systematic review of prognostic tests. AHRQ Publication No. 12-EHC084-EF. Chapter 12 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published in a special supplement to the Journal of General Internal Medicine, July 2012.

Appendix: Test Performance Metrics

The basic definitions of test performance are based upon a 2×2 cross tabulation of the "true disease status" and the test results (Appendix Table). Ostensibly simple, the information in a 2×2 table can be summarized in several ways—some of which are mathematically equivalent (e.g., sensitivity/specificity vs. positive/negative likelihood ratios)—and these measures and their application can be confusing in practice.[1] The basic measures are described briefly below; for a discussion on their relative merits and drawbacks, see Tatsioni et al.[2]

Appendix Table. 2×2 table used in the calculation of test performance measures

		True Disease Status	
		Disease*	**No Disease***
(Index) Medical Test	Suggestive of disease ("positive")	"TP" (="true positives")	"FP" (="false positives")
	Not suggestive of disease ("negative")	"FN" (="false negatives")	"TN" (= "true negatives")

"TP" = true positive; "FN" = false negative; "FP" = false positive; "TN" = true negative. The quotation marks are retained to stress that, in the calculation of the basic measures reviewed here, we assume that the reference standard test has negligible misclassification rates for practical purposes.

- Sensitivity: "TP"/("TP" + "FN")
- Specificity: "TN"/("FP" + "TN")
- Positive likelihood ratio (LR+): Sensitivity/(1–Specificity)
- Negative likelihood ratio (LR-): (1–Sensitivity)/Specificity
- Diagnostic odds ratio: ("TP" * "TN") / ("FP" * "FN")
- Positive predictive value: "TP"/("TP" + "FP") = (Sensitivity * Prevalence)/(Sensitivity * Prevalence + (1-Prevalence)*(1-Specificity))
- Negative predictive value: "TN"/("TN" + "FN") = (Specificity * (1-Prevalence))/(Specificity *(1- Prevalence) + Prevalence*(1-Sensitivity))

*This is typically ascertained by a reference test. The reference test is assumed to have negligible misclassification of the true disease status.

Sensitivity and Specificity

Sensitivity, also known as the true positive rate, is the probability of testing positive for diseased patients. It expresses the ability of a medical test to maximize true positives. Specificity, or true negative rate, is the probability of testing negative for non-diseased patients. It expresses the ability of a test to minimize false positives.

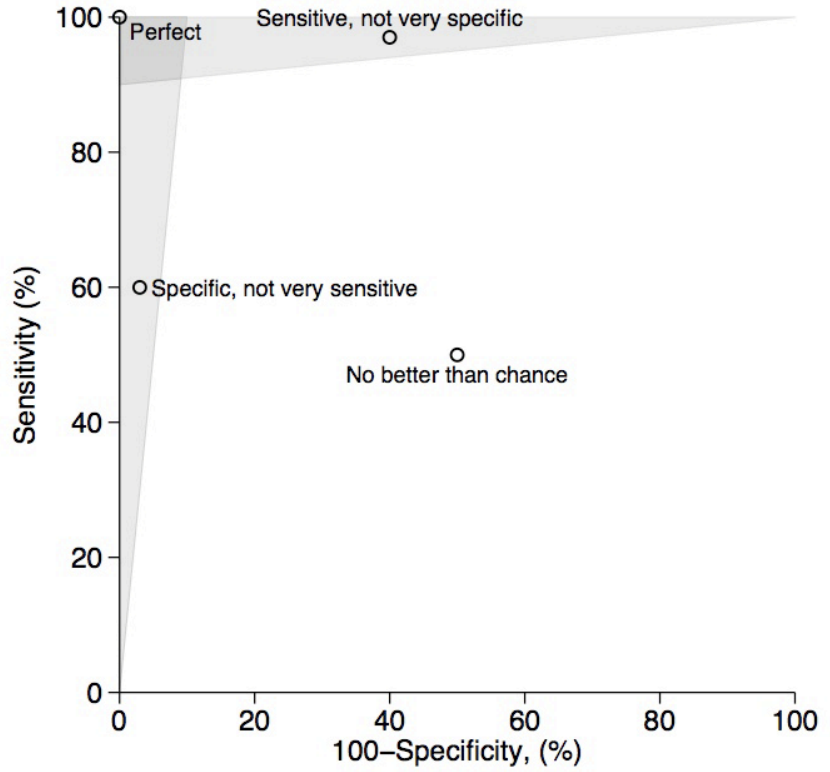
The two measures have clear clinical interpretations, but are not as useful clinically as the positive and negative predictive values (see below). Sensitivity and specificity are negatively correlated with each other with respect to diagnostic thresholds. If the threshold (cutpoint) for test positive is set higher—say when the test provides a continuously valued result—sensitivity will decrease while specificity will increase. On the other hand, if the threshold is lower, we will see an increase in sensitivity and a corresponding decrease in specificity. This non-independence of sensitivity and specificity across explicit or implicit diagnostic thresholds poses challenges for quantitative synthesis.

When several thresholds have been considered for a single set of data, a receiver operating characteristic (ROC) curve could be obtained by plotting sensitivity vs. 1–specificity. (Appendix Figure). The ROC curve depicts the observed patterns of sensitivity and specificity at different

thresholds, as well as the negative correlation between the two measures. As we will discuss below, one way to summarize diagnostic accuracy data is to calculate a summary ROC curve.

Note that the closer a study point is to the upper left corner of the plot, the better its diagnostic ability.

Appendix Figure. Typical plot of sensitivity versus 100 percent specificity



Four hypothetical studies are depicted in the square sensitivity/100 percent–specificity plot. The closer a study is to the upper-left corner of the plot, the better its diagnostic ability. Studies lying on the major diagonal of the plot have no diagnostic ability (no better than chance). Studies lying on the left shaded area have positive likelihood ratio (LR+) of 10 or more. Studies lying on the top shaded area have negative likelihood ratio (LR-) of 0.1 or less. Studies lying on the intersection of the grey areas (darker grey polygon) have both LR+>10 and LR-<0.1. Screening tests typically operate in the less shaded areas, whereas confirmatory tests used to rule out a diagnosis often operate near or in the top shaded area. The systematic reviewer must be familiar with the mentioned measures and their interpretation: the same medical test can be used in different settings and roles, and different measures will best capture its performance each time.

Positive and Negative Likelihood Ratios

The *positive* and *negative likelihood ratios* (LR+ and LR-, respectively) quantify the change in the certainty of the "diagnosis" conferred by test results. More specifically, the likelihood ratios transform the *pretest odds* to the *posttest odds* of a given (positive or negative) diagnosis:

$$posttest\ odds = pretest\ odds \times LR$$

For a positive result with the medical test, the positive likelihood ratio would be used in the above relationship; for a negative result with the medical test portable monitor, the negative likelihood ratio would be used.

If a given medical test has very good ability to predict the "true disease status," its positive likelihood ratio will be high (i.e., will greatly increase the odds of a positive diagnosis) and its negative likelihood ratio will be low (i.e., will diminish substantially the likelihood of the positive diagnosis). A completely non-informative portable monitor would have likelihood ratios equal to 1 (i.e., does not transform the pre-test odds substantially in the equation above). Typically, a positive likelihood ratio of 10 or more and a negative likelihood ratio of 0.1 or less are considered to represent informative tests.[3] We note that other, more lenient boundaries for LR+ and LR- can be used[3] and that the choice of the boundaries is a subjective decision. It is interesting to note that studies with high LR+ and low LR- can be readily identified in the square sensitivity/100 percent-specificity plot, as shown in the Appendix Figure above.

Diagnostic Odds Ratio

The diagnostic odds ratio (DOR) describes the odds of a positive test in those with disease relative to the odds of a positive test in those without disease.[4] It can be computed in terms of sensitivity and specificity as well as in terms of positive and negative likelihood ratios (DOR = LR+/LR-). Thus this single measure includes information about both sensitivity and specificity and tends to be reasonably constant despite diagnostic threshold. However, it is impossible to use diagnostic odds ratios to weigh sensitivity and specificity separately, and to distinguish between tests with high sensitivity and low specificity and tests with low sensitivity and high specificity.

Another disadvantage is that it is difficult for clinicians to understand and apply, limiting its clinical value. This is partly because they are not often exposed to diagnostic odds ratios. A diagnostic odds ratio is similar to an odds ratio that measures strength of association in an observational study or effect size in a trial. However, contrary to the typical effect size magnitudes of such odds ratios (often between 0.5 and 2), diagnostic odds ratios can attain much larger values (often greater than 100).

Positive and Negative Predictive Values

Positive predictive value is the probability of disease given a positive test and negative predictive value is the probability of no disease following a negative test. These values are highly useful for clinical purposes because they give the clinician an indication of the likelihood of disease or a specific event such as death following the results of the medical test.

Positive and negative predictive values depend upon disease prevalence, which is unlikely to be consistent among studies. Therefore, they are often calculated for a range of plausible prevalence values and tabulated or plotted in graphs.

References

1. Loong TW. Understanding sensitivity and specificity with the right side of the brain. BMJ 2003; 327(7417):716-9.
2. Tatsioni A, Zarin DA, Aronson N, Samson DJ, Flamm CR, Schmid C, et al. Challenges in systematic reviews of diagnostic technologies. Ann Intern Med 2005; 142(12 Pt 2):1048-55.
3. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. JAMA 1994; 271(9):703-7.
4. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol 2003; 56(11):1129-35.