This document (PDF) is a compilation of the following draft chapters of the:

Methods Guide for Medical Test Reviews

Source: Agency for Healthcare Research and Quality. Methods Guide for Medical Test Reviews [posted November 2010]. Rockville, MD. Available at: http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayProduct&productID=454

Methods Guide for Medical Test reviews

Paper 1

Introduction to the Methods Guide for Medical Test Reviews

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

AHRQ Publication No. xx-EHCxxx-EF
<Month Year>

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, through a collaborative effort within the Evidence-based Practice Center Program, have developed a Methods Guide for Medical Test Reviews. We intend that these documents will serve as a resource for our EPCs as well as for other investigators interested in conducting systematic reviews on medical tests.

This Medical Test Methods guide is intended to be a practical guide for those who prepare and use systematic reviews on medical tests. This document complements the EPC Methods Guide on Comparative Effectiveness Reviews (http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318), which focuses on methods to assess the effectiveness of treatments and interventions. The guidance here applies the same principles for assessing treatments to the issues and challenges in assessing medical tests and 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. We provide guidance in stepwise fashion for those conducting a systematic review.

The *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 *Medical Test Methods Guide* and the Effective Health Care Program can be made at www.effectivehealthcare.ahrq.gov.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

Paper 1. Introduction to the Methods Guide for Medical Test Reviews

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 these questions.

Systematic reviews developed under the EPC Program (sometimes labeled “evidence reports” or “health 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] 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* (hereafter referred to as the *General Methods Guide*). Not only has the present *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 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 a puzzle-like fashion.

In addition to encouraging a high standard for excellence, usefulness, and efficiency in EPC reviews, this *Medical Test Methods Guide* is designed to promote consistency in how specific issues are addressed across the various systematic reviews produced by EPC investigators. Even though consistency in approach may not always guarantee that a particular task in review development is done in an ideal fashion, 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.

[a] A MEDLINE® search using the keyword “test.mp” demonstrates a doubling of the number of citations approximately every 10 years since 1960.

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.

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. However, often context affects 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—that medical tests (and therefore assessments of those tests) are about improving patient outcomes—may seem to reflect an obvious sentiment. If so, then it is a sentiment more honored in the breach than in the observance. 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 the *sine qua non* of test value (i.e., if the performance characteristics are good, then the test should be promoted). For example, a test with sensitivity and specificity in the high 90-percent range may not improve the likelihood of a good patient outcome if the underlying condition prevalence 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 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 P-1.

Figure P-1. Causal chain diagram

Test → Result → Categorization (e.g., high risk, disease present, disease progression) → Decision → Patient outcome

In rare cases, a test will have been evaluated in a comprehensive clinical trial in which every relevant outcome was 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[b] are among the most commonly employed. 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, such tests also encompass 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 primarily because many treatments lead directly to the intended result (or to adverse effects), whereas tests may have several steps between the performance of the test and the outcome of clinical importance. Moreover, the diagnostic and management process can have myriad options and is thus more varied and more difficult to standardize than many treatment plans.[1] One consequence 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.[2]

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 (involves a process that can be reproduced), consistent (is similar across EPC reports), tractable (can be performed within resource constraints), and useful (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 through a properly designed randomized controlled trial (RCT)[3-6] that compares patient management using the test versus one or more alternative strategies. In practice, such trials are not routinely performed because they are often deemed unattainable.[4,6] In those uncommon circumstances where a medical test is evaluated in an RCT, the reader is referred to other relevant guidance documents, including AHRQ's *General Methods Guide*.[7]

Key Insights 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 key ideas emerge from this literature. The first is the recognition that a medical test used to discriminate between the

[b] Here the term "medical tests" is used as an umbrella 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).

presence or absence of a specific clinical condition can be likened to an electronic signal detector.[9-11] This has opened the way to applying signal detection theory, including the notions of sensitivity, specificity, and the application of Bayes rule, to calculate disease probabilities for positive or negative test results.[9-11]

The second insight 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. Study objectives, the terms used to convey those objectives, and relevant examples are listed provided in Table I-1.

Table I-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 liklihood ratios Positive and negative predictive value Test yield Receiver operating characteristic (ROC) curve
Ability of 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 the test to benefit society as a whole	Societal efficacy	Incremental cost-effectiveness

Analytic Frameworks

While the preceding provides a way to classify test evaluation studies according to their objective, 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 1991, Woolf described a conceptual model that he termed the "Evidence Model" for use during clinical practice guideline development.[12] He proposed the model as a means of visually clarifying the relationship between health care interventions and outcomes, and for guiding the review process. Writing as the science advisor to the United States Preventive Services Task Force (USPSTF) in 1994, Woolf described this same model as the "analytic framework,"[13] which he proposed as an approach to "keep the analytic process on track and to avoid unnecessary inefficiencies." He described the framework as a means to create the questions for the review and to determine the nature of the evidence necessary for addressing the questions.

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.[14]

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 to decisionmaking. As noted below, the current standard approach to classifying contexts of use is the PICOTS typology.[c]

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.[7] Although not specifically recommending a visual representation of the framework, the Cochrane Collaboration also organizes key questions using a similar framework.[15]

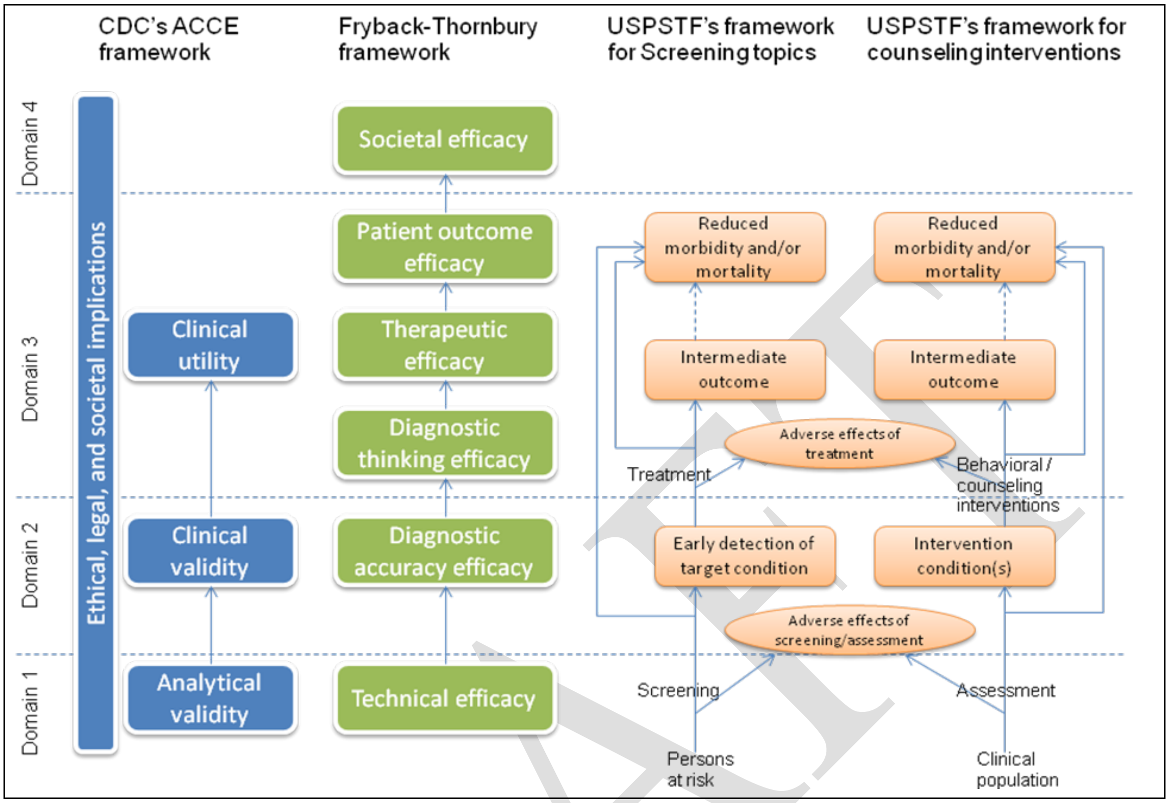
A Note on Terminology

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 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 I-1 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,[16-17] (2) the Fryback and Thornbury model, one of the most widely used and well-known of all the proposed organizing frameworks,[18] and (3) the USPSTF model for assessing screening and counseling interventions.[19]

[c] For more on the PICOTS typology, *see* Paper 2.

Figure I-1. A mapping across three major organizing frameworks for evaluating clinical tests


CDC's ACCE framework
Fryback-Thornbury framework
USPSTF's framework for Screening topics
USPSTF's framework for counseling interventions
Domain 4
Domain 3
Domain 2
Domain 1
Ethical, legal, and societal implications
Clinical utility
Clinical validity
Analytical validity
Societal efficacy
Patient outcome efficacy
Therapeutic efficacy
Diagnostic thinking efficacy
Diagnostic accuracy efficacy
Technical efficacy
Reduced morbidity and/or mortality
Intermediate outcome
Adverse effects of treatment
Treatment
Behavioral / counseling interventions
Early detection of target condition
Intervention condition(s)
Adverse effects of screening/assessment
Screening
Assessment
Persons at risk
Clinical population

Notes: 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 formalism that has proven extremely useful for the evaluation of therapies, and which also applies to the evaluation of medical tests, is the PICOTS typology. The PICOTS 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.[8]

The EPC Program, reflecting the systematic review community as a whole, occasionally uses variations of the PICOTS typology (Table I-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 on 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 I-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, desicional, 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 Guide*

As noted above, this *Medical Test Methods Guide* complements AHRQ's *General Methods Guide*),[7] 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.

Papers 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 Papers 2 and 3, which describe how to determine the key questions, as well as in Papers 5 and 6, which describe how to assess the quality and applicability of studies.

Papers 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 (Paper 4) requires unique strategies, which are discussed briefly. Assessing individual study quality (Paper 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 Paper 5, we note several approaches to assessing the limitations of a study of a medical test and recommend an approach.

Assessing applicability (Paper 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 (Paper 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 Paper 7, we discuss the challenges and applications of grading the strength of a body of test evidence.

Papers 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 (Paper 9). In reviewing the application of modeling in clinical test evidence reviews, we focus in Paper 10 on evaluating the circumstances under which a formal modeling exercise may be a particularly useful component of an evidence review.

Finally, in Papers 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 papers, those papers focus on methods for evaluating tests to determine the current presence of disease, as with screening or diagnostic tests. Papers 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 for clarifying the context relevant to the review, and (2) use of 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 *Medical Test Methods Guide* seeks to provide practical guidance to achieving the goal of clarity, consistency, tractability, and usefulness.

References

1. 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.

2. 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.

3. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet 2000;356:1844-7.

4. 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.

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

6. 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.

7. Agency for Healthcare Research and Quality. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville, MD: Agency for Healthcare Research and Quality. Available at: http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318. Accessed September 20, 2010.

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

9. 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.

10. Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis. Science 1959;130:9-21.

11. Yerushalmy J. Statistical problems in assessing methods of medical diagnosis, with special reference to x-ray techniques. Public Health Rep 1947;62:1432-49.

12. 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.

13. 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.

14. 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: www.ahrq.gov/clinic/uspstf08/methods/procmanual.htm. Accessed September 3, 2009.

15. 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.

16. 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.

17. 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.

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

19. 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.

Methods Guide for Medical Test reviews

Paper 2

Developing the Topic and Structuring the Review: Utility of PICOTS, Analytic Frameworks, Decision Trees, and Other Frameworks

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

AHRQ Publication No. xx-EHCxxx-EF
<Month Year>

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, through a collaborative effort within the Evidence-based Practice Center Program, have developed a Methods Guide for Medical Test Reviews. We intend that these documents will serve as a resource for our EPCs as well as for other investigators interested in conducting systematic reviews on medical tests.

This Medical Test Methods guide is intended to be a practical guide for those who prepare and use systematic reviews on medical tests. This document complements the EPC Methods Guide on Comparative Effectiveness Reviews (http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318), which focuses on methods to assess the effectiveness of treatments and interventions. The guidance here applies the same principles for assessing treatments to the issues and challenges in assessing medical tests and 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. We provide guidance in stepwise fashion for those conducting a systematic review.

The *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 *Medical Test Methods Guide* and the Effective Health Care Program can be made at www.effectivehealthcare.ahrq.gov.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

Paper 2. Developing the Topic and Structuring the Review: Utility of PICOTS, Analytic Frameworks, Decision Trees, and Other Frameworks

> *"[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. Developing the topic includes understanding and clarifying how a test might be of value in practice and establishing the key questions to guide decisionmaking related to the claim. This typically involves specifying the clinical context in which the test might be used. 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 EPCs develop and refine the topic, the structure the review should follow ideally becomes clearer. Moreover, success at this stage reduces the chance of major changes in the scope of the review and minimizes rework.

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 achieving 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 might fail to specify the patients who may benefit from the test—from the "worried well" without evidence of deficit to those with frank impairment and loss of function in daily living. Similarly, the request for a review of tests for prostate cancer might neglect to consider the role of such tests in clinical decisionmaking, such as guiding the decision to biopsy.

Because of the indirect impact of medical tests on clinical outcomes, a second problem is how to identify the intermediate outcomes that 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, EPCs should engage in a direct dialogue with the primary requestors or relevant users of the review (herein denoted “stakeholders”) to understand the objectives of the review in practical terms; in particular, EPC investigators should understand the sorts of decisions that the review is likely to affect. Such a discussion also 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., 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. EPCs 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. Including a wide variety of stakeholders and experts (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* (Paper 1), is a typology for defining particular contextual issues, and this formalism can be useful in focusing discussions with stakeholders.

It is important to recognize that the process of topic refinement is iterative. Despite the best efforts of all participants, the topic may change even as the review is being conducted. EPCs should consider at the outset how such a situation will be addressed.[2-4]

Principle 2: Develop an Analytic Framework

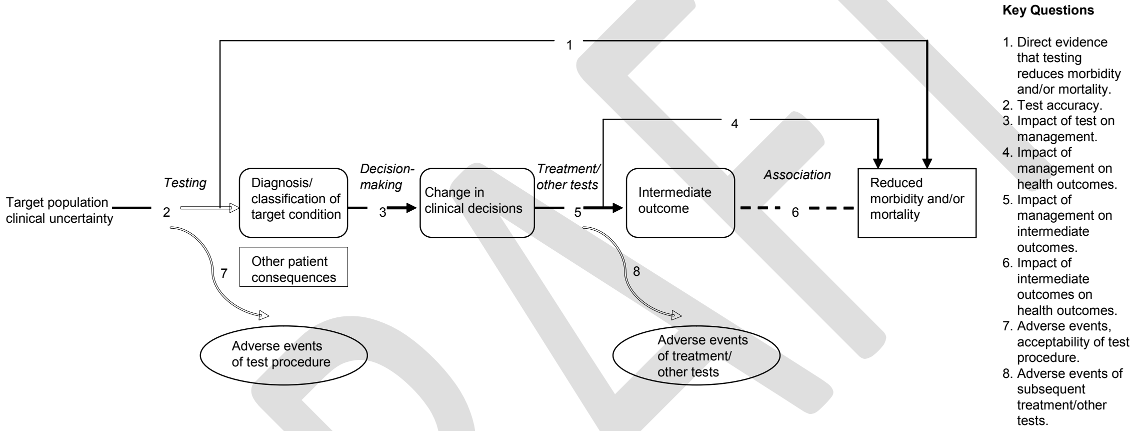
The term “analytic framework” (sometimes called a causal pathway) is used here to denote a specific form of graphical representation that specifies a path from the intervention or test of interest to all important health outcomes, including intervening steps and intermediate outcomes.[5] Each linkage 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. 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.[6]

However, systematic reviews of medical tests present challenges not encountered in reviews of therapeutic interventions. The impact of medical tests on important outcomes is, by nature, more indirect, and there are more potential pathways by which a medical test may affect important outcomes. Because of the often-convoluted linkage to clinical outcomes, research studies mostly focus on intermediate outcomes such as diagnostic accuracy. The analytic framework can help users to understand how these intermediate outcomes fit in the pathway to influencing clinical outcomes, and to consider whether these downstream issues may be relevant to the review.

Harris and colleagues have described the value of the analytic framework in assessing screening tests for the U.S. Preventive Services Task Force (USPSTF).[7] 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


Target population clinical uncertainty
Testing
2
Diagnosis/ classification of target condition
Decision-making
3
Change in clinical decisions
Treatment/ other tests
5
Intermediate outcome
Association
6
Reduced morbidity and/or mortality
1
4
7
Other patient consequences
Adverse events of test procedure
8
Adverse events of treatment/ other tests
Key Questions
1. Direct evidence that testing reduces morbidity and/or mortality.
2. Test accuracy.
3. Impact of test on management.
4. Impact of management on health outcomes.
5. Impact of management on intermediate outcomes.
6. Impact of intermediate outcomes on health outcomes.
7. Adverse events, acceptability of test procedure.
8. Adverse events of subsequent treatment/other tests.

Adapted from Harris et al., 2001[7]

In summarizing evidence, reviewers should remember that 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.

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, the specific potential outcome consequences of altered decisions. EPCs 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. Constructing decision trees may help clarify key questions by identifying which indices of diagnostic accuracy are relevant to the clinical problem and which range of possible pathways and outcomes (see Paper 3) practically and logically flow from a test strategy. Lord et al., describe how decision trees may be used for defining which steps and outcomes may differ with different test strategies, and thus what are 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.[8]

One example of how constructing decision trees can be useful comes from a review of noninvasive tests for carotid artery disease.[9] This review 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.[9-10] Further examples are provided in the Illustrations, below.

Principle 4: Sometimes it is Sufficient to Focus Exclusively on Accuracy Studies

Once EPCs have diagrammed the decision tree by which diagnostic accuracy may affect intermediate and clinical outcomes, reviewers can 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 as the old test *and* the new test's value derives from avoiding the old test's adverse effects (i.e., because the new test is safer or less invasive) or higher costs. 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.[11]

Particular questions that EPCs may consider in reviewing analytic frameworks and decision trees to determine if diagnostic accuracy studies alone are adequate include the following:

1. Are extra cases detected by the new, more sensitive test similarly responsive to treatment?
2. Are trials available that selected patients with 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 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., 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.

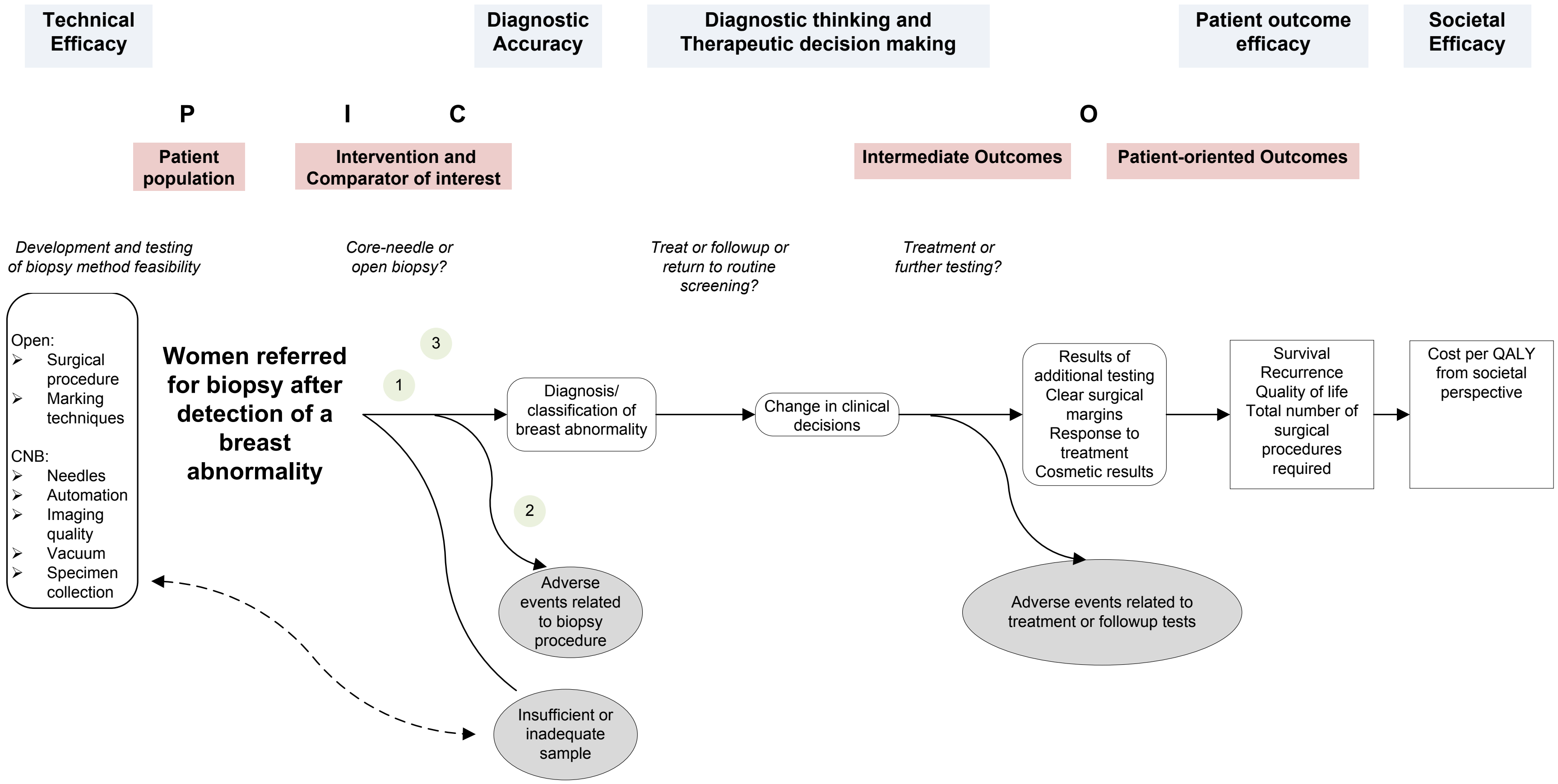
Principle 5: Other Frameworks May be Helpful

Various other frameworks (generally termed "organizing frameworks," as described briefly in the Introduction to this *Medical Test Methods Guide* [Paper 1]) relate to categorical features of medical tests and medical test studies, and reviewers may find these frameworks useful. 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.[12]

These frameworks have been defined for a variety of purposes. Some researchers, such as Van Den Bruel and colleagues, proposed that these frameworks are 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).[13] 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."[12]

We suggest that rather than being a hierarchy of evidence, organizational frameworks are useful in categorizing key questions and which types of studies would be most useful for specific questions in the review. They may be useful in clustering studies to be reviewed together, and this 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.

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


Technical Efficacy
Diagnostic Accuracy
Diagnostic thinking and Therapeutic decision making
Patient outcome efficacy
Societal Efficacy
P
I
C
O
Patient population
Intervention and Comparator of interest
Intermediate Outcomes
Patient-oriented Outcomes
Development and testing of biopsy method feasibility
Core-needle or open biopsy?
Treat or followup or return to routine screening?
Treatment or further testing?
Open:
Surgical procedure
Marking techniques
CNB:
Needles
Automation
Imaging quality
Vacuum
Specimen collection
Women referred for biopsy after detection of a breast abnormality
3
1
2
Diagnosis/ classification of breast abnormality
Change in clinical decisions
Results of additional testing
Clear surgical margins
Response to treatment
Cosmetic results
Survival
Recurrence
Quality of life
Total number of surgical procedures required
Cost per QALY from societal perspective
Adverse events related to biopsy procedure
Insufficient or inadequate sample
Adverse events related to treatment or followup tests

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

Illustrations

To illustrate the principles above, we describe three examples. In each case the initial claim was at least somewhat ambiguous, and through the use of the PICOTS typology, the analytic framework, and simple decision trees, the systematic reviewers were able to work with stakeholders to clarify the objective and analytic approach to the evidence review (Table 2-1).

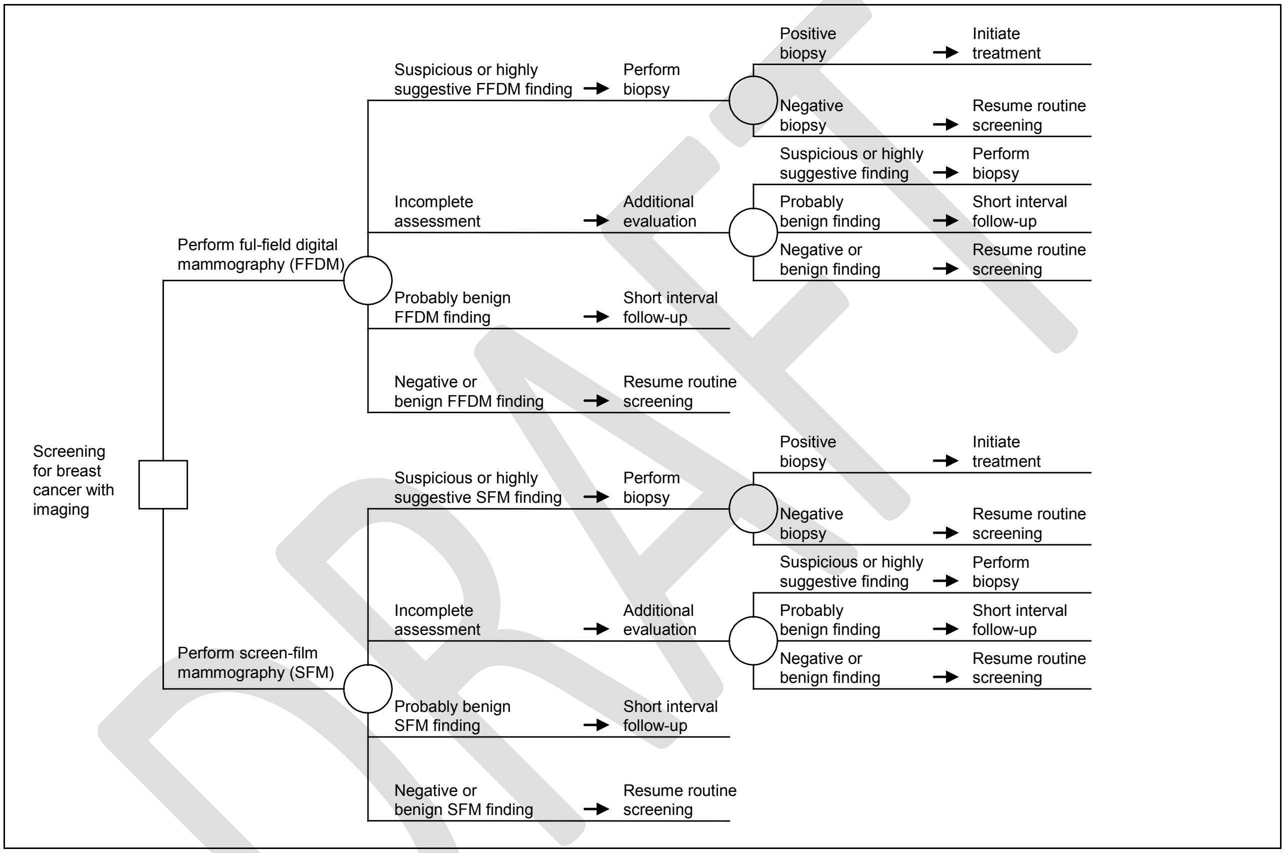
Table 2-1. Examples of initially ambiguous claims that were clarified through the process of topic development

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 accuracy indices: sensitivity, diagnostic yield, recall rate; similar types of management decisions and outcomes for index and comparator test-and-treat strategies	Key accuracy indices: 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 accuracy indices: 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[14]	Seidenfeld et al., 2008[15]	Samson et al., 2002[16]

Abbreviations: 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.[14] Specifying PICOTS elements and constructing an analytic framework were straightforward, with the latter resembling Figure 2-2 in form. In addition, a simple decision tree was drawn (Figure 2-3) which revealed that the management decisions for both screening strategies were similar. The decision tree also showed that the key indices of test performance were sensitivity, diagnostic yield, and recall rate, and given the symmetry of the tree and stakeholder input indicating that the outcomes of a breast cancer identified with one modality or the other was the same, downstream treatment outcomes were not a critical issue. These insights were useful as the project moved to abstracting and synthesizing the evidence, which focused on accuracy and recall rates. As a note, the reviewers concluded that FFDM and SFM had comparable accuracy and led to comparable outcomes; however, storing and manipulating images was much easier for FFDM than for SFM.

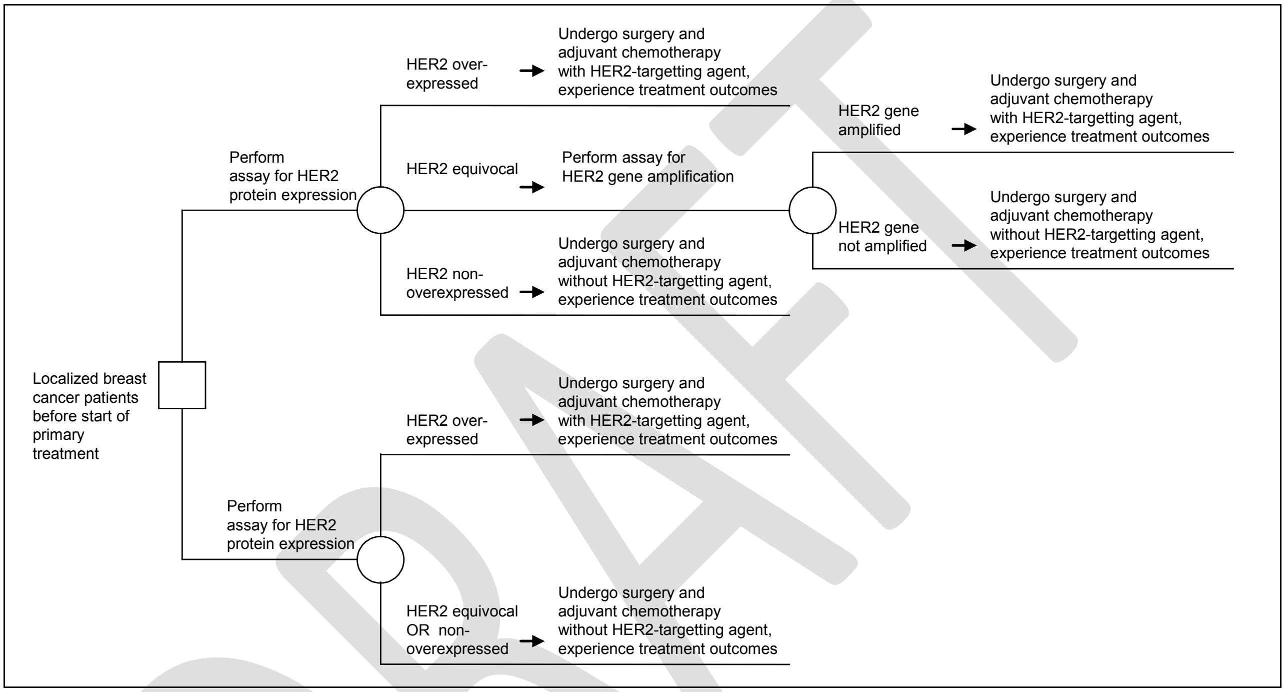
Figure 2-3. Replacement test example: full-field digital mammography versus screen-film mammography*



* 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.[15] The HER2 gene amplification assay has been promoted as an add-on to HER2 protein expression assay. Specifically, individuals with equivocal HER2 protein expression would be followed up with a measure of 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, a decision tree was constructed (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.

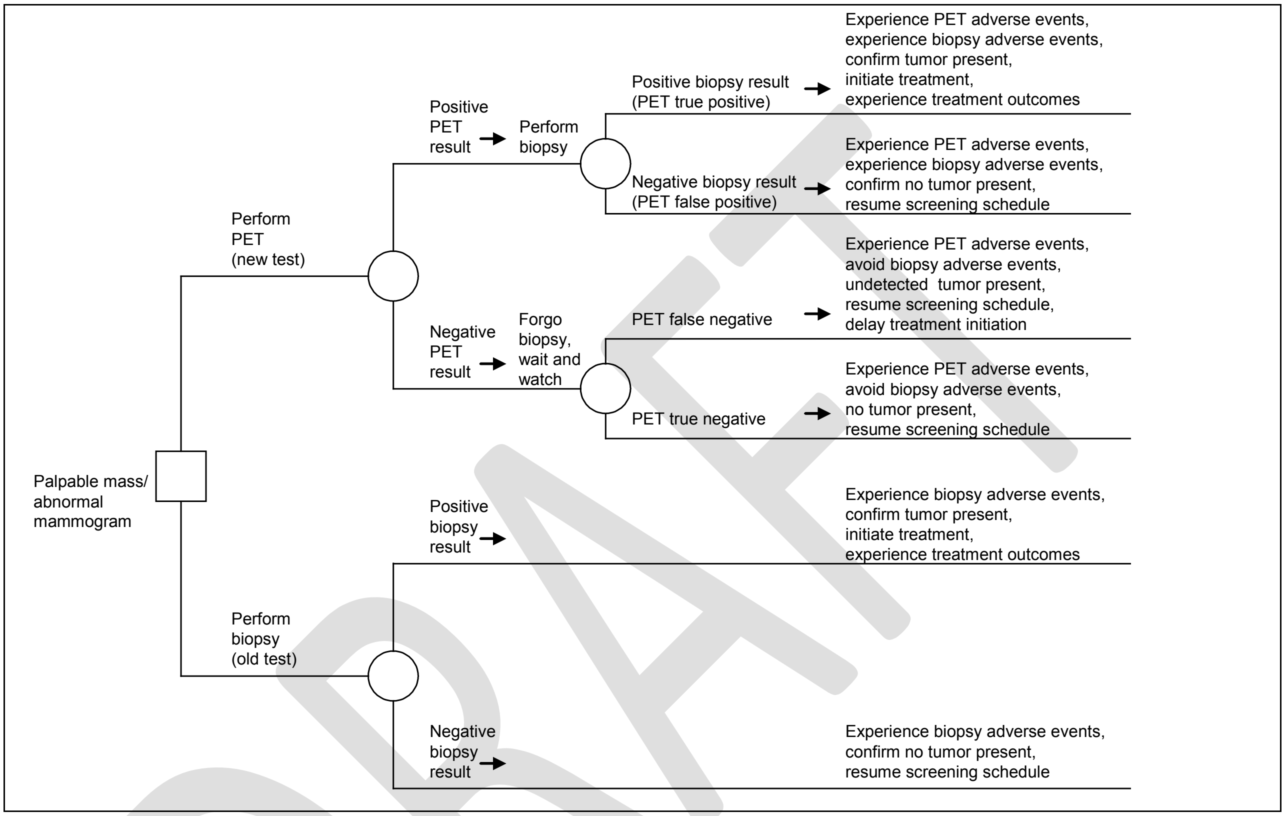
Figure 2-4. Add-on test example: HER2 protein expression assay followed by HER2 gene amplification assay to select patients for HER2-targetted therapy*


Localized breast cancer patients before start of primary treatment
Perform assay for HER2 protein expression
HER2 over-expressed
Undergo surgery and adjuvant chemotherapy with HER2-targetting agent, experience treatment outcomes
HER2 equivocal
Perform assay for HER2 gene amplification
HER2 gene amplified
Undergo surgery and adjuvant chemotherapy with HER2-targetting agent, experience treatment outcomes
HER2 gene not amplified
Undergo surgery and adjuvant chemotherapy without HER2-targetting agent, experience treatment outcomes
HER2 non-overexpressed
Undergo surgery and adjuvant chemotherapy without HER2-targetting agent, experience treatment outcomes
Perform assay for HER2 protein expression
HER2 over-expressed
Undergo surgery and adjuvant chemotherapy with HER2-targetting agent, experience treatment outcomes
HER2 equivocal OR non-overexpressed
Undergo surgery and adjuvant chemotherapy without HER2-targetting agent, experience treatment outcomes

Abbreviation: 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.[16] 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 utility of the analytic framework was limited as several possible testing strategies might be relevant and that is not represented explicitly in an analytic framework. A decision tree was then constructed (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 helps us to see 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 review 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*


Palpable mass/ abnormal mammogram
Perform PET (new test)
Positive PET result
Perform biopsy
Positive biopsy result (PET true positive)
Experience PET adverse events, experience biopsy adverse events, confirm tumor present, initiate treatment, experience treatment outcomes
Negative biopsy result (PET false positive)
Experience PET adverse events, experience biopsy adverse events, confirm no tumor present, resume screening schedule
Negative PET result
Forgo biopsy, wait and watch
PET false negative
Experience PET adverse events, avoid biopsy adverse events, undetected tumor present, resume screening schedule, delay treatment initiation
PET true negative
Experience PET adverse events, avoid biopsy adverse events, no tumor present, resume screening schedule
Perform biopsy (old test)
Positive biopsy result
Experience biopsy adverse events, confirm tumor present, initiate treatment, experience treatment outcomes
Negative biopsy result
Experience biopsy adverse events, confirm no tumor present, resume screening schedule

* Figure taken from Samson et al., 2002.[16]

This case illustrates when a more formal decision analysis may be useful, specifically when 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 evaluate efficiently the relative 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 as the old test *and* the new test's value derives from avoiding the old test's adverse effects (i.e., because the new test is safer or less invasive) or higher costs.

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. p. 80-90.

2. 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 October 4, 2010.

3. 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.

4. 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.

5. 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.

6. Helfand M, 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.

7. 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.

8. 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.

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

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

11. 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.

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

13. 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.

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

15. 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 July 21, 2010.

16. 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.

Methods Guide for Medical Test reviews

Paper 3

Considering the Range of Decision-Relevant Effects

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

AHRQ Publication No. xx-EHCxxx-EF
<Month Year>

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, through a collaborative effort within the Evidence-based Practice Center Program, have developed a Methods Guide for Medical Test Reviews. We intend that these documents will serve as a resource for our EPCs as well as for other investigators interested in conducting systematic reviews on medical tests.

This Medical Test Methods guide is intended to be a practical guide for those who prepare and use systematic reviews on medical tests. This document complements the EPC Methods Guide on Comparative Effectiveness Reviews (http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318), which focuses on methods to assess the effectiveness of treatments and interventions. The guidance here applies the same principles for assessing treatments to the issues and challenges in assessing medical tests and 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. We provide guidance in stepwise fashion for those conducting a systematic review.

The *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 *Medical Test Methods Guide* and the Effective Health Care Program can be made at www.effectivehealthcare.ahrq.gov.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

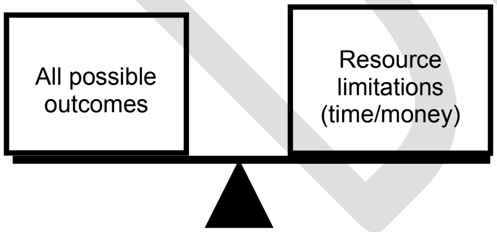
Paper 3. Considering the Range of Decision-Relevant Effects

In this paper, we describe the range of decision-relevant effects, or outcomes, that medical tests have and how these outcomes may be incorporated into a systematic review. “Decision-relevant” refers to those outcomes that result from the testing encounter and have an impact on decisions downstream. The outcomes to be discussed are those that are relevant to screening tests, diagnostic tests, and prognostic tests, although prognostic tests are discussed separately in Paper 12 in this *Medical Test Methods Guide*. We also briefly address unique issues that might arise if the test in question is a genetic test; this topic is explored in greater detail in Paper 11. Other topics considered here include the challenges involved in encompassing a range of outcomes in a systematic review, a framework for generating potential outcomes for inclusion, the role of stakeholders in choosing outcomes for evaluation, and a way to prioritize the outcomes that should be considered. 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 working in Evidence-base Practice Centers (EPCs) are tasked with choosing the outcomes to consider in a systematic review of a medical test. Resource limitations require judicious selection from among all possible outcomes, which necessitates setting priorities for the outcomes to include. If EPCs do not explore the full range of outcomes at the outset of the project, the likelihood of excluding important outcomes is high, and the systematic review may miss outcomes relevant to stakeholders. However, if the initially broad range of outcomes is not carefully reduced, the quality of the review will be threatened by resource limitations (Figure 3-1).

Figure 3-1. Balancing outcomes against resources



If EPCs do not adopt a methodical approach to considering the range of outcomes that might be covered in a review, they may inadvertently exclude important outcomes. On the other hand, if EPCs attempt to cover all outcomes without carefully setting priorities, they may end up with an overly ambitious review. A misstep of either type can result in a suboptimal review—the first type of review may be incomplete, and the second may be too broad to provide meaningful insights.

Principles for Addressing the Challenges

We recommend a two-step approach, applying two principles in sequence, for selecting the outcomes to be included in a review of a medical test. The first step is to catalog potential outcomes methodically, and the second is to solicit input from stakeholders. Below is a description of a conceptual approach to identifying outcomes to ensure that relevant outcomes are not overlooked.

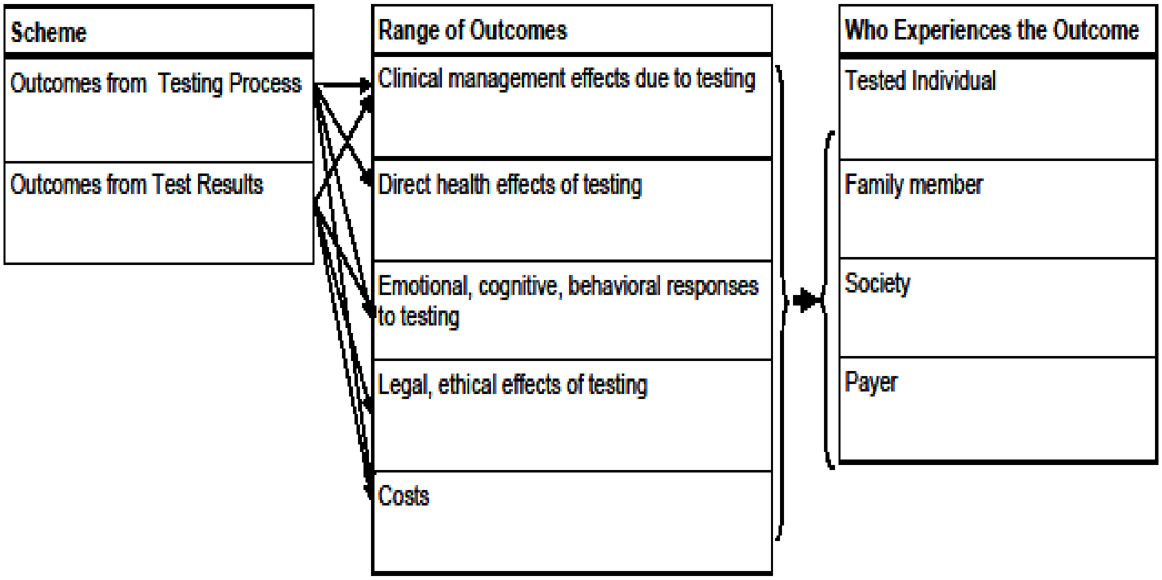
Principle 1: Catalog Outcomes Methodically

Paper 2 describes frameworks for designing systematic reviews of medical tests that include consideration of PICOTs (Patient population, Intervention, Comparator, Outcomes, Timing, Setting). Here we present another framework specifically for thinking about the outcomes attributable to using a test in a clinical setting. Outcomes may be broadly 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 (harms and benefits from the test procedure). Outcomes attributable to the test results are more plentiful and include the patient's response to the test results and outcomes deriving from how the patient and clinician act upon the results.

Bossuyt and McCaffery recently described a useful framework for thinking about patient outcomes attributable to medical testing.[1] They classified outcomes into three groups—outcomes that result from (1) clinical management based on the test results, (2) the direct health effects of testing, and (3) the patients' emotional, 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 appropriate depending on the test under consideration, and (5) the costs of the test. These five categories of outcomes can be associated with the testing process or the test result, or with both. For example, a medical test for HIV may have behavioral responses associated with the testing process: the act of getting an HIV test is associated with getting other tests, such as for hepatitis C. However, behavioral responses to the test result are also conceivable, including high-risk sexual behavior if the test is negative.

Reviewers should 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 (e.g., 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 that 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 illustrates these additional considerations.

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


Scheme
Outcomes from Testing Process
Outcomes from Test Results
Range of Outcomes
Clinical management effects due to testing
Direct health effects of testing
Emotional, cognitive, behavioral responses to testing
Legal, ethical effects of testing
Costs
Who Experiences the Outcome
Tested Individual
Family member
Society
Payer

In summary, the range of outcomes that might be included in a systematic review of a medical test is wide. We encourage EPCs to think systematically through this range of outcomes and consider the testing process, the test results, the range of associated outcomes, and the parties that may experience the outcome. These considerations may differ depending on the type of test under evaluation, as discussed below.

Principle 2: Solicit Input From Stakeholders

As described above, the range of outcomes that EPCs might include in a systematic review of a medical test is broad, and expecting such reviews to include all possible outcomes is unrealistic due to time and resource limitations. The Agency for Healthcare Research and Quality's (AHRQ's) *General Methods Guide* recommends that stakeholders be involved at several steps in the systematic review process.[2] We describe additional considerations regarding the role of stakeholders in reviews of medical tests, as their input is 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. We suggest that the choice of outcomes depends largely on the needs of stakeholders and how they intend to use the review. The stakeholders (or sponsors) who submit requests for evidence reports from the EPCs represent many different interests. Therefore, the outcomes they consider essential will also vary.

For example, 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.[3-5] 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. The outcomes EGAPP considers to be relevant are correspondingly 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 EPC has a responsibility to work with stakeholders to refine the questions carefully so that the task can be accomplished.

Other stakeholders (e.g., professional societies such as the American College of Physicians) may be most interested in systematic reviews 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 they may be less interested in outcomes that are more relevant to payers, such as cost-shifting to accommodate costs of testing and downstream costs.

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

EPCs have a role in helping stakeholders understand the breadth of outcomes that could be considered and to think through the clinical questions. Conversely, EPCs also have the responsibility of focusing their key questions so that the selected outcomes can be addressed in accordance with the resources allocated. These choices will depend on the context within which the evidence review is being done. Investigators should 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

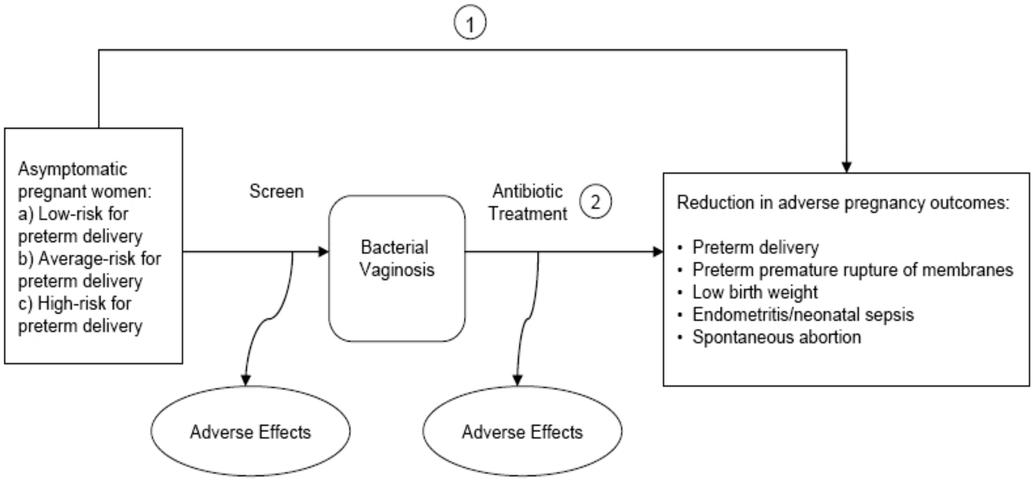
To explain these points in greater detail, we describe three examples: one each of a screening test, a diagnostic test, and a prognostic test. We assume that after EPC investigators have done the mapping proposed above (Figure 3-2), they will discuss the range of outcomes that might be included in a systematic review with stakeholders. In discussing these examples, we consider outcomes that result from the process of testing, outcomes associated with the results of testing, and outcomes that affect the tested individual and others. We conclude with a discussion of additional considerations for evaluating a genetic test. In these illustrations, we are not suggesting that the reviewers should necessarily include any or all of the potential outcomes in their evaluation; we are simply demonstrating how one might go through the exercise of considering each class of potentially relevant outcomes.

Example of a screening test. Screening tests are used to detect disease in individuals who are asymptomatic or who have unrecognized symptoms.[6] In essence, screening tests should be able to separate individuals with the disease of interest from those without the disease and should be used when there is a treatment available and where early treatment is known to improve outcomes.

The United States Preventive Services Task Force (USPSTF) is an independent panel of experts that systematically reviews the evidence and develops recommendations for use of clinical preventive services in the United States. The USPSTF makes recommendations about the use of screening tests such as colonoscopy or mammography. An EPC is sometimes tasked with preparing the supporting systematic review of the evidence.[7-8] Other stakeholders obviously have an 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[9] and the American Cancer Society's recommendations for early cancer detection.[10]

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.[11] This 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


1
Asymptomatic pregnant women:
a) Low-risk for preterm delivery
b) Average-risk for preterm delivery
c) High-risk for preterm delivery
Screen
Bacterial Vaginosis
Antibiotic Treatment
2
Reduction in adverse pregnancy outcomes:
• Preterm delivery
• Preterm premature rupture of membranes
• Low birth weight
• Endometritis/neonatal sepsis
• Spontaneous abortion
Adverse Effects
Adverse Effects

Critical Key Questions:

1. Does screening for bacterial vaginosis during pregnancy in asymptomatic women reduce adverse pregnancy outcomes for those at:
   a. low-risk for preterm delivery?
   b. average-risk for preterm delivery?
   c. high-risk for preterm delivery?
2. Does treatment of bacterial vaginosis during pregnancy in asymptomatic women reduce adverse pregnancy outcomes for those at:
   a. low-risk for preterm delivery
   b. average-risk for preterm delivery
   c. high-risk for preterm delivery

Clinical management effects. The authors of this review 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 include adverse effects of therapies and the beneficial effects of reduction in adverse pregnancy outcomes, such as preterm delivery. The authors might also have explicitly included an outcome that examines whether 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 effects are unlikely in this example. The screening test (a vaginal swab) should not result in any injury; neither does the test confer any direct benefit because the testing procedure does not treat the infection. The process of screening may, however, include added contact with health care providers, which may confer some direct benefit. Thus, the authors might have included this as a decision-relevant effect.

Emotional, cognitive, and behavioral effects. The authors might also have looked at emotional, cognitive, and behavioral effects from the screening process or from the screening test results. It may have been appropriate to consider outcomes that are associated with screening but are not the result of antibiotic therapy. In thinking about these effects, consideration may be given to the emotional or cognitive effects of the testing process (likely to be few in this example) and the effects stemming from 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 itself. As with any measure, the EPC may require that the instrument used to measure emotional response be a validated and appropriate instrument.

Legal and ethical effects. 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. The authors of this review did not consider the costs of the test to the patient as an outcome, probably because such costs are unlikely to be very important in this example.

Parties experiencing the effects. The authors of this review on screening for bacterial considered the effects of screening for bacterial vaginosis 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 since 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 screening tests are used in asymptomatic or presymptomatic people, diagnostic tests are applied to confirm or refute disease in symptomatic individuals. 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 the use of diagnostic tests because many are costly. EPCs have been involved in many systematic reviews of diagnostic tests.

We discuss the example of a systematic review from outside the EPCs that addressed the diagnostic value of 64-slice computed tomography (CT) in comparison to conventional coronary angiography.[12] 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. This was a very narrowly focused review since an assessment of accuracy alone cannot address the question of whether or when CT angiography should replace conventional angiography. If an EPC were to assess the effectiveness of CT angiography, the investigators should consider the full range of outcomes.

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 identification of occluded coronary arteries. 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 which has associated benefits and harms. All of these may be appropriate outcomes to include in evaluating a diagnostic test. If one test under consideration reports more or fewer occluded coronary arteries (correctly or not) than another, this will be reflected in more or fewer clinical management interventions and their resulting outcomes.

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 or colonoscopy) 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, cognitive, and behavioral effects. The testing process itself is unlikely to have significant emotional consequences since it is not an invasive test and is generally comfortable for the tested individual (unlike other radiographic procedures such as magnetic resonance imaging). 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 in riskier behavior. These behavioral effects are likely to affect the family members of the tested individuals 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. 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, for example, the tested person were an airline pilot. However, this outcome is not expected to differ between CT and conventional angiography.

Costs. 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.

Example of a prognostic test. Prognostic tests are tests 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 diagnostic test, but the results are applied with a different purpose. Given that this is the case, additional considerations for outcomes should be included in reviews of prognostic tests. For example, consider the use of spirometry for predicting prognosis in individuals with chronic obstructive pulmonary disease (COPD). The test is commonly used for making the diagnosis of COPD and monitoring 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.[13] The discussion below focuses on one of their key questions, which was whether prediction of prognosis with spirometry, with or without clinical indicators, is more accurate than prediction based on clinical indicators alone. Investigators were interested in predicting 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 (e.g., some malignancies, some dementing illnesses), most prognostic information can be used to alter the course of treatment. In the present example, spirometry may suggest a high likelihood of progressing to respiratory failure and prompt initiation of processes to avert this possibility (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 of respiratory failure is high, patients and their physicians may choose not to proceed with a 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 they are 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, cognitive, and 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 to the results of testing, however, are likely; they may be even more pronounced for prognostic tests than for screening or medical tests because the test may yield more specific information about mortality risk than is usual from a diagnostic test. This could have a range of effects on behavior, including efforts to alter prognosis (e.g., smoking cessation). Test results with prognostic information would be expected to affect family members as well.

Legal and ethical effects. 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 the example being considered here, it is unlikely that the prognostic information from spirometry would actually raise legal issues, but in other cases, such as a test that demonstrates widely metastatic cancer, legal or ethical issues might arise. 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. The costs of the test to the insurer and the patient, relative to the costs of collecting information from a history and physical examination, could be of interest to stakeholders.

Additional considerations involved in evaluating genetic tests. Paper 11 describes in detail unique issues regarding evaluation of genetic tests. With respect to relevant outcomes, we note only a few considerations here. Most prominent is the effect of genetic testing on family members. Genetic information about the tested individual has direct bearing on family members who share genes. Emotional and behavioral outcomes need to be considered, as well as 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 the possible impact of testing 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 systematic reviews of medical tests, EPCs should remember these key points:

- Consider both outcomes relevant to the testing process and outcomes relevant to the test results.
- Consider inclusion of outcomes in all five domains: clinical management effects; direct test effects; emotional, cognitive, and behavioral effects; legal and ethical effects; and costs.
- As part of the process of choosing the outcomes for inclusion, consider to whom the outcomes are most relevant.
- Given resource limitations, prioritize which outcomes to include. This decision depends on the needs of the stakeholders, 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(5):E30-8.

2. Agency for Healthcare Research and Quality. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville, MD: Agency for Healthcare Research and Quality. Available at: http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318. Accessed September 20, 2010.

3. 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. (Prepared by the Duke Evidence-based Practice Center under Contract No. 290-02-0025.) AHRQ Publication No. 07-E002. Rockville, MD: Agency for Healthcare Research and Quality. January 2007 Available at: http://www.ahrq.gov/downloads/pub/evidence/pdf/cyp450/cyp450.pdf. Accessed July 21, 2010.

4. Bonis PA, Trikalinos TA, Chung M, et al. Hereditary Nonpolyposis Colorectal Cancer: Diagnostic Strategies and Their Implications. Evidence Report/Technology Assessment No. 150 (Prepared by Tufts-New England Medical Center Evidence-based Practice Center under Contract No. 290-02-0022). AHRQ Publication No. 07-E008. Rockville, MD: Agency for Healthcare Research and Quality. May 2007. Available at: http://www.ahrq.gov/downloads/pub/evidence/pdf/hnpcc/hnpcc.pdf. Accessed September 30, 2010.

5. 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 July 2, 2010.

6. Wilson JMG, Jungner G. Principles and practice of screening for disease. WHO Chronicle 1968;22(11):473.

7. Hillier TA, Vesco KK, Pedula KL, et al. Screening for gestational diabetes mellitus: a systematic review for the U.S. Preventive Services Task Force. Ann Intern Med 2008;148(10):766-75.

8. Whitlock EP, Lin JS, Liles E, et al. Screening for colorectal cancer: a targeted, updated systematic review for the U.S. Preventive Services Task Force. Ann Intern Med 2008;149(9):638-58.

9. Waxman AG. Guidelines for cervical screening: history and scientific rationale. Clin Obstet Gynecol 2005;48:77-97.

10. 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. Cances J Clin 2008;58(3):161-79.

11. Nygren P, Fu R, Freeman M, et al. Evidence on the benefits and harms of screening and treating pregnant women who are asymptomatic for bacterial vaginosis: an update review for the U.S. Preventive Services Task Force. Ann Intern Med 2008;148(3):220-33.

12. 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(11):1386-93.

13. Wilt TJ, Niewoehner D, Kim C-B, et al. Use of Spirometry for Case Finding, Diagnosis, and Management of Chronic Obstructive Pulmonary Disease (COPD). Evidence Report/Technology Assessment No. 121 (Prepared by the Minnesota Evidence-based Practice Center under Contract No. 290-02-0009.) AHRQ Publication No. 05-E017-2. Rockville, MD. Agency for Healthcare Research and Quality. September 2005. Available at: http://www.ahrq.gov/downloads/pub/evidence/pdf/spirocopd/spiro.pdf. Accessed July 23, 2010.

Methods Guide for Medical Test reviews

Paper 4

Searching for Studies

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

AHRQ Publication No. xx-EHCxxx-EF
<Month Year>

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, through a collaborative effort within the Evidence-based Practice Center Program, have developed a Methods Guide for Medical Test Reviews. We intend that these documents will serve as a resource for our EPCs as well as for other investigators interested in conducting systematic reviews on medical tests.

This Medical Test Methods guide is intended to be a practical guide for those who prepare and use systematic reviews on medical tests. This document complements the EPC Methods Guide on Comparative Effectiveness Reviews (http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318), which focuses on methods to assess the effectiveness of treatments and interventions. The guidance here applies the same principles for assessing treatments to the issues and challenges in assessing medical tests and 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. We provide guidance in stepwise fashion for those conducting a systematic review.

The *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 *Medical Test Methods Guide* and the Effective Health Care Program can be made at www.effectivehealthcare.ahrq.gov.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

Paper 4. Searching for Studies

Locating all published studies relevant to the key questions is a goal of all systematic reviews. Inevitably, Evidence-based Practic Centers (EPCs) 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 problems to identify all relevant studies, taking into account the usual constraints on time and resources.

With studies of medical tests, locating all available studies is especially important because the results of studies of medical tests themselves tend to be highly variable.[1-2] In the face of such challenges, searches need to use multiple approaches to have high recall (sensitivity). Unfortunately, sensitivity usually comes at the cost of relevance (specificity). Any systematic review of medical tests is likely to involve a good deal of human labor identifying relevant articles from large batches of potentially relevant articles to be sure that none is missed. In this paper, we discuss some of the challenges in identifying studies, focusing on medical tests.

Common Challenges

Systematic reviews of test strategies for a given condition require a search on each of the relevant test strategies under consideration. In conducting the search, an EPC may use one of two approaches. The EPC may search on all possible tests used to evaluate the given disease, which requires knowing all the possible test strategies available, or the EPC 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 diagnostic concept may not be necessary.[3-4]

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®.[1,3-6] In particular, one filter[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[2,6,11-12] 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; [12]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.[11] In fact, Kastner and colleagues[13] recently reviewed 22 systematic reviews of diagnostic accuracy published in 2006 to determine whether the PubMed Clinical Queries 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 EPCs not rely on them exclusively.

Principle 2: Do not rely on controlled vocabulary alone

It is important to use all known variants of the test name when searching, 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.[4]

Because indexing of medical tests is variable, 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.[3] Filters may suggest the sort of textwords that may be appropriate. As always—but in particular with searches for studies of medical tests—we advise EPCs to search more than one database and to tailor search strategies to each individual database.[14]

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.[2,6,11-12] Even an initial sensitive search is likely to miss relevant articles, and following cited references from relevant articles (hand searching) and identifying articles that cite key studies remain important sources of citations.[15]

Because the FDA regulates many medical tests as medical devices, another potential source of information is regulatory documents. Reviewers who know the name of specific tests can search for regulatory documents at the FDA’s Device website: http://www.accessdata.fda.gov/scripts/cdrh/devicesatfda/index.cfm.

Illustration

In the AHRQ report, *Testing for BNP and NT-proBNP in the Diagnosis and Prognosis of Heart Failure*,[16] 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*,[17] 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 are:

- Currently, 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, EPCs 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.

References

1. Bachmann LM, Coray R, Estermann P, et al. Identifying diagnostic studies in MEDLINE: reducing the number needed to read. J Am Med Inform Assoc 2002;9(6):653-8.

2. 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.

3. Deville WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. J Clin Epidemiol 2000;53(1):65-9.

4. van der Weijden T, CJ IJ, Dinant GJ, et al. Identifying relevant diagnostic studies in MEDLINE. The diagnostic value of the erythrocyte sedimentation rate (ESR) and dipstick as an example. Fam Pract 1997;14(3):204-8.

5. Haynes RB, Wilczynski N, McKibbon KA, et al. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. J Am Med Inform Assoc 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 Info Libr J 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, Estermann P, Kronenberg C, et al. Identifying diatnostic accuracy studies in EMBASE. J Med Lib Assoc 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 Med 2005;3:7.

11. Doust JA, Pietrzak E, Sanders S, 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.

12. 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 Info Libr J 2003;20(3):150-9.

13. Kastner M, Wilczynski NL, McKibbon AK, et al. Diagnostic test systematic reviews: bibliographic search filters ("Clinical Queries") for diagnostic accuracy studies perform well. J Clin Epidemiol 2009;62(9):974-81.

14. Honest H, Bachmann LM, Khan K. Electronic searching of the literature for systematic reviews of screening and diagnostic tests for preterm birth. Eur J Obstet Gynecol Reprod Biol 2003;107(1):19-23.

15. 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.

16. Balion C, Santaguida P, Hill S, 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: www.ahrq.gov/downloads/pub/evidence/pdf/bnp/bnp.pdf. Accessed April 16, 2009.

17. 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 April 16, 2009.

Methods Guide for Medical Test reviews

Paper 5

Assessing Individual Study Limitations (Risk of Bias) as a Domain of Quality

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

AHRQ Publication No. xx-EHCxxx-EF
<Month Year>

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, through a collaborative effort within the Evidence-based Practice Center Program, have developed a Methods Guide for Medical Test Reviews. We intend that these documents will serve as a resource for our EPCs as well as for other investigators interested in conducting systematic reviews on medical tests.

This Medical Test Methods guide is intended to be a practical guide for those who prepare and use systematic reviews on medical tests. This document complements the EPC Methods Guide on Comparative Effectiveness Reviews (http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318), which focuses on methods to assess the effectiveness of treatments and interventions. The guidance here applies the same principles for assessing treatments to the issues and challenges in assessing medical tests and 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. We provide guidance in stepwise fashion for those conducting a systematic review.

The *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 *Medical Test Methods Guide* and the Effective Health Care Program can be made at www.effectivehealthcare.ahrq.gov.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

Paper 5. Assessing Individual Study Limitations (Risk of Bias) as a Domain of Quality

After identifying all studies to be included in the assessment of the benefits and harms of a medical test, systematic reviewers in EPCs must critically consider the contribution of each study to answering the research question of interest. A variety of study designs and approaches can be used to evaluate medical tests, and these study designs and approaches can affect the estimates and interpretation of the study results. Some differences lead to systematic bias such that estimates of test performance will differ from their true values. Other differences in study design and conduct can give rise to heterogenity across studies, which can limit applicability and interpretation. The merits of an individual study must be appraised for several elements, including the methodological limitations and study design, the direction and magnitude of results, the sample size of the study, directness of comparisons and outcomes, and the relevance of results. All of these elements reflect our understanding of the overall "quality" of the study being evaluated.

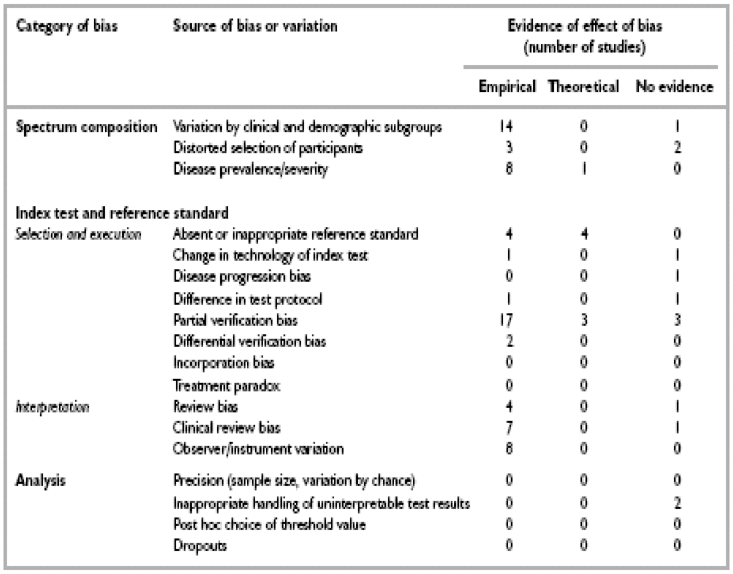
In this paper we focus on one aspect of study quality. Specifically, this paper addresses the study limitations that may affect the internal validity of the results. Internal validity refers to the confidence that the results are a "true" representation of the state of the world for the group studied. Both random error (imprecision) and systematic error (bias) can affect the internal validity of a study. In a systematic review, however, the potential for random error is best considered across studies (rather than within an individual study) when assessing the precision of combining results to estimate a summary effect measure. In contrast, when considering the potential for systematic error, which is most consistent with the internal validity of individual studies, there is the need to evaluate studies individually for this risk of bias. By definition, systematic error (bias) may lead to a constant over- or underestimation of the test performance. As such, there is the need to evaluate systematic error in individual studies within a systematic review in an attempt to identify the specific sources of error (e.g., a consistent failure to randomize adequately or to blind outcome assessors adequately). External validity concerns issues of applicability or generalizability and is discussed in Paper 6.

Evaluation of individual study systematic error is often described as an assessment of "risk of bias" or "study limitations" and focuses on factors traditionally understood as internal validity. For the purposes of this paper we will use the term "risk of bias" to reflect the appraisal of individual studies for specific internal validity criteria. For example, the standardized QUADAS instrument (Quality Assessment of Diagnostic Accuracy Studies), which is comprised of 14 criteria, is used to rate the study on risk of bias. Consideration is given to the number (e.g., 6 from 14 criteria) or specific types of risk of bias criteria to then determine the category for study limitations. For practical purposes, it can be useful to use these assessments to categorize studies according to their potential risk of bias (e.g., high, fair, or low risk of bias).

It is important to note that, however defined, study limitations can be assessed within an individual study or across several studies. When grading the strength of evidence across studies (see Paper 7), "study limitations" is one of five domains of quality considered.

Table 5-1 summarizes the empirical evidence from a literature review assessing the effects of specific types of biases (systematic error) in diagnostic test studies. This review evaluated literature from 1966 to 2000.[1] Although this review identified the relevant evidence (from 55 eligible studies) to show the impact of specific biases on medical test results, no conclusions could be drawn about the direction or relative magnitude of effect for these specific biases. This would suggest that more research is required to better estimate the impact of the specific biases on studies evaluating medical tests. However, the limitations of the literature establishing empirical evidence for the impact of biases in medical tests does not differ from that of intervention studies.

Table 5-1. Empirical evidence on effects of specific types of bias in diagnostic test studies[1]



Category of bias	Source of bias or variation	Evidence of effect of bias (number of studies)		
		Empirical	Theoretical	No evidence
Spectrum composition	Variation by clinical and demographic subgroups	14	0	1
	Distorted selection of participants	3	0	2
	Disease prevalence/severity	8	1	0
Index test and reference standard				
Selection and execution	Absent or inappropriate reference standard	4	4	0
	Change in technology of index test	1	0	1
	Disease progression bias	0	0	1
	Difference in test protocol	1	0	1
	Partial verification bias	17	3	3
	Differential verification bias	2	0	0
	Incorporation bias	0	0	0
	Treatment paradox	0	0	0
Interpretation	Review bias	4	0	1
	Clinical review bias	7	0	1
	Observer/instrument variation	8	0	0
Analysis	Precision (sample size, variation by chance)	0	0	0
	Inappropriate handling of uninterpretable test results	0	0	2
	Post hoc choice of threshold value	0	0	0
	Dropouts	0	0	0

Study design and conduct elements 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 the criteria used for rating the quality of intervention studies (see the chapter on "Assessing the Quality and Applicability of Included Studies" in AHRQ's *General Methods Guide*[2]). However, diagnostic accuracy studies differ from intervention studies in the most appropriate study design to assess the various measurements of accuracy outcomes, as well as in the various potential biases that must be considered (e.g., complete ascertainment of true disease status, adequacy of reference standard, and spectrum effect). Because of these unique challenges, this chapter will focus on assessing the risk of bias of individual studies of medical test performance.

Common Challenges

In assessing the risk of bias in diagnostic accuracy studies, several common challenges arise. The first is identifying the appropriate criteria to use. A number of instruments are available for assessing risk of bias in medical test performance studies; however, these scales and checklists include criteria 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] Moreover, available checklists and scales may not address some study weaknesses (e.g., basic study design attributes such as randomization or blinding).

Other common challenges are not unique to medical test studies. EPCs must consider how to apply the criteria and how to deal with inadequate reporting. Once criteria are selected, there is always some subjectivity in determining how well a study meets the criteria. Similarly, assessment of an individual criterion may be determined by various ratings for risk of bias (e.g., yes/no or other), which could lead to issues of rationale or consistency when combining multiple criteria to give an overall assessment, and by whether some criteria are weighted more heavily than others (i.e., a “fatal flaw”). Although judgment and pragmatic choices will lead to the final methods used to assess risk of bias (selecting specific criteria, categorize studies into high or low risk, and incorporating study limitations in grading of the strength of evidence), it will be important to provide justification and transparency in how the appraisal was undertaken.

Although inadequacy of reporting in itself does not lead to systematic bias, adequate assessment of important risk of bias criteria is limited by what was reported; thus, fairly or unfairly, studies with less meticulous reporting may be assessed as having been less meticulously performed and as not deserving the degree of attention given to well-reported studies. In such cases, when a study is otherwise judged otherwise to make a potentially important contribution, reporting questions may need to be addressed to the study’s authors.

Principles for Addressing the Challenges

Principle 1: Identify Relevant Sources of Bias

Systematic error may be introduced into a medical test performance study in numerous ways. Table 5-2 summarizes the most common sources of such bias.[1,4]

Table 5-2. Commonly reported sources of systematic error in diagnostic accuracy studies

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	The 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 in practice, 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 can be the result of differences in test execution.

Source of systematic bias	**Description**
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	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	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 or standards 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 (reading process)	
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	The availability of information on 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 of the determinants of 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 overoptimistic 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.

Although the study features listed above all have the potential to cause systematic bias, some may also be a source of random error (such as observer variability) or generate issues related to applicability (such as variation in test execution or technology, using an inappropriate reference standard, or spectrum effect). Also, features related to analysis of data at the individual study level may not be relevant when conducting a systematic review. For example, arbitrary choice of threshold value may not bias a systematic review if the study provides sufficient data to allow the EPC investigators to perform their own reanalysis.

EPCs should consider which of the sources of bias listed above are most likely to systematically bias the results of the study and identify criteria that will adequately address the biases. EPCs may also identify particular issues that will be considered at other stages of the review process, such as assessing applicability (*see* Paper 6) or grading the strength of evidence across studies (*see* Paper 7)

Principle 2: Consider Using Validated Criteria to Address Relevant Sources of Bias

One published tool that includes validated criteria is the QUADAS scale (Table 5-3).[5-7] This tool contains elements of study limitations beyond those concerned with risk of systematic bias; it also includes questions related to reporting. In applying QUADAS or other instruments, EPCs need to clarify how particular items apply to the systematic review at hand.

Table 5-3. Criteria within the QUADAS scale for assessing quality of diagnostic accuracy studies[7]

1)	Was the spectrum of patients representative of the patients who will receive the test in practice?
2)	Were the selection criteria clearly described?
3)	Is the reference standard likely to correctly classify the target intervention?
4)	Is the time period between reference standard and index test short enough to be reasonably sure the target condition did not change between the two tests?
5)	Did the whole sample or a random selection of the sample receive verification using a reference standard of diagnosis?
6)	Did the patients receive the same reference standard regardless of the index test?
7)	Was the reference standard independent of the index test (i.e., the index test did not form part of the reference standard)?
8)	Was the execution of the index test described in sufficient detail to permit replication of the test?
9)	Was the execution of the reference standard described in sufficient detail to permit its replication?
10)	Were the index test results interpreted without knowledge of the results of the reference standard?
11)	Were the reference standard results interpreted without knowledge of the index test?
12)	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?
13)	Were uninterpretable/intermediate test results reported?
14)	Were withdrawals for the study explained?

We recommend that EPCs identify criteria that assess the risk of systematic error that have been validated to some degree from an instrument like QUADAS, putting aside other items that assess applicability or random error to be considered at a different stage of the review (*see* Papers 6 and 7). In addition to deleting irrelevant assessments, systematic reviewers may also need to add additional criteria, which may be identified from other standardized checklists. Other standardized scales or reporting standards include the Standards for Reporting of Diagnostic Accuracy (STARD)[8-10] and the Strengthening the Reporting of Genetic Association Studies (STREGA),[11] (an extension of the Strengthening the Reporting of Observational Studies in Epidemiology [STROBE]).[12]

Principle 3: Make Decisions Transparent

There is little empiric evidence to inform decisions such as handling of inadequate reporting and methods for applying and summarizing criteria. We make suggestions below and recommend that EPCs consider a priori how they will handle these issues and document these decisions.

Consistent with previous EPC guidance and other published recommendations,[1,13] we suggest summarizing study limitations across multiple items for a single study into simple categories; following the guidance given in AHRQ's General Methods Guide,[1] we propose using the terms "good," "fair," and "poor" for this purpose. Table 5-4 illustrates how the application of these three categories may be interpreted in the context of diagnostic accuracy studies.

Table 5-4. Categorizing individual studies into general quality classes (adapted from AHRQ's *General Methods Guide*[1])

Category	Application to randomized controlled trials	Application to medical test performance studies
Good. 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, analysis by intention-to-treat; and key study features are described clearly, including the population, setting, interventions, comparison groups, measurement of outcomes, and reasons for dropouts.	RCTs are considered a high study design type, but studies that include consecutive patients representative of the intended sample for whom diagnostic uncertainty exists may also meet this standard. A "good" study avoids the multiple biases to which medical test studies are subject (e.g., use of an inadequate reference standard, verifcation bias), and key study features are clearly described, including the comparison groups, measurement of outcomes, and the characteristics of patients who failed to be have actual state (diagnosis or prognosis) verified.
Fair. 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 good quality, but no flaw is likely to cause major bias. The study may be missing information, making it difficult to assess limitations and potential problems.	Similar to RCTs (neither "good" nor "poor").
Poor. Significant flaws that imply biases of various types that may invalidate the results	The study has serious errors in design, analysis, or reporting; large amounts of missing information; or discrepancies in 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).

Once all criteria have been rated for each study, it is advised that these are reported in their entirety (either in an appendix or within figures). In order to determine the overall "risk of bias" based upon individual criteria, EPCs must carefully consider what criteria would qualify a study as poor, such as the presence of a certain percentage of "low" criteria ratings, or perhaps presence of a fatal flaw or flaws. Input from clinical and methodological experts on the project's Technical Expert Panel (TEP) may be helpful. Regardless, the rationale for the criteria and approach should be established a priori and detailed within the methods section of the review.

EPCs must also carefully consider how to handle inadequate reporting. Inadequate reporting, in and of itself, does not induce systematic bias, but limits the ability to assess the risk of bias. Some systematic reviewers may "assume the worst," while others may prefer to "give the benefit of the doubt." Some instruments, such as QUADAS, provide some guidance as to the criteria to use to indicate that an item is not clear due to reporting inadequacy. Again, when a study is otherwise deemed to make a potentially important contribution to the review, issues of reporting may be resolved by contacting study authors. In any case, EPCs 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 reporting family history and the factors that were likely to affect accuracy.[14-15] In the context of this review, the index test was patients' self-report of their family history, and the reference standard test could include verification of the relatives' status from either medical records or a disease or death registry. 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 EPC provided contextual examples of how reviewers had adapted each QUADAS item for the review. As noted in Table 5-5, partial verification bias was defined in the context of self-reported family history as the index test, and verification in the relatives (by either direct contact or health record or disease registry) was the reference test. Decision rules for rating this quality criterion as "yes," "no," or "unclear" are explicitly given.

Table 5-5. How verification bias was interpreted for the family history example[14-15]

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 occurs when not all of the study participants receive the reference standard (in our context, confirmation of the TRUE disease status of the relative). This is a form of selection bias. Sometimes the reason only a part of the sample receives the reference standard is because knowledge of the index test results **influence** the decision to perform the reference standard. Please note that, in the context of family history, the reference standard can only be applied to the family members or relatives. The self report of 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 truth of the 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).

Modified QUADAS item (*Topic/Bias*)	Interpretation
	3) Relatives who are **DECEASED**; as such they 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.

Abbreviation: QUADAS = Quality Assessment of Diagnostic Accuracy Studies.

The EPC presented results of applying the adapted QUADAS in tabular form as a percentage of the studies that scored yes, no, or unclear.

Summary

Assessing the overall quality of an individual study involves assessing 1) the size of the study, 2) the direction and degree of findings, 3) the relevance of the study, and 4) the risk of bias (systematic error) and study limitations (internal validity) of the study. Here we focus on the evaluation of risk of bias of an individul study as a distinctly important quality assessment of studies of medical test performance.

Key points are:

- Reviewers should select criteria that assess the risk of systematic bias when assessing study limitations that are particularly relevant to the test under evaluation. A comprehensive list of important sources of systematic biases is provided.
- Categorizing individual studies as "good," "fair," or "poor" with respect to quality (risk of bias) is a useful way to proceed.
- Methods for determining an overall categorization for the study limitations should be established a priori and documented clearly.
- We recommend separately evaluating the size and direction of findings in the context of determining the strength of a body of evidence (*see* Paper 7), and evaluating the relevance of the study in the context of assessing applicability (*see* Paper 6).

References

1. Whiting P, Rutjes AW, 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.

2. Agency for Healthcare Research and Quality. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville, MD: Agency for Healthcare Research and Quality. Available at: http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-

guides-reviews-and-reports/?pageaction=displayproduct&productid=318. Accessed September 20, 2010.

3. Whiting P, Rutjes AW, 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.

4. 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.

5. Leeflang MM, Deeks JJ, Gatsonis C, et al. Systematic reviews of diagnostic test accuracy. Ann Intern Med 2008;149:889-97.

6. Centre for Reviews and Dissemination. Systematic Reviews: CRD's Guidance for Undertaking Reviews in Health Care. Centre for Reviews and Dissemination: York, UK; 2009.

7. Whiting P, Rutjes AW, Reitsma JB, et al. 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.

8. 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.

9. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Br Med J 2003;326:41-44.

10. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Clin Chem 2003;49:1-6.

11. Little J, Higgins JP, Ioannidis JP, et al. STrengthening the REporting of Genetic Association Studies (STREGA): an extension of the STROBE statement. PLoS Medicine / Public Library of Science 2009;6(2):e22.

12. 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. J Clin Epidemiol 2008;61(4):344-9.

13. Higgins JPT, Altman DG, 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 of Systematic Reviews of Intervention. Version 5.0.1 (updated September 2008). The Cochrane Collaboration, 2008. Available from www.cochrane-handbook.org. Accessed September 3, 2009.

14. Qureshi N, Wilson B, Santaguida P, et al. NIH State-of-the-Science Conference: Family History and Improving Health. Evidence Report/Technology Assessment No. 186. (Prepared by the McMaster University Evidence-based Practice Center, under Contract No. 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 October 5, 2010.

15. 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.

Methods Guide for Medical Test reviews

Paper 6

Assessing Applicability

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

AHRQ Publication No. xx-EHCxxx-EF
<Month Year>

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, through a collaborative effort within the Evidence-based Practice Center Program, have developed a Methods Guide for Medical Test Reviews. We intend that these documents will serve as a resource for our EPCs as well as for other investigators interested in conducting systematic reviews on medical tests.

This Medical Test Methods guide is intended to be a practical guide for those who prepare and use systematic reviews on medical tests. This document complements the EPC Methods Guide on Comparative Effectiveness Reviews (http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318), which focuses on methods to assess the effectiveness of treatments and interventions. The guidance here applies the same principles for assessing treatments to the issues and challenges in assessing medical tests and 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. We provide guidance in stepwise fashion for those conducting a systematic review.

The *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 *Medical Test Methods Guide* and the Effective Health Care Program can be made at www.effectivehealthcare.ahrq.gov.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

Paper 6. Assessing Applicability

Reviews of medical test studies, like most systematic reviews, are conducted for a practical purpose: to support clinicians, patients, and policymakers—decisionmakers—in making informed decisions. To make informed decisions, decisionmakers need not merely to understand whether a particular test strategy is worthwhile in some context, but whether it is worthwhile in the specific context—patient and disease characteristics, downstream management options, setting, and so on—relevant to their particular decisions. Is this test robust over a wide range of patient types and scenarios of use, or is it relevant only to a narrow set of circumstances?

As systematic reviewers, we approach this concern of decisonmakers about the applicability of *tests* to *contexts* by focusing our attention on the applicability of *evidence about tests* to one or more *key questions,* in which context is (or should be) stipulated (*see* Paper 2). To the extent that available evidence is applicable to a particular context of interest, and the evidence is supportive of the use of the test in that context, then the test is also applicable to that context.

AHRQ's *General Methods Guide*[1] describes four steps for assessing and reporting applicability of individual studies and of a body of evidence. These steps are relevant for evaluating medical tests, but with some important considerations, which we highlight here.

As with assessing interventions, it is useful to distinguish the assessment of applicability, or external validity, from that of quality, often considered synonymous with internal validity.[1] Consider, as is often the case, that the accuracy of a medical test is highly sensitive to the severity of disease—that is, there is a spectrum *effect*.[2] Imagine a study that purports to address the accuracy of a particular test for sepsis in a general population of individuals with abnormal white blood cell count, but in which only extreme subjects with white blood cell counts higher than twice normal are chosen for study. In this study, the selection of subjects introduces a spectrum *bias*, leading to estimates of operating characteristics that are likely to have low external validity, or low applicability to the "general" population of individuals with abnormal white blood cell counts;[3] in many reviews, such studies may be deemed to be of low "quality" (*see* Paper 5). However, in reality, the assessment of the test may be valid for a particular context, and the task at hand is to describe the context so that the utility of the test is clearly understandable for the end-users. As recommended when assessing applicability for systematic reviews on interventions,[1] applicability should be reported separately from quality or strength of evidence, and should not be assessed by a universal rating scheme.

Applicability is relevant to all elements in the causal chain in which test use is purportedly linked to health outcomes. We may ask, as for stress treadmill testing, "Is this test equally accurate for diagnosing coronary heart disease in women as in men?," or "How often does a thyrotropin-stimulating hormone test identify a thyroid condition that needs treatment among individuals with a complaint of fatigue across the age spectrum?" We may also ask about the applicability of studies that speak to more distal steps in the causal chain, such as, "Does fetal fibronectin testing of women with preterm labor symptoms (versus in routine clinical care) change the care they receive?" Or, for an appropriate clinical study, we may ask about its applicability to an overarching question, such as, "Does use of PET scanning for individuals with mild cognitive impairment lead to better survival or quality of life?" In this document, we focus on assessing applicability of studies of test accuracy since these are the outcomes that are most unique to medical tests; however, the principles are relevant to all links of the causal chain, as well as to an overarching question about the value of testing on overall health outcomes.

Common Challenges

Evaluating the applicability of a body of evidence regarding medical tests presents several common challenges to reviewers.

Lack of Clarity in Key Questions

Early formulations of key questions may not provide clear context for determining the applicability of a study. For example, for test accuracy questions, not all potentially relevant contextual factors are stipulated in the key questions. This complicates further decisions in the review process. First, which studies should be included and which excluded? The issue typically arises either because the study evaluated a context different from that described in the key question, or because the study failed to provide sufficient information to assess applicability. The reviewer is faced with deciding when these deviations from ideal are minor, and when they are more crucial and are likely to affect test performance, clinical decisionmaking, and health outcomes in some significant way.

Another point in the review process where lack of clarity in key questions affects assessments of applicability is in the presentation, analysis, and summarization of data. What study features should be included in evidence tables? How should aggregate tables and meta-analyses approach studies that describe tests in somewhat different contexts? The challenge of when to “lump” and when to “split” is evident in the complexity of the evidence concerning lipid and triglyceride screening for prevention of cardiovascular disease. While it is possible to compile data relating cholesterol subtypes and total levels to health risks, we know clinically that they are interrelated and that effects may vary with age and gender. The reviewer must consider how to organize the data by important clinical or policy contexts so as to maximize the relevance of the findings to decisionmakers.

Review Covers a Wide Range of Contexts

A second common challenge faced by reviewers is that the key questions relate to the wide range of potential applications of one or more tests singly or in various combinations, and across multiple populations and settings. In addition to the problems noted above, when the review is intended to be broad, it is tempting to “let the evidence speak” by, for example, examining test accuracy as a function of the numerous possibly relevant contextual factors. The problem with this approach is that it introduces a potential bias in that some factors that appear to influence test accuracy or subsequent outcomes may actually be spurious. This is similar to the challenge of evaluating efficacy of treatments in subgroups and is addressed in the *General Methods Guide*.[1]

Tests are Rapidly Evolving

A third major challenge to assessing applicability especially relevant to medical tests is that, even more that treatments, tests are often changing rapidly, both in degree (enhancements in existing technologies) and type (incorporate substantively new technologies). The literature often contains evidence about tests that are not yet broadly available or are no longer common in clinical use. Secular 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 that it will produce a new version that provides only a continuous, quantitative measure. In this situation, reviewers must weigh how best to capture data relating the two versions of the test and decide whether there is merit 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.

Principles for Addressing the Challenges

The root cause of the challenges above 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, and analysis/summarization. In applying the principles described below, the PICOTS typology can be useful (*see* Papers 1 and 2).

Principle 1: Identify Important Contextual Factors

In an ideal review, all possible factors related to impact of a test use on health outcomes should be considered. However, this is usually not practical, and so some tractable list of factors must be considered before initiating a detailed review. First, consider factors that may affect the step in the causal chain of direct relvance to the key question (e.g., for assessing accuracy of cardiac MRI for atherosclerosis, slice thickness is a relevant factor in assessing applicability). However, even if the key question relates only to test accuracy, it is also important to consider factors that may 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 Principle 1, consider contextual issues that are especially relevant to tests. Factors that may be less obvious but of significant importance in the interpretation of medical tests are time effects and secular trends. Factors that may be especially relevant are listed below.

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, and 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, for example, would need to be examined by gestational dates and varied cut-offs for those stages in gestation. Awareness of these changes guides review parameters such as date range selection and eligible test type for the included literature helps frame categorization and discussion of the review 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 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 in general use, to change modestly over time.

Secular trends in population characteristics can have indirect effects on applicability when the population characteristic changes 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 criteria to diagnose acquired immune deficiency syndrome (AIDS), the transition from the cystometrically defined condition detrusor instability or overactivity to the symptom complex “overactive bladder,” and the continuous refinement of classifications of mental health conditions recorded in the *Diagnostic and Statistical Manual* updates. If the condition being tested for changes criteria, 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. Preventive care is an important component of understanding current practice to consider 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 is 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 (*see* Papers 1 and 2) to identify the range of possible factors that might affect applicability and consider the hidden sources of limitations noted above. Second, review the list of applicability factors with stakeholders to ensure common vantage points and 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 the process of designing the process and abstracting data for an evidence review.

Core characteristics for assessing and describing the applicability of medical tests are organized by the elements of PICOTS in Table 6-1. This often requires generous use of the "NR" entry ("not reported") in evidence and summary tables.

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

PICOTS element	Core characteristics to document	Challenges	Example	Potential systematic approaches
Population	Method of identification/selection Inclusion & exclusion criteria Demographic characteristics of those included Prevalence of condition Spectrum of disease detected	Source of population not described Study population poorly specified Key characteristics not reported	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 case 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 How conducted By whom Cut-off/diagnostic thresholds applied Skill of assesors when interpretation of test required	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 critical and not assessed Or include but: – Contact authors for clarification – Flag version of test or deficits in reporting in

PICOTS element	**Core characteristics to document**	**Challenges**	**Example**	**Potential systematic approaches**
				tables/text – Discuss implications – Model cut-offs and conduct sensitivity analyses
Comparator	Gold standard vs "alloy" standard Alternate or "usual" test 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
Outcome of use of the test	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	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	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 Place in the sequence of care Timing of assessment of disease status and outcomes	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

PICOTS element	**Core characteristics to document**	**Challenges**	**Example**	**Potential systematic approaches**
				– Discuss implications – Note need for challenge to be addressed in future research
Setting	Primary care vs. specialty care Hospital-based Routine processing vs. specialized lab or facility Specialized personnel	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

Abbreviations: CT = computed tomography; DVT = deep venous thromboembolism.

Principle 2: Be Prepared to Deal With Additional Factors Affecting Applicability

Despite best efforts, some contextual factors that appear to be 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 was 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 rises only to the level of a hypothesis.

Principle 3: Restriction of Scope may be Appropriate

An important decision is how to deal with studies that do not directly apply to the context described in the key questions, or when important details that allow the reviewers to assess applicability are missing. For instance, if the goal for a review is to synthesize the literature on the use of individual risk prediction tools for estimating risk of myocardial infarction or stroke among diabetics, the reviewer is faced with the decision whether to systematically exclude studies that were not limited to people with diabetes, or those that did not include some proportion of diabetics, or those that included diabetics but did not publish results stratified by the presence or absence of diabetes. This decision is particularly challenging when restriction leads to a scant body of evidence.

In general, if the review is intended to apply to 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), then excluding studies is appropriate if they (a) fail to address that characteristic of interest, (b) do not include relevant individuals or settings, (c) do not stratify results by the feature of interest, or (d) do not include analysis of effect measure modification by that characteristic. 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 the review 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 exclusion based on consensus, and inclusion but 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. 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.

Illustrations

Our work on the 2002 Cervical Cancer Screening Summary of the Evidence for the US Preventive Services Task Force[4] illustrates several of the challenges in determining applicability: the literature included many studies that did not use gold standards or testing of normals; many studies could not relate cytologic results to final histopathologic status; and few data were available from the target "use case" populations. The evidence review also encountered significant examples of changes in secular trends and availability and format of medical tests.

When the update[4] was intitiated, much had changed since the prior 1996 review: liquid-based cervical cytology was making rapid inroads into practice; resources for reviewing conventional Pap smear tests were under strain from a relative shortage of cytotechnologists in the workforce and restrictions on the volume of slides they could read each day; several new technologies had entered the market 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 the use of HPV testing to enhance accuracy or to triage which indiviudals needed evaluation with colposcopy and biopsies to evaluate for cervical dyplasia and cancer. No randomized controlled trials (RCTs) were available using, comparing, or adding new tests or technologies to prior conventional care. Some USPSTF members were interested in teens and younger women; however, in topic development, the spot searches and expert consensus was that the literature was insufficient to support a specific focus on the subgroup of women in their teens and twenties. Key questions related to stopping screening focused on the subgroups of older women and those after hysterectomy. We chose an exhaustive approach to describe the broadest possible picture of the state-of-the-science in cervical cancer screening, emphasizing the changes outlined above. We also sought to gather data for indirect comparisons among techniques, as well as examining outcomes of screening in two specific subgroups—older women and those who have had hysterectomy.

Because no data were available comparing the effects of new screening tools or strategies on outcomes, the report focused on medical test characteristics, 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. In fact, one of the companies involved had purchased the intellectual property rights of another with the plan to shelve the competitor's technology and reduce competition. This was publicly available knowledge that might have simplified the work and the findings. 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 confused users of the report. Only in retrospect after the decision to include all test was made were we able to see that this approach did not substantially add to understanding the findings because diagnostic characteristics of those tests that were not longer available or available only in restricted settings were not meaningfully superior..

We restricted *a priori* on these requirements: the study reported on tests obtained for screening; results were compared with a colposcopy and/or histology reference standard; the reference standard was assessed within 3 months of screening test, and data from the publication allowed completion of a 2-by-2 table, preferably for each level of dyplasia and presence of cancer. Overwhelmingly, failure to meet criteria resulted from failure to use a reference standard. Most excluded studies used split-sample correlations. Thus, there was little to report beyond the lack of high quality evidence to guide care.

Although clearly describing the dearth of information available to inform decisions, the review was not able to provide more information for care providers and women about expected usefulness of the tests. As a means of remediation, not planned in advance, we used prior USPSTF meta-analysis data on conventional Pap medical test performance,[5] along with the one included paper about liquid cytology,[6] to model the potential implications of its use; specifically, we explored the issue of overburdening care systems with detection of low-grade dysplasia while not substantively enhancing detection of severe disease or cancer.[7] The projections from the report have since been validated in prospective studies. In the time since the 2002 USPSTF Cervical Cancer Screening update,[4] liquid cytology has become the most widely used form of Pap testing, adopted by more than 80 percent of practices within 5 years of its introduction.

On the topic of HPV testing for the virus that causes dysplasia, the we identified 65 articles, of which 13 met the inclusion criteria. We compiled meta-estimates of the test performance characteristics including likelihood ratios and examples of sequential use of tests to demonstrate the potential influence of HPV testing in varied scenarios for underlying disease prevlance. Our intention was that these models would approximate expected outcomes across a wider range of settings than was represented in the literature, enhancing the usefulness of findings and improving applicability. In our desire to be exhaustive, however, we reviewed tests of single or combinations of virus types and methods of testing (polymerase chair reaction) that did not reflect tests likely to be used outside of specialized research settings. The data were confusing since some data understated performance compared with tools that were available and some overstated potential benefits by including highly sensitive tests not available for clinical use. Although the review discussed these nuances, we now recognize the report might have been clearer and would have been far more efficiently conducted had the team reviewed only those tests with current or likely near-term clinical availability, which could have been determined on the basis of patent and FDA regulatory decisions that indicated what products would be available and been designed for use in clinical care.

For two of the subgroup analyses (age and hysterectomy), it was necessary to include publications with information about underlying incidence and prevalence in order to provide context, as well as to drive modeling efforts. These data helped improve understanding 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. Rather than restrict the literature, it was preferable not to use more restricted search terms in order to ensure that the most applicable evidence was found for the two target groups, even inside of larger studies.

As this illustration of challenges highlights, applicability of a report can be well served by restricting inclusion of marginally related or outdated studies and is rarely enhanced by uncritically extrapolating the 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 use of restrictions will benefit from seeking explicit guidance from clinical, medical testing, and statistical experts about applicability challenges. Review teams need to familiarize themselves with the availability and contemporary clinical use of the test, 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.

However, often the target of interest is 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—to be exhaustive or to narrow the eligible literature—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.

Summary

Key points are:

- Applicability is the ability of an individual study or a body of evidence to provide information to guide real-world decisions about the likely performance characteristics or outcomes of a particular medical test for a particular setting.
- Systematic reviews typically face one or more of the following challenges: (1) publications in the literature report insufficient detail to permit assessment of key elements of applicability; (2) the relevant literature is large and needs to be grouped or organized in ways that help convey applicability and that anticipate differences in test performance or outcomes across groups; and (3) medical tests are rapidly changing and the literature may reflect tests that are not yet available, or that are no longer in use.
- Using a PICOTS checklist approach to develop an inventory of applicability factors is key for designing and scoping reviews, summarizing large bodies of evidence and their relevance to smaller subgroups, and identifying potentially concerning trends over time that may influence interpretation of the literature.
- It is necessary to be prepared for the uncovering of relevant contextual factors after a substantial volume of literature has been reviewed. It is crucial to recognize that like any relationship discovered *a posteriori*, this may reflect a spurious association. 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 rises only to the level of a hypothesis.
- When the intended context and population for whom applicability is of interest are clear and focused in advance, EPCs should shape the review at the level of tightly focused inclusion and exclusion criteria. We also recommend that EPCs carefully weigh the implications of reviewing data for older and possibly outmoded test methods. The temptation to be exhaustive can dilute the focus.
- If a review is intended to be exhaustive, EPCs can use the factors that influence applicability to provide a structure within key question results and discussion for organizing findings. It may be useful to organize summary tables by key factors, for example, grouping data by sex when possible and indicating where this was not possible, or grouping by version of the test when there is more than one method in the literature.

- The importance of contextual factors that may affect applicability can be evaluated using evidence synthesis methods, such as meta-regression and decision modeling (*see* Papers 8 and 10).

References

1. Agency for Healthcare Research and Quality. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville, MD: Agency for Healthcare Research and Quality. Available at: http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318. Accessed September 20, 2010.

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

3. 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. p. 80-90.

4. Hartmann KE, Hall SA, Nanda K, et al. Cervical Cancer Screening: A Summary of the Evidence. (Prepared for the US Preventive Services Task Force under Contract No. 290-97-0011). AHRQ Publication No. 03-515A. Rockville (MD) Agency for Healthcare Research and Quality, January 2002. Available at: http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=es25. Accessed on October 13, 2010.

5. McCrory DC, Matchar DB, Bastian L, et al. Evaluation of Cervical Cytology. Evidence Report/Technology Assessment No. 5. (Prepared by Duke University under Contract No. 290-97-0014.) AHCPR Publication No. 99-E010. Rockville, MD: Agency for Health Care Policy and Research. February 1999. Avaliable at: http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=erta5. Accessed on October 13, 2010. 1999.

6. 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.

7. Hartmann KE, Nanda K, Hall S, et al. Technologic advances for evaluation of cervical cytology: is newer better? Obstet Gynecol Surv 2001;56(12):765-74.

Methods Guide for Medical Test reviews

Paper 7

Grading the Strength of a Body of Evidence

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

AHRQ Publication No. xx-EHCxxx-EF
<Month Year>

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, through a collaborative effort within the Evidence-based Practice Center Program, have developed a Methods Guide for Medical Test Reviews. We intend that these documents will serve as a resource for our EPCs as well as for other investigators interested in conducting systematic reviews on medical tests.

This Medical Test Methods guide is intended to be a practical guide for those who prepare and use systematic reviews on medical tests. This document complements the EPC Methods Guide on Comparative Effectiveness Reviews (http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318), which focuses on methods to assess the effectiveness of treatments and interventions. The guidance here applies the same principles for assessing treatments to the issues and challenges in assessing medical tests and 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. We provide guidance in stepwise fashion for those conducting a systematic review.

The *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 *Medical Test Methods Guide* and the Effective Health Care Program can be made at www.effectivehealthcare.ahrq.gov.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

Paper 7. Grading the Strength of a Body of Evidence

"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 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 they might make 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 grades.

Guidance to Evidence-base Practice Centers (EPCs) on assessing the strength of a body of evidence when comparing medical interventions has been described in AHRQ's *General Methods Guide*.[1,3] That guidance is based on the principles identified by the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) working group[4-5] with minor adaptations. When assessing the strength of evidence, systematic reviewers should consider four "required" domains—risk of bias, consistency, directness, and precision[5]—as well as the additional domains of publication bias, dose-response association, existence of plausible unmeasured confounders, and strength of association (i.e., magnitude of effect).

EPCs grade the strength of evidence for each key question addressed in a systematic review. The process of defining the important intermediate and clinical outcomes of interest for medical tests is further described in Paper 3. Because most medical 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 medical test is studied in the context of a clinical trial (e.g., test versus do not test) with clinical outcomes as the study endpoint, the reader is referred to 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.

In this paper, we outline the particular challenges that systematic reviewers face in in grading the strength of a body of evidence on medical test performance; we then propose principles for addressing these challenges.

Common Challenges

Medical test studies commonly focus on test performance, 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 medical tests,[6-10] we identified common challenges that reviewers face when assessing the strength of a body of evidence on medical 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 medical test, where evidence is 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 medical test (*see* Paper 2), and it is the reason we must assess the strength of the body of evidence for each link in the chain. The quality of the body of evidence regarding the overarching question of whether a test will improve clinical outcomes depends on the quality of the body of evidence for the weakest link in this chain.

A second challenge related to the indirect nature of medical test evidence is that the assessment of the quality of a body of evidence for one key question may be affected by other issues in the evidence chain. For example, how we judge a test's performance for the presence of a particular diagnosis may be affected by disease prevalence and downstream treatment effects, including adverse effects. Consider the precision of estimates of test performance in terms of confidence intervals. 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 may not necessarily translate into a clinically meaningful impact. 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.[8]

Table 7-1. Example of the impact of precision or imprecision of sensitivity on negative predictive value

Type of biopsy	**Postbiopsy probability of having cancer after a negative core-needle biopsy result[a]**			
	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%

[a] 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.

Principles for Addressing the Challenges

Principle 1: GRADE-Required Domains can be Adapted to Assess a Body of Evidence on Medical Test Performance

To assess a body of evidence related to medical test performance, we can adapt the GRADE-required domains of risk of bias, consistency, directness, and precision. 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 medical test performance and strategies for assessing the risk of systematic error in such studies, are discussed in detail in Paper 5.

Consistency concerns homogeneity in direction and magnitude of results across different studies. The concept can be similarly applied to medical 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 medical 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. Spread of data points on the ROC curve is one method of assessing consistency of diagnostic accuracy among studies. As with intervention studies, residual unexplained heterogeneity—not explained by different study designs, methodologic quality of studies, diversity in subject characteristics, or study context—should reduce the strength of a body of evidence.

Directness, according to AHRQ's *General Methods Guide*, concerns whether the evidence being assessed "reflects a single, direct link between the interventions of interest [medical tests] and the ultimate health outcome under consideration."[1] In the case where a key question concerns the performance of a medical test, there are unlikely to be any intermediate outcomes that would reduce the directness from the test being evaluated to the accuracy outcome. While the systematic review may conclude direct evidence for accuracy outcomes, the burden then shifts to the decision-makers to consider how the accuracy outcome relates to important clinical outcomes. Directness also applies to comparing interventions. In the case of medical tests, it is important to consider whether the tests in a study are being used in a similar way as they are used in practice. For example, a study may compare the use of d-dimer test accuracy compared to venous ultrasound for venous thromboembolism, but the comparison of interest may actually be the use of the d-dimer test as a triage for venous ultrasound.

Precision refers to the width of confidence intervals for diagnostic accuracy measurements and is integrally related to sample size.[1] Difficulties arise when a test may have narrow confidence intervals for one outcome (e.g., true negatives) but not for another outcome (e.g., true positives). Grading of precision should be based on more than one measure (good precision if all important measures have reasonable precision, fair if some but not all have reasonable precision, or poor if none has reasonable precision).

Before assessing the precision of diagnostic accuracy, reviewers should consider how wide confidence intervals for one measure of accuracy may translate into 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, above, or it may require more formal analysis as with a decision model (*see* Paper 10). If the impact of imprecision on clinical outcomes is negligible, the grade for precision should not be downgraded.

Principle 2: Additional Domains can be Adapted to Assess a Body of Evidence on Medical Test Performance

Under certain circumstances, additional domains—such as publication bias, dose-response association, existence of plausible unmeasured confounders, and strength of association—can also be adapted to assess a body of evidence related to medical test accuracy, as shown in Table 7-2.

Table 7-2. Additional domains and their definitions (adapted from the *General Methods Manual*)[1,3]

Domain	Definition and elements	Application to evaluation of medical test performance evidence
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 assymetry in funnel plots should raise suspicion for publication bias.	Publication bias can influence ratings of consistency, precision, 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 (dose, duration, adherence).	The dose-response association may support an underlying mechanism of detection and potential relevance to some tests that have continuous outcomes and possibly multiple cutoffs (e.g., gene expression, serum PSA levels, and ventilation/perfusion scanning).
Plausible unmeasured confounding that would decrease observed effect	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 been even larger. In such a case, an EPC may wish to upgrade the level of evidence.	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.

Abbreviations: EPC = Evidence-based Practice Center; PSA = prostate-specific antigen.

Principle 3: Methods for Grading Intervention Studies can be can be Adapted for Studies Evaluating Broader Medical Test Outcomes

A body of evidence evaluating broader medical 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 intervention outcomes. Grading issues in this type of medical 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, application of the grading criteria should not signifcantly differ from the methods used for intervention evidence.

An unresolved issue remains 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 surrogate 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 upon the benefits and harms of available treatments. The benefits or harms of accuracy outcomes such as true negatives or false negatives depend on the cognitive or emotional outcomes resulting from the knowledge itself, as outlined in Paper 3.

Currently there is insufficient evidence to suggest how very indirect evidence, such as when only accuracy outcome studies are available, should be reflected in the overall grade assigned to a body of evidence. Some have suggested that there are cases in which accuracy outcomes may be sufficient to conclude that there is or is not a benefit on clinical outcomes,[11] as discussed in Paper 2. At this point, we recommend that EPCs assess direct evidence, when available, using the established GRADE criteria. When only indirect evidence on surrogate 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 4: Multiple Domains Should be Incorporated Into an Overall Assessment in a Transparent Way

The overall strength of evidence grades reflect a global assessment of the required domains and any additional domains as needed into an overall summary grade—high, moderate, low, or insufficient. 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. The GRADE approach weights various required domains to arrive at a summary score and thus advocates a quantitative approach. 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 EPCs 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

Consider the following illustration, which is by no means prescriptive but highlights some of the challenges and improvements in grading a body of medical test evidence. An AHRQ-sponsored systematic review addressed the key question of whether testing for the Factor V Leiden (FVL) mutation, alone and in combination with the prothrombin G202102A mutation, leads to improved clinical outcomes in adults with a personal history of venous thromboembolism (VTE) or in adult family members of mutation-positive individuals (Key Question 1).[9] Additional key questions included the analytic validity, clinical validity, and clinical utility of testing for the FVL mutation alone or in combination with prothrombin G202102A testing. After developing an analytic framework and performing a quality assessment of individual studies, the reviewers graded the strength of evidence for each key question. They focused on whether testing improved clinical outcomes (an overarching issue) and then evaluated the analytic validity, clinical validity, and clinical utility of these medical tests.

The review authors assessed aggregate risk of bias based on the number of studies with the strongest design and the quality of the best available evidence, including evaluation of the limitations affecting individual study quality. They also evaluated the remaining mandatory domains and some additional domains; these included the certainty regarding the directness of the observed effects in the studies, the consistency of the evidence, the precision and strength of association observed (measures of relative risk greater than 2 or less than 0.5 were considered as strong evidence of association), and the possibility of publication bias and the selective reporting of outcomes.[9] The reviewers considered specific domains and additional domains, but some were considered more relevant than others for specific key questions. Lack of directness was weighted more heavily for the question on clinical outcomes, whereas strength of association was weighted more heavily for the questions on predictive ability. The team found no direct evidence that addressed the primary objective; hence, they graded the evidence as “insufficient” for the overarching key question on clinical outcomes.

On another key question regarding whether the heterozygous presence of the FVL mutation alone predicts the risk of recurrent VTE in individuals (probands) who have had VTE, the team graded the strength of evidence as “moderate” because of inconsistency of the results among included studies and the lack of directness as some of these studies had not been designed to address the question directly. Thus, directness and consistency were the domains that primarily resulted in a summary grade of “moderate” on this key question.

By contrast, because most of the studies used a solid reference standard, and the test performance characteristics, including sensitivity and specificity, were excellent, the authors graded the strength of evidence about analytic validity as “high.”[9] Although not clarified in the review, further discussion with the authors revealed that the consistency and directness of the evidence were the domains that resulted in this summary grade of “high.”

The reviewers also concluded that the evidence that homozygosity for FVL in family members predicted VTE (clinical validity) was “high.” This summary grade was primarily driven by the domain of consistency and the additional domain of strength of association (odds ratios > 10) across studies.

The evidence was graded “low” on the key question of whether patient management by physicians may change on the basis of the results of testing for FVL or prothrombin G20210A and improve VTE-related outcomes in individuals who have had VTE or in the probands’ family members who have been tested (clinical utility). This summary grade was primarily driven by the lack of direct evidence.

Summary

Grading the strength of a body of medical test evidence involves challenges over and above those related to grading the evidence from health care 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—medical test performance—and described the challenges involved in assessing other links in the chain, and the relationship between assessing one link and another, less fully.

No one system for grading the strength of evidence for diagnostic medical 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 on medical tests more useful for decisionmakers.

Key points:

- Reviewers grading the strength of a body of evidence on medical tests should consider the domains of risk of bias, directness, consistency, and precision, which are also routinely used to grade evidence on non-test interventions.
- Given that most evidence regarding the clinical value of medical tests is indirect, it is essential that an analytic framework be developed to clarify the key questions; the strength of evidence for each link in that framework (i.e., corresponding to each key question) should be graded separately (e.g., for test performance alone).
- 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 weighted the most 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. Available at: http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-

guides-reviews-and-reports/?pageaction=displayproduct&productid=318. Accessed September 20, 2010.

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. 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 July 1, 2010.

7. 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: www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=erta1. Accessed July 1, 2010.

8. 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 July 1, 2010.

9. 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 July 2, 2010.

10. West S, King V, Carey TS, et al. Systems to Rate the Strength of Scientific Evidence. Evidence Report/Technology Assessment No. 47 (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.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=erta47. Accessed July 2, 2010.

11. 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.

Paper 8

Meta-analysis of Test Performance Evidence When There is a “Gold Standard” Reference Standard

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

AHRQ Publication No. xx-EHCxxx-EF
<Month Year>

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, through a collaborative effort within the Evidence-based Practice Center Program, have developed a Methods Guide for Medical Test Reviews. We intend that these documents will serve as a resource for our EPCs as well as for other investigators interested in conducting systematic reviews on medical tests.

This Medical Test Methods guide is intended to be a practical guide for those who prepare and use systematic reviews on medical tests. This document complements the EPC Methods Guide on Comparative Effectiveness Reviews (http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318), which focuses on methods to assess the effectiveness of treatments and interventions. The guidance here applies the same principles for assessing treatments to the issues and challenges in assessing medical tests and 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. We provide guidance in stepwise fashion for those conducting a systematic review.

The *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 *Medical Test Methods Guide* and the Effective Health Care Program can be made at www.effectivehealthcare.ahrq.gov.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

Paper 8. Meta-analysis of Test Performance Evidence When There is a "Gold Standard" Reference Standard

Meta-analysis and related methodologies are an important part of systematic reviews of medical tests. Quantitative analyses are not required for a systematic review; several very informative and well-conducted reviews do not include any such analyses. However, when possible, systematic reviewers appropriately strive to perform meaningful quantitative analyses to provide summary estimates for key quantities, or to explore and explain observed heterogeneity in the results of identified studies.

Syntheses of medical test data tend to focus on test performance; that is, on the ability to discern the presence of a particular condition or level of risk ("accuracy" studies). Thus, much of the attention on statistical issues relevant to synthesizing medical test evidence focuses on summarizing test performance data. It is important to be aware that key clinical questions driving an evidence synthesis (e.g., Is this test alone, or some test/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 papers in this *Medical Test Methods Guide* (*see* Papers 2, 6, and 10). For our purposes, we assume that the primary goal of synthesis is to combine and interpret test performance data.

It is beyond the scope of this *Medical Test Methods Guide* to provide a detailed description of statistical methodologies for meta-analysis of medical test performance and the assumptions they invoke, or to discuss the practical aspects of their implementation. 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, as well as with the common measures of medical test performance. For example, we do not review challenges posed by methodological or reporting shortcomings of test performance studies.[1] The Standards for Reporting of Diagnostic Accuracy (STARD) initiative published a 25-item checklist that aims to improve reporting of medical test studies.[2] We refer readers to Paper 5 of this *Medical Test Methods AnGuide* and to several methodological and empirical explorations that discuss the effects of bias and variation on the performance of medical tests.[3-5]

Common Challenges

Summarizing medical test performance data in general, or for a subgroup of patients, is complicated because medical test performance studies can differ on the definition of test positivity, may apply varying criteria for "truth," and often involve heterogeneous patient populations. This presents two common challenges.

First, the simplest challenge is summarizing studies of test performance when the reported sensitivities and specificities are not variable (non-heterogeneous) across studies. Non-heterogeneity can appear when there is no explicit or implicit variation in the threshold for a "positive" test result, as well as for other reasons.

A second challenge is summarizing test performance when there is substantial variation (heterogeneity) in sensitivity or specificity estimates across studies. This often-encountered situation can be secondary to explicit or implicit variation in the threshold for a “positive” test result; heterogeneity in populations, reference standards, or index tests; study design; chance; or bias. For example, when the test threshold changes, it is expected that sensitivity and specificity will also change, and in opposite directions.[a]

Principles for Addressing the Challenges

In this paper, we describe our approach to summarizing commonly encountered types of diagnostic data by addressing the two challenges mentioned above: (1) The case of non-heterogeneous sensitivity and specificity, which can be encountered when there is no variation in the threshold of positive tests; and (2) the case of heterogeneous sensitivity or specificity, which can appear when there is variation in the use and reporting of thresholds, among other things. We also discuss how to assess and explore heterogeneity using illustrative examples. We propose that reviewers follow two general principles.

Principle 1: Favor the Simplest Analysis That Properly Summarizes the Data

Meta-analysis of test performance data may require more complex methods than are needed for meta-analyses of intervention studies. This is because there are many metrics that describe test performance from different points of view, and one has to choose among these metrics. For most analyses, we resort to modeling sensitivity and specificity rather than other metrics, as will be discussed below. Furthermore, test performance is most often specified as a two-dimensional problem; that is, it consists of an analysis of sensitivity and specificity, or of quantities derived from them. A large number of analytical options are available to the meta-analyst, some more suitable to certain situations than others. We will discuss a general approach for choosing among the available options, which should of course be critically applied to the topic at hand.

Principle 2: Explore Any Variability in Study Results With Graphs and Suitable Analyses

One of the most important uses of meta-analysis is to quantify and explore reasons for between-study heterogeneity. Meta-regression-based analyses and subgroup analyses are naturally amenable to this purpose.

One of several possible approaches to meta-analysis of diagnostic data is presented in Box 8-1 and expanded in the section that follows.[6]

[a] Changes in the threshold of a test can be explicit when, for example, a cutoff value across a continuum of measurements is employed; or changes can be implicit when, for example, tests have a more qualitative interpretation, such as many imaging tests (e.g., chest radiography used for screening versus confirmation of tuberculosis). Notably, this is a frequent source of heterogeneity for test performance studies, and so the issue of coping with heterogeneity will be addressed here.

Box 8-1. Algorithm for meta-analysis of diagnostic studies when there is a “gold standard”

Step 1: Decide whether separate meta-analyses of sensitivity and specificity are suitable. This is true if there is no evidence of variability in the sensitivity or specificity across studies (e.g., by observing forest plots, or with standard heterogeneity assessment methods).[a] This can happen when all studies use the same explicit threshold for positive tests, but also in other cases.

Step 2: If there is substantial variation in sensitivity or specificity, consider methods that analyze them jointly. This can be the case when studies use different explicit or implicit thresholds for positive tests. Fit the bivariate model (see text). If there is evidence that the correlation between sensitivity and specificity is

a. *Negative** (i.e., as expected), present:
    i. Summary point estimates of sensitivity and specificity[a]
    ii. Summary receiver operating characteristic curve (SROC)
    iii. Both of the above
b. *Positive*** (i.e., in the opposite than expected direction), then for this collection of studies consider summarizing as per Step 1.
c. *Zero*, then the results of the bivariate model would be similar to the separate meta-analyses of Step 1.

Step 3: If more than one threshold is reported per study, this has to be taken into account in the quantitative analyses. Tentatively, we encourage both qualitative analysis via graphs and quantitative analyses via proper methods.

Step 4: Explore the impact of study characteristics on summary results in the context of the primary methodology used to summarize studies (Steps 1, 2, or 3):

a. Meta-regression-based analyses
b. Subgroup analyses

* Based on the posterior distribution of the correlation parameter, or perhaps based on other, external knowledge about the test.

** It is not uncommon to have a positive median or mean in the posterior distribution of the correlation. This is contrary to what is expected, and may be the result of confounding (omission of an important covariate).

Step 1: Decide whether separate meta-analyses of sensitivity and specificity are suitable. When there is no obvious variability in the sensitivity and specificity across studies, then there is no use for more advanced methods, which were specifically developed to take into account how sensitivity and specificity are related across their range. Among other cases, this situation can arise when all studies use the same explicit or implicit threshold for positive tests.

Most measures of test performance (with the exception of positive and negative predictive value) can be quantitatively combined using standard methodologies discussed in AHRQ’s *General Methods Guide*.[7] Specifically, for the different measures:

Sensitivity and specificity. Sensitivity and specificity could be combined separately in the absence of variation in diagnostic thresholds when there is not substantial between-study heterogeneity. The naive approach of summing over all true positives (negatives) and diseased (non-diseased) subjects across studies and calculating “pooled” sensitivity (specificity) from the ratio of the corresponding sums should be avoided. This approach does not take into account between-study variability, or the potential bias caused by unbalanced ratios between diseased and non-diseased patients.[8-9]

Standard meta-analysis techniques using a fixed-effect or random-effects model could be used to obtain summary estimates by combining logit-transformed sensitivity and specificity and then transforming back to the original scale.[b] Such an approach assumes that the logits of sensitivity and specificity approximately follow a normal distribution. It is also possible to combine sensitivity and specificity using the number of test positives and negatives directly in a random-effects logistic model assuming a binomial distribution. There is evidence that a binomial approach performs better than the normal approximation.[10]

Positive and negative predictive values. As mentioned above, predictive values are dependent on prevalence. Because prevalence is often variable, and because many medical test studies have a case-control design (where prevalence cannot be estimated), it is rarely meaningful to combine predictive values across studies. Instead, the summary sensitivity and specificity can be used to calculate the "overall" positive and negative predictive values along a range of prevalence values. The calculations can be tabulated or plotted.

Positive and negative likelihood ratios. Positive and negative likelihood ratios could also be combined in the absence of threshold variation. Some methodological authors and medical journals give explicit guidance to combine likelihood ratios using standard fixed-effect or random-effects meta-analysis methods.[11] However, others note that this practice may result in problems under specific conditions.[12] For example, "summary" likelihood ratios calculated in this way may correspond to "summary" sensitivity and specificity values outside the 0 to 1 range.[c] Instead of summarizing likelihood ratios, these authors suggest to calculate the "summary" likelihood ratios from summary sensitivities and specificities obtained from bivariate analyses (as described in the following paragraphs). However, the actual numerical differences in the calculated quantities with alternate methods are generally small and do not result in different clinical conclusions.[13]

Diagnostic odds ratio. The synthesis of diagnostic odds ratios is straightforward and follows standard meta-analysis methods.[7,14] The diagnostic odds ratio is closely linked to sensitivity, specificity, and likelihood ratios, and can easily be included in meta-regression models to explore the impact of explanatory variables on between-study heterogeneity. In addition to challenges in interpreting diagnostic odds ratios, a disadvantage to using them is that they do not allow separate weighting of true positive and false positive rates.

[b] Note that neither of the above methods takes into account the correlation in sensitivity and specificity. In fact, this may not be a problem because in the absence of variation in thresholds, sensitivity and specificity may not be negatively correlated. However, even in this case, sensitivity and specificity are paired (they are still estimated from the same study), and this could be taken into account by modeling them jointly with a bivariate random effects model.[20] Although these models are developed with the goal of accounting for variation in threshold, they would equally apply to the situation of no variation in threshold. Generally, summary sensitivity and specificity estimated from separate random-effects meta-analyses of sensitivity and specificity are similar to those estimated from the bivariate model.[8] The advantage of the bivariate model is that a joint confidence region (instead of separate confidence interval) could be constructed for sensitivity and specificity to evaluate the two measures simultaneously and efficiently.

[c] This can happen not only when the correlation between LR+ and LR- is ignored, but also when one analyzes LR+ and LR- jointly with bivariate models.[12]

In such a situation:

- Reviewers can combine sensitivity and specificity in separate meta-analyses (preferably with a binomial method). However, it is also possible to combine them in a bivariate model.[d]
- It is recommended to back-calculate positive and negative predictive values from summary estimates of sensitivity and specificity (obtained with the bivariate method), rather than meta-analyzing them directly (in separate or joint models).
- It may be preferable to back-calculate "summary" positive and negative likelihood ratios from summary estimates of sensitivity and specificity (obtained with the bivariate method [see below]), rather than meta-analyzing them directly (in separate or joint models).
- One can summarize diagnostic odds ratios using standard meta-analysis methods.
- Generally, we recommend using a random-effects model as discussed in the *General Methods Guide*.[7]

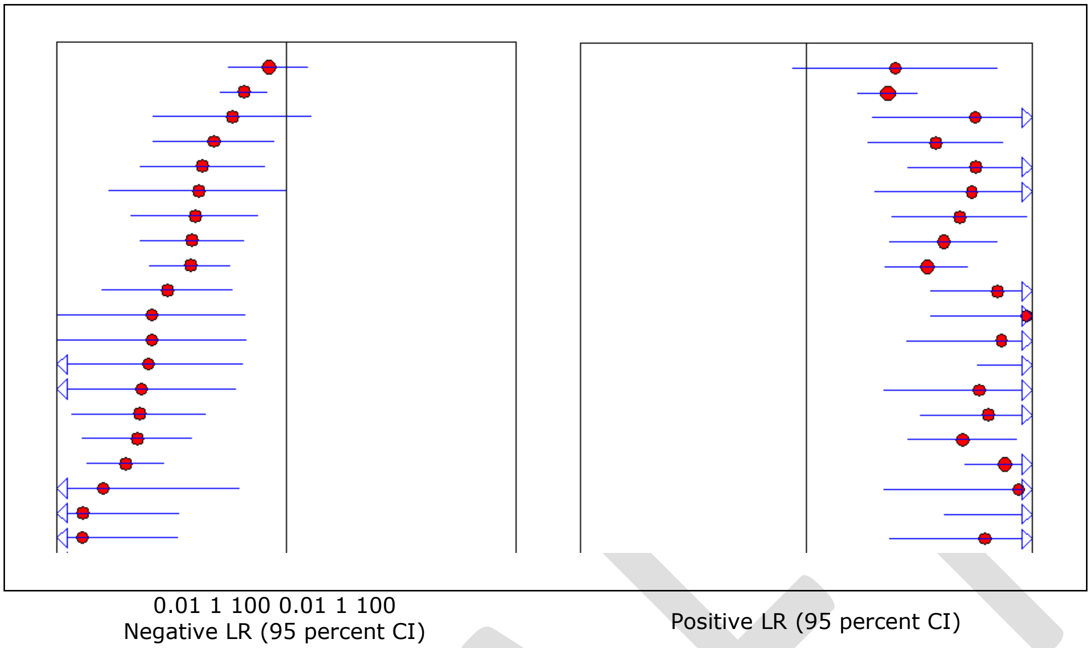
Step 2: Consider joint analysis of sensitivity and specificity when each study reports a single threshold. When there is obvious variability in the sensitivity and specificity across studies, then there is no use for more advanced methods, which were specifically developed to take into account how sensitivity and specificity are related across their range. Among other cases, this can happen when all studies use the same explicit or implicit threshold for positive tests.

In the most common scenario, there is obvious variability in the sensitivity and specificity across studies. This variability may be secondary to different explicit or implicit thresholds across studies for positive tests, composition of populations, heterogeneity in the reference standard or the index test, biases, or chance. For example, one cannot be sure that there is no variation in thresholds (e.g., for qualitative tests, such as many imaging tests). If elevated values indicate disease, a high threshold leads to low sensitivity and high specificity, while a low threshold produces high sensitivity and low specificity. This variation in threshold induces a negative correlation between sensitivity and specificity. Combining sensitivity and specificity separately (ignoring the negative correlation between them) would generally underestimate test accuracy given variation in threshold.[9] The preferred method is to model sensitivity and specificity simultaneously, accounting for tradeoff variations and for the correlation between sensitivity and specificity.

Graphing paired sets of accuracy results (sensitivity/specificity or likelihood ratios) using a forest plot, ideally in descending or ascending order of accuracy, may help illustrate the extent of variability across studies (Figure 8-1).

[d] If there is no variability in sensitivity and specificity across studies, their correlation is close to zero, and the results of the bivariate model should be similar to those from separate meta-analyses.

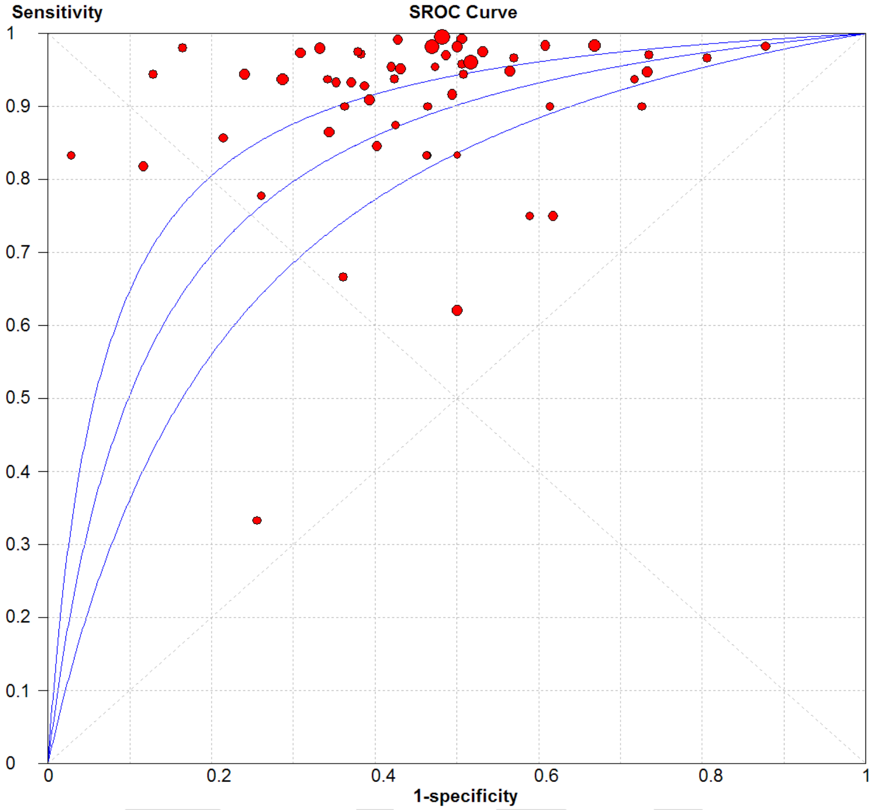
Figure 8-1. Side-by-side forest plots of “paired” negative and positive likelihood ratios



CI = confidence interval; LR = likelihood ratio. The plot represents the twenty diagnostic studies from the meta-analysis by Gupta and colleagues[32] on the ability of endometrial thickness measured by pelvic ultrasonography for diagnosing endometrial cancer. Studies shown in descending order of LR- (left panel) along with with the corresponding LR+ (right panel). Note the inverse relationship between positive and negative LRs, which is suggestive of a threshold effect.

In addition, plotting accuracy studies on a summary receiver operating characteristic (SROC) curve plane, with sensitivity on the vertical axis and 1–specificity on the horizontal, is expected to result in a “shoulder arm” pattern when a threshold effect is present. Figure 8-2 shows a typical example.

Figure 8-2. SROC curve with a “shoulder arm” pattern



SROC = summary receiver operating characteristics. SROC curve includes all studies (n = 57) included in the meta-analysis by Gupta and colleagues[32] evaluating the ability of endometrial thickness measured by pelvic ultrasonography for diagnosing endometrial cancer. The “shoulder arm” appearance suggests the presence of a “threshold effect.”

The common case is when each study reports *a single pair of sensitivity and specificity at a given threshold* (with different studies reporting different thresholds). Several statistical models have been developed to summarize medical test performance for this case. 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. We discuss both cases, emphasizing the former. Also, we discuss the synthesis of sensitivity and specificity and recommend that other measures be back-calculated from summary sensitivity and specificity estimates, as when combining studies that use the same explicit threshold.

To model the relationship between specificity and sensitivity, it is recommended always to plot the data (sensitivity versus 1–specificity) to visually assess the relationship between the two measures first. The early development of the various models has generally followed the idea to construct an SROC curve. More comprehensive approaches (random intercept, hierarchical SROC [HSROC], and bivariate sensitivity-specificity models) have been developed to address many limitations of the early models. Below is a brief comment on the strengths and limitations of the aforementioned approaches. The details of the parameterization of these models are not reviewed here.

SROC models. The most used SROC model[e] was developed by Littenberg and Moses[15-16] and attempts to explicitly model the testing threshold effect. It is essentially a regression (weighted or unweighted variations exist) of the difference of the logit-transformed true and false positive rates (sensitivity and 1–specificity, respectively) versus their sum. The back-transformed regression line, plotted in the sensitivity versus 1–specificity space, is the SROC. When there is no threshold effect, the model can also provide an estimate of the summary diagnostic odds ratio. The model is easy to implement and can be performed in most statistical packages; thus, it has been routinely used to summarize diagnostic accuracy. However, this model has major limitations, including:

- It is a fixed-effect model, which ignores unexplained variation among studies.
- It does not account for the correlation between sensitivity and specificity.
- It does not account for the variation associated with the independent variable in the model.
- It does not weight studies optimally in estimation, and therefore inferences on the effects of covariates are wrong (i.e., this model should not be used to explore effects of covariates).
- It uses an arbitrary continuity correction of 0.5 when there are empty cells in the 2-by-2 matrix.

Random intercept models. It is straightforward to extend the Littenberg and Moses model to a random-effects model, which is parameterized as a random intercept model. This model is an improvement over the Littenberg and Moses model and has been used in combining medical tests.[17-18] The random intercept model can incorporate study-level covariates. However, it does not address all the limitations listed above. It does not take into account the correlation between sensitivity and specificity and the variation associated with the independent variable in the model.

Hamza et al.[10] compared the bias and confidence interval coverage probabilities of the random intercept model and the bivariate model. So far, there is no study comparing the random intercept model and the hierarchical SROC (HSROC) model. In the Hamza et al. study,[10] the random intercept model had better coverage probabilities than the bivariate model when there are few studies (fewer than 10). However, it yields biased parameter estimates. For more than 10 studies, the bivariate model yields less biased estimates and has higher coverage probabilities.[f]

HSROC and bivariate models. The HSROC model and the closely related bivariate model address all the limitations of the simple SROC approach. Recently, Harbord et al.[8] compared the performance of the HSROC/bivariate model to several simpler models[g] and concluded that

[e] The SROC model was first proposed by Kardaun and Kardaun,[33] but it was not straightforward to implement and did not gain wide use.

[f] Hamza et al.[10] compared in simulations the performance of the bivariate random effects models with the aforementioned random intercept model estimating summary log diagnostic odds ratio examining bias, mean squared error, and coverage probabilities. When the number of studies (N) ≥ 10, the model using binomial distribution is always preferred. When the number of studies = 10, the model using binomial distribution provides a better unbiased estimate, but the coverage probability and mean squared error is often outperformed by the random intercept model.

[g] The random intercept model was not examined in this study.

HSROC/bivariate models are necessary to address variation in threshold. The HSROC and bivariate models can incorporate study-level covariates.

Rutter and Gatsonis[19] describe the HSROC model that is motivated by an ordinal regression approach and is constructed in terms of positivity threshold and accuracy parameters. The numbers of positive tests are evaluated using a binomial distribution, and a Bayesian approach is used to obtain model parameter estimates through Markov chain Monte Carlo (MCMC).[h] The Bayesian approach provides a flexible modeling and estimating framework but at the price of more programming, simulation, evaluation of model adequacy, and synthesis of simulation results.[19]

Reitsma et al.[20] and Arends et al.[21] advocate the approach of bivariate random effects model to analyze sensitivity and specificity jointly. These models preserve the two-dimensional nature of the data (sensitivity/specificity pairs), produce summary estimates of sensitivity and specificity, and incorporate any possible (negative) correlation between the two measures.[i]

Harbord et al.[22] showed that the HSROC model and the bivariate random-effects model are very closely related, and in the absence of study-level covariates, they are different parameterizations of the same model. The different parameterizations may reflect a difference in selecting a summary measure for medical test accuracy. Namely, the HSROC model naturally leads to an SROC curve when there is a threshold effect but little heterogeneity in the accuracy parameter; and the bivariate model parameterization naturally produces a summary sensitivity and specificity, together with a joint confidence or prediction region for both together. However, all other commonly used measures for medical test accuracy could readily be obtained from each model with proper calculations. Rutter and Gatsonis[19] provided formulas to estimate summary sensitivity, specificity, and likelihood ratios; several types of SROC curves could be produced from the bivariate model based on regressing logit sensitivity on logit specificity, regressing logit specificity on logit sensitivity, or using other options,[21] in addition to likelihood ratios and diagnostic odds ratios.[20]

Chappell et al.[6] discuss that if there is substantial evidence that the correlation between sensitivity and specificity is *positive* (opposite from what is theoretically expected), then an SROC curve will be meaningless, and perhaps an important covariate has not been taken into account.[j] In that event, they suggest that one performs independent analyses of sensitivity and specificity to try to identify a likely explanation.

[h] Macaskill[34] implemented the same model in a classical framework using the SAS® procedure NLMIXED and produced comparable results without the degree of complexity inherent in the MCMC simulation.

[i] The bivariate random effects model can be parameterized assuming that the logits of sensitivity and specificity follow a bivariate normal distribution,[20,21] alternatively it can model the number of positive tests using the binomial distribution.[21,35] One disadvantage of the former model is that an arbitrary correction of 0.5 is required to avoid undefined log odds when the data are sparse with zero positive tests. The model using binomial distribution generally performs better than the model using normal distribution for logits of sensitivity and specificity, though the difference is not always practically relevant when the study sample sizes are large.[10,21]

[j] Chappell et al.[6] point out that they have still to encounter an example where the correlation between sensitivity and specificity is, e.g., positive with probability >0.95. However, they comment that it is not uncommon for the median or mean of the posterior density of this correlation to be positive. In that case, the model is probably not well estimated.

Choice of models. We make the following recommendations about selecting the appropriate model:

1) There is adequate theoretical motivation to discourage the use of the Littenberg and Moses model to draw inferences (compare subgroups of studies or perform direct or indirect comparisons between two or more index tests).

2) We do not recommend the routine use of the random intercept model.

3) Based on currently available empirical evidence, we encourage the use of the HSROC model or the bivariate model (preferably using the binomial distribution). These models are quite complex. Nevertheless, the complexity of the diagnostic data demands such models.[k]

4) When there is no covariate in the model, the HSROC and bivariate models are equivalent. The primary interests of inference may help guide the choice of model. That is, if the investigators are mostly interested in an SROC curve (or the research question makes an SROC curve a better choice), then the HSROC model is more convenient to use. If the research question makes a summary sensitivity or specificity the most appropriate measure, the bivariate model would be more convenient. For example, if the reported sensitivity and specificity has a wide range of values, an SROC would be appropriate. On the other hand, if the reported sensitivity and specificity are very similar to each other and located only in a small portion of the ROC space, then it would not be appropriate to extrapolate and construct a full ROC curve, and summary sensitivity or specificity are better choices by using the bivariate model. Harbord et al.[22] also explained how the difference in design and conduct of the included diagnostic accuracy studies may affect the choice of the model. For example, "spectrum effects," where the subjects included in a study are not representative of the patients who will receive the test in practice,[23] "might be expected to affect test accuracy rather than 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."[22] Investigators are encouraged to look at study characteristics and evaluate how these study characteristics could affect the diagnostic accuracy, which in turn might affect the choice of model. Further research is also needed to evaluate how study characteristics are associated with the performance of these models.

5) When there are covariates in the model, the HSROC model allows direct evaluation of the difference in accuracy or threshold parameters or both. Bivariate models, on the other hand, allow for direct evaluation of the difference in sensitivity or specificity or both. In addition, the HSROC model can be more easily extended to

[k] If the systematic reviewers opt not to use a Bayesian approach, they could use the SAS NLMIXED procedure or the Stata® *xtmelogit* or *gllamm* commands. Note that the proper syntax for these commands can become complicated in that subtle variations may result in fitting a different model.

include a covariate to affect the degree of asymmetry of the SROC curve.[22] Investigators are encouraged to consider these factors in model choice.

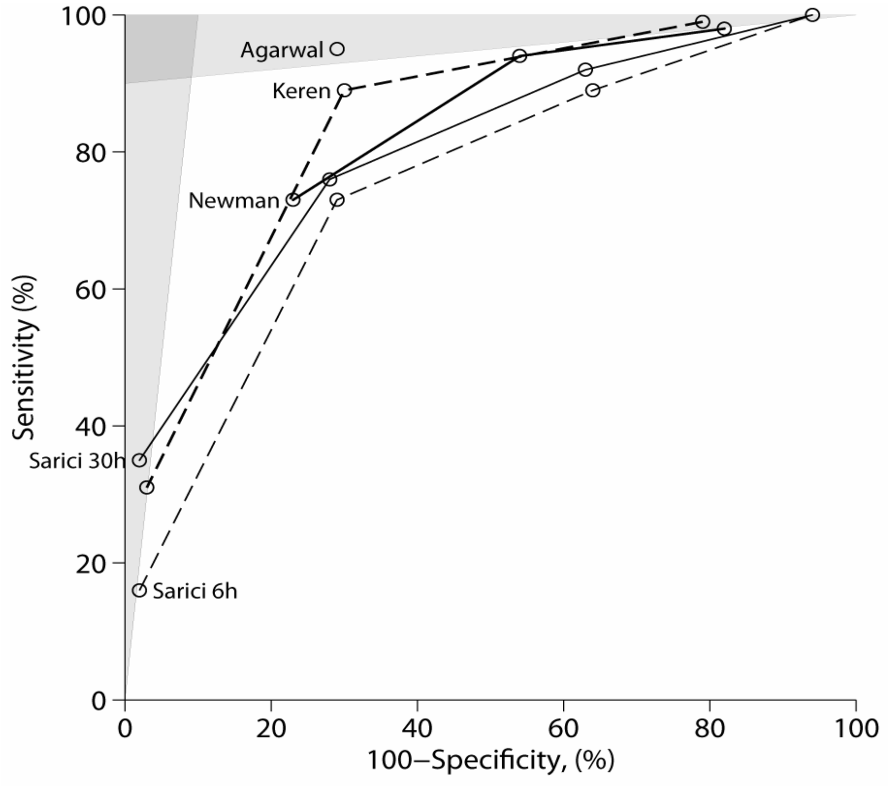
Step 3: Consider joint analysis of sensitivity and specificity when studies report multiple thresholds. It is not uncommon for some studies to report multiple sensitivity/specificity pairs. One option is to decide on a single threshold from each study and apply the aforementioned methods. To some extent, the setting in which the test is used can guide the selection of the threshold. For example, in some cases, the threshold that gives the highest sensitivity may be appropriate in medical tests to rule out disease. Another option is to use all available thresholds per study, as detailed below.

Statistical models. An extension of the HSROC model has been developed to analyze sensitivity and specificity data reported at more than one threshold.[24] This model explicitly uses latent variables and is more challenging to implement than the HSROC and bivariate models discussed above. Related models have been recently proposed in the literature.[25] Further, if each study reports enough data on sensitivity and specificity to construct a ROC curve, Kester and Buntinx[26] have proposed a so far little-used method to combine whole ROC curves.

Choice of models. Both models are theoretically motivated. The Dukic and Gatsonis model[24] is more elaborate and more technical in its implementation than the Kester and Buntinx variant.[26] 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 always to perform such analyses.[1] At a minimum, we suggest that systematic reviewers perform explorations in a qualitative, graphical depiction of the data in the ROC space. This will provide a qualitative summary and highlight similarities and differences among the studies. An example of such a graph is in Figure 8-3, which illustrates the diagnostic ability of early measurements of total serum bilirubin (TSB) to identify postdischarge TSB above the 95th 10-hour-specific percentile.[27]

[1] Systematic reviewers are 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.

Figure 8-3. Diagnostic ability of early TSB measurements



TSB = total serum bilirubin. Ability of early TSB measurements to identify postdischarge TSB above the 95 percent 10-hour-specific percentile. Sensitivity/100 percent minus specificity pairs from the same study (obtained with different cutoffs for the early TSB measurement) are connected with lines. These lines are reconstructed based on the reported cutoffs and are not perfect representations of the actual ROC curves in each study (they lack granularity). Studies listed on the left shaded area have an LR+ of at least 10. Studies listed on the top shaded area have an LR- of at most 0.1. Studies listed at the intersection of the gray areas (darker gray polygon) have both LR+ of at least 10 and LR- of 0.1 or less.[27]

Step 4. Explore between-study heterogeneity. Heterogeneity is common in systematic reviews of medical tests due to differences in threshold or cutoff values used by different studies of an index test to define the presence of the target condition. The difference in threshold can be explicit; for example, studies that use different numerical cutoff values of blood glucose level to determine the presence of diabetes. Or the difference in threshold can be implicit, as when decision on a positive test depends on subjective judgment and interpretation; for example, the abnormality of radiographs. This "threshold effect" is valid only if all the studies use an equivalent index test; that is, in the case of a blood test, one manufacturer's test method gives the same values as another manufacturer's test method for the same sample. In practice, this is not often the case, and special attention to method differences is required before heterogeneity due to a threshold effect can be investigated.

Exploring heterogeneity using hierarchical models. Other than accounting for the presence of a threshold effect, HSROC and bivariate models provide flexible ways to test and explore between-study heterogeneity simultaneously. For either model, the formulation includes a variance parameter to test the presence and magnitude of between-study heterogeneity. Both models allow for inclusion of study characteristics to explore how study characteristics could explain sources of heterogeneity across studies and allow for exploration in two measures at once. As mentioned above, an HSROC model allows for direct evaluation of the heterogeneity in accuracy or threshold parameters or both, and study characteristics for each parameter may be the same or different. Bivariate models, on the other hand, allow for direct evaluation of heterogeneity in sensitivity or specificity or both, and again, study characteristics for each measure could be different. After inclusion of study characteristics, the estimate of variance parameter provides an estimate of the residual variance and helps evaluate the amount of heterogeneity explained by these characteristics by comparing the estimate of variance parameter before including the study characteristics. Factors reflecting differences in patient population and method in patient selection, methods of verification and interpretation of results, clinical setting, disease severity, etc., are common considerations for the source of heterogeneity. Investigators are encouraged to use these models to test and explore heterogeneity, especially when they have chosen these methods for combining studies.

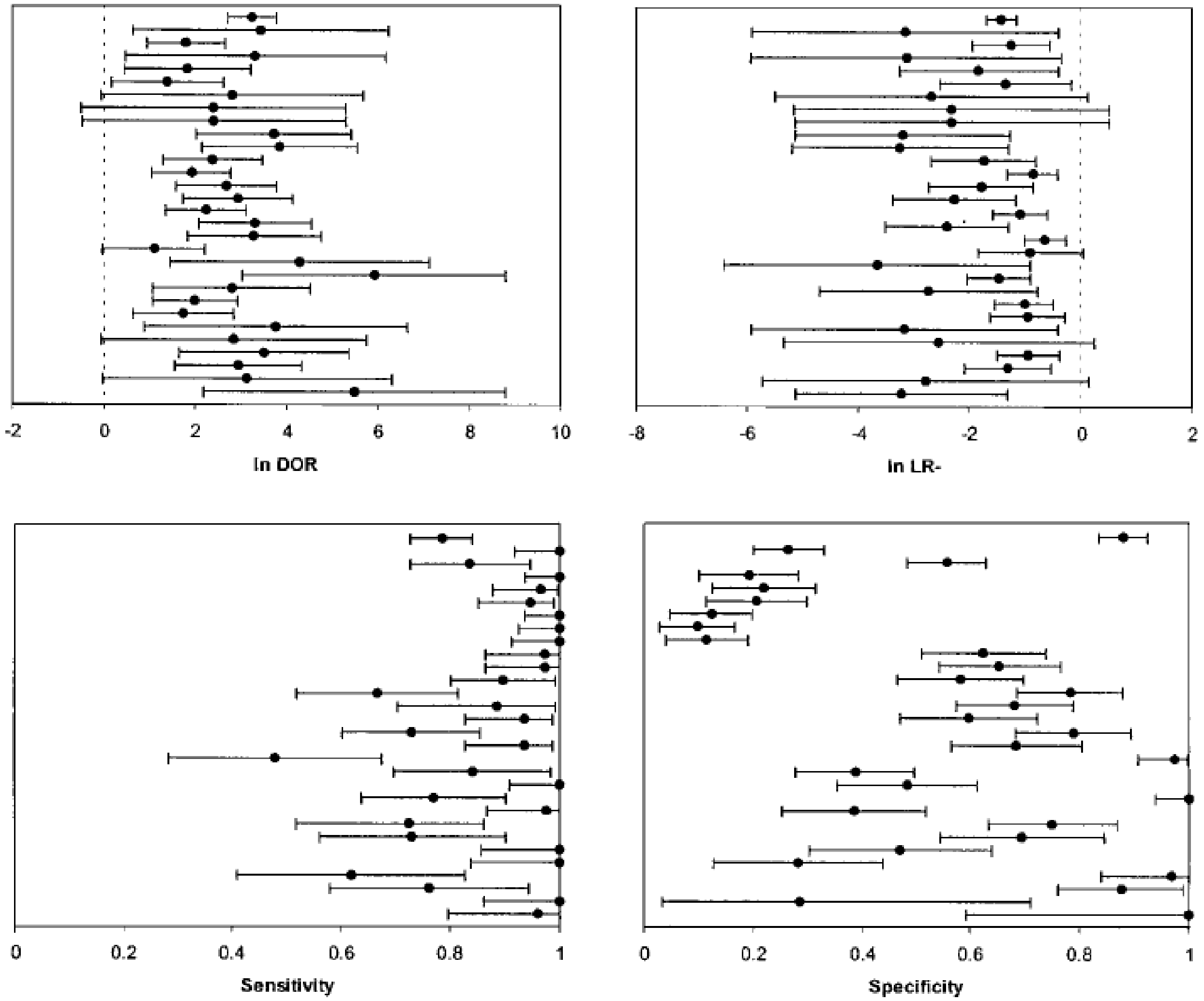
Illustrations

We briefly demonstrate the above with two applied examples. Example 1 is on D-dimer assays for the diagnosis of venous thromboembolism[28] and shows heterogeneity due to "threshold effect"; it has been discussed by Lijmer et al.[29] Example 2 is from an EPC evidence report on the ability of serial creatine kinase-MB measurements to diagnosis acute cardiac ischemia[30-31] and shows heterogeneity for another reason.

Example 1: D-dimers for diagnosis of venous thromboembolism. D-dimers are fragments specific to fibrin degradation in blood or plasma and can be used to diagnose venous thromboembolism. Figure 8-4 presents forest plots of the (log-transformed) diagnostic odds ratio, negative likelihood ratio,[m] sensitivity, and specificity for the D-dimers example.[29] As discussed above, one could, in theory, synthesize each of these measures. Note that sensitivity and specificity are more heterogeneous than the diagnostic odds ratio or the negative likelihood ratio (this can be verified by formal testing for heterogeneity). This may be due to threshold variation in these studies (from 25 to 550 ng/mL, when stated; Figure 8-5), or due to other reasons.[29]

[m] This is a diagnostic test aiming to identify the majority of patients with venous thromboembolism. Therefore, the negative likelihood ratio is arguably quite important to note.

Figure 8-4. The log-odds ratio (ln DOR), log-likelihood ratio of a negative test (ln LR-), sensitivity, and specificity of 30 evaluations of D-dimer assays for detecting venous thromboembolism



In Figure 8-4,[29] points indicate estimates; horizontal lines are 95 percent confidence intervals for the estimates. Such plots help to appreciate the extent of heterogeneity in different measures of medical test performance. Heterogeneity is much more pronounced for sensitivity and specificity rather than for the diagnostic odds ratio (log-transformed, ln DOR) or the negative likelihood ratio (lon transformed, ln LR-). As discussed in the previous section, these measures express different aspects of test performance, and one could synthesize any of them. Recall that the summary diagnostic odds ratio gives an overall indication of ability of the D-dimers to distinguish venous thromboembolism from other conditions, but it does not inform on whether it performs better in people with the disease or in people without the disease. As noted in the text above, pooling likelihood ratios may result in "summary" likelihood ratios that do not correspond to meaningful "summary" sensitivity/specificity estimates. In this example, there is large variability in the diagnostic threshold used in the different studies. Therefore, univariate (separate) syntheses of sensitivity and specificity are probably not meaningful. Heterogeneity in the thresholds is one of the potential explanations for the observed between-study heterogeneity in this example.

Figure 8-5. Variation in threshold in the D-dimer example

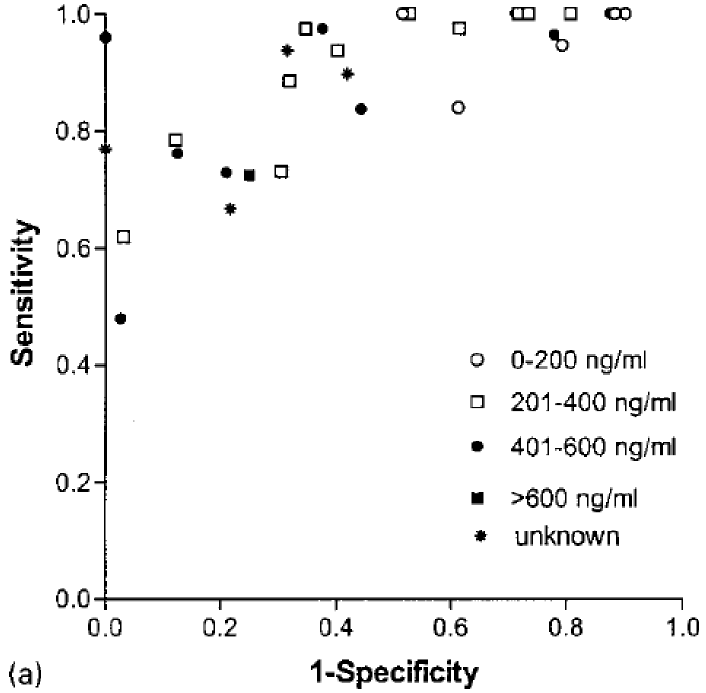


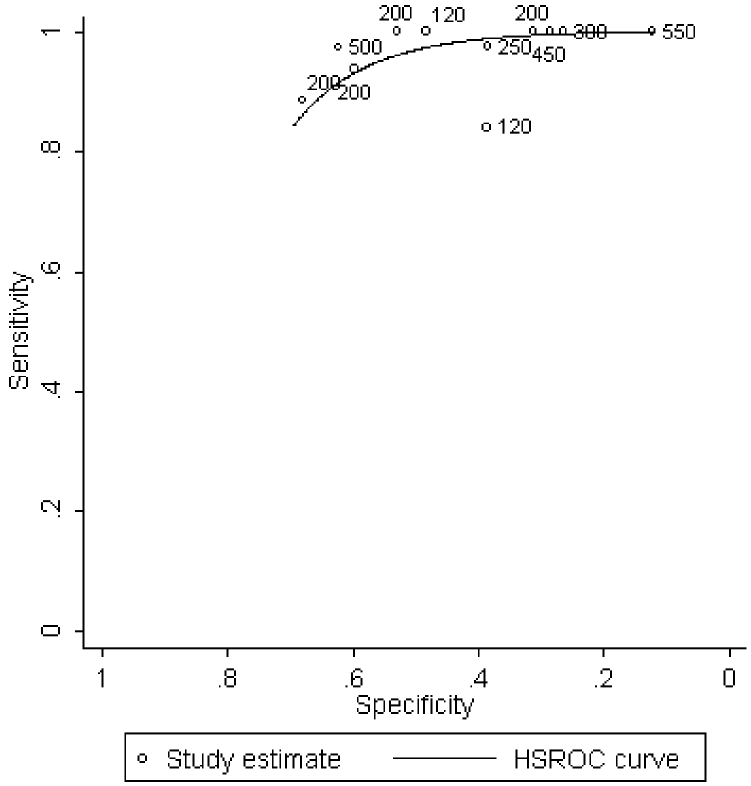
Figure 8-5 is from Lijmer et al.,[29] with data from Becker et al.[28] Shown is a sensitivity/1–specificity plot for the D-dimers example. Different markers denote different thresholds for the test. The "shoulder" image that the distribution of studies follows is seen in many meta-analyses where there is threshold-induced heterogeneity.

Since there is evident variation in the thresholds for studies of D-dimers, it is more appropriate to summarize the performance of the test using an HSROC curve rather than provide summary sensitivities and specificities.[n] The HSROC describes the tradeoff between sensitivity and specificity in the various studies. Shown below is the HSROC curve for the ELISA-based D-dimer tests in 11 studies (Figure 8-6). (For simplicity, we selected the highest threshold from two studies that reported multiple ELISA thresholds.)

Interpreting the HSROC curve (or the summary estimates) in this case is straightforward. This test has very good diagnostic ability, and it appropriately focuses on minimizing false negative diagnoses. It is also informative to "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 8-7 shows such an example using the summary sensitivity and specificity of the 11 studies of Figure 8-6.

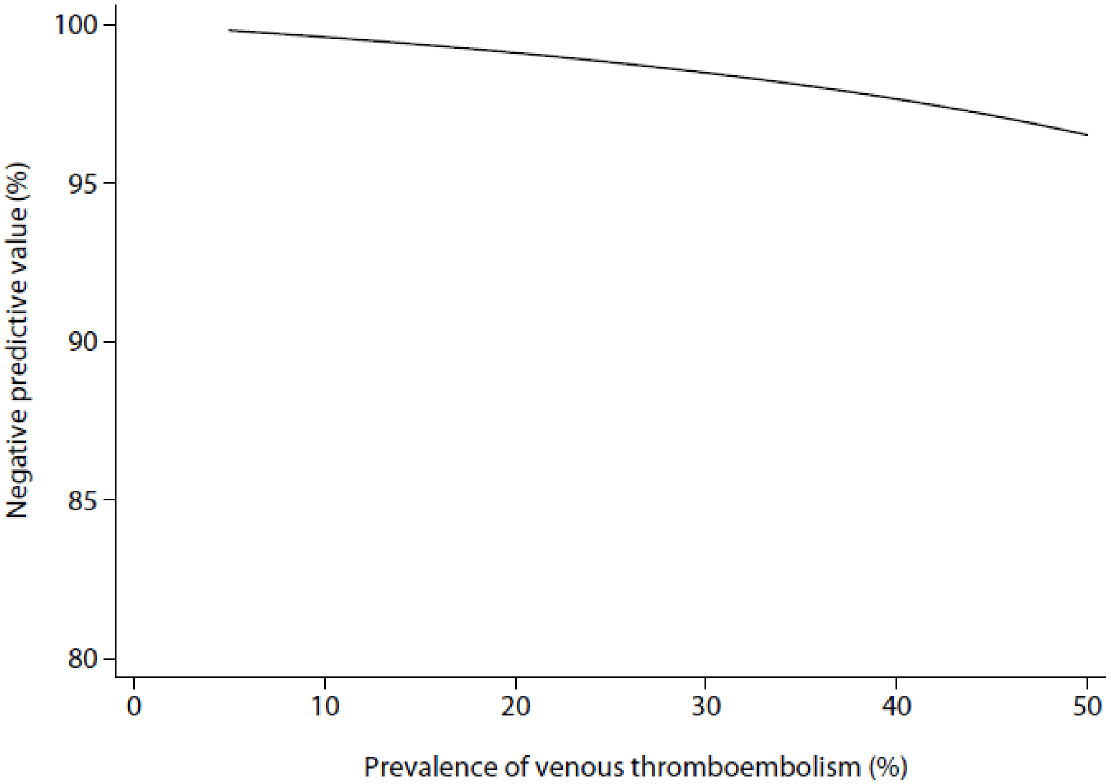
[n] Because sensitivity and specificity change across thresholds, it is unclear what a "summary" sensitivity and specificity would mean.

Figure 8-6. HSROC for the ELISA-based D-dimer tests



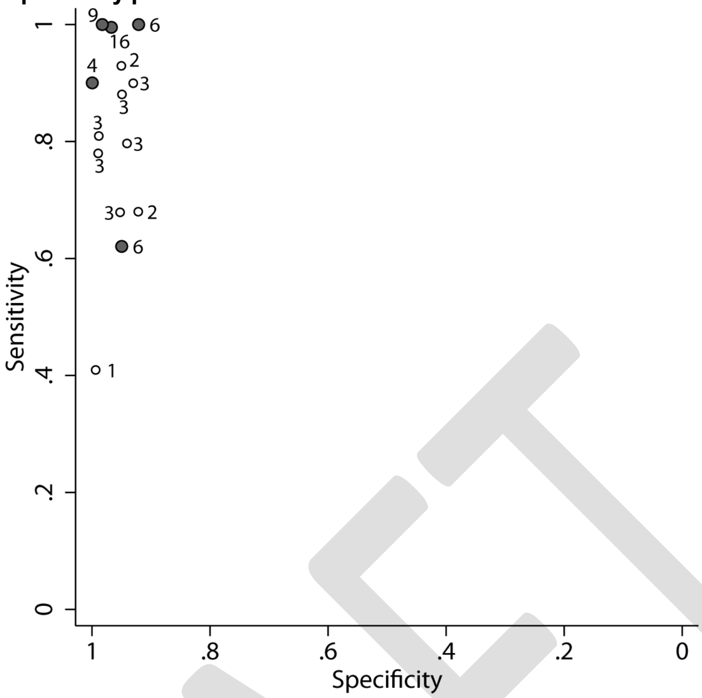
HSROC = hierarchical summary receiver-operator curve. Note that the horizontal axis is essentially 1–specificity (because it is reversed). Also shown is the threshold (in ng/mL) that was used in each study. This analysis used the *metandi* module for Stata (Roger Harbord's wrapper command for the more general *xtmelogit* and *gllamm* commands in Stata). This analysis is not included in the methodological paper by Lijmer et al.[29]

Figure 8-7. Calculated negative predictive value for the ELISA-based D-dimer test if the sensitivity and specificity are fixed at a summary value (x and y, respectively) and prevalence of venous thromboembolism is between 5 and 50 percent



Example 2: 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.[30-31] Figure 8-8 shows the 14 eligible studies along with how many hours after symptom onset the last measurement was taken. It is evident that there is heterogeneity in the sensitivity, and that sensitivity increases with longer time from symptom onset.

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



Note that since it is reversed, the horizontal axis effectively represents 1–specificity as usual for an ROC plot. Numbers are the number of hours after symptom onset that the last CK-MB measurement was taken. Filled circles are studies with a last measurement taken more than 3 hours after 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 the remaining studies(Table 8-1).[o] We use a bivariate multilevel model with exact binomial likelihoods, as discussed above. In the fixed-effect part of the model, we include an indicator variable that codes whether the last measurement was earlier than 3 hours from symptom onset. We allow this variable to have different effects on the summary sensitivity and specificity.[p] This is essentially a bivariate meta-regression.

Table 8-1. Comparison of diagnostic performance of studies according to timing of the last serial CK-MB measurement for diagnosis of acute cardiac ischemia

	≤ 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

The corresponding meta-regression can be specified for the HSROC model, in which case the results will be expressed with respect to the accuracy and threshold parameters (more difficult to interpret for a clinical audience).

[o] This analysis was not performed in the evidence report. It is performed here for illustration.

[p] Analyses are performed in Intercooled Stata 10, using the *xtmelogit* command. The same results are obtained with the *gllamm* command in earlier versions of Stata. The same model can be run in other statistical languages. Care has to be taken to specify the model correctly.

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

Notably, the evidence report described above included systematic reviews of several technologies (tests), including the aforementioned example. To interpret and contextualize the findings of these systematic reviews, the researchers used decision modeling analyses that compared 17 technologies and 4 test combinations.[31]

Summary

Key points are:

- Summarizing the performance of diagnostic tests is a multidimensional problem, as there are several quantities of interest that are, in principle, correlated and contribute complementary information.
- Separate meta-analyses of sensitivity and specificity are suitable if there is no evidence of variability in the sensitivity or specificity across studies. This can happen when all studies use the same explicit threshold for positive tests, but also in other cases.
- If there is substantial variation in sensitivity or specificity, one should consider methods that analyze these quantities jointly. The bivariate meta-analysis of sensitvity and specificity and the closely related HSROC method are theoretically motivated.
- If more than one threshold is reported per study, this has to be taken into account in the quantitative analyses. Tentatively, we encourage both qualitative analysis via graphs and quantitative analyses via appropriate methods.
- Reviewers 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. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003;138(1):W1-12.

2. 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.

3. 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.

4. Rutjes AW, Reitsma JB, Di NM, et al. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006;174(4):469-76.

5. Whiting P, Rutjes AW, 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.

6. Chappell FM, Raab GM, Wardlaw JM. When are summary ROC curves appropriate for diagnostic meta-analyses? Stat Med 2009;28(21):2653-68.

7. Agency for Healthcare Research and Quality. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville, MD: Agency for Healthcare Research and Quality. Available at: http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318. Accessed September 20, 2010.

8. Harbord RM, Whiting P, Sterne JA, 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.

9. Irwig L, Tosteson AN, Gatsonis C, et al. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 1994;120(8):667-76.

10. 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.

11. Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. BMJ 2004;329(7458):168-9.

12. Zwinderman AH, Bossuyt PM. We should not pool diagnostic likelihood ratios in systematic reviews. Stat Med 2008;27(5):687-97.

13. Simel DL, Bossuyt PM. Differences between univariate and bivariate models for summarizing diagnostic accuracy may not be large. J Clin Epidemiol 2009;62(12):1292-300.

14. Glas AS, Lijmer JG, Prins MH, et al. The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol 2003;56(11):1129-35.

15. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Making 1993;13(4):313-21.

16. 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):1293-316.

17. Oei EH, Nikken JJ, Verstijnen AC, et al. MR imaging of the menisci and cruciate ligaments: a systematic review. Radiology 2003;226(3):837-48.

18. 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.

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

20. Reitsma JB, Glas AS, Rutjes AW, et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005;58(10):982-90.

21. Arends LR, Hamza TH, van Houwelingen JC, et al. Bivariate random effects meta-analysis of ROC curves. Med Decis Making 2008;28(5):621-38.

22. Harbord RM, Deeks JJ, Egger M, et al. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics 2007;8(2):239-51.

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

24. Dukic V, Gatsonis C. Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. Biometrics 2003;59(4):936-46.

25. Hamza TH, Arends LR, van Houwelingen HC, et al. Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds. BMC Med Res Methodol 2009;9:73.

26. Kester AD, Buntinx F. Meta-analysis of ROC curves. Med Decis Making 2000;20(4):430-9.

27. Trikalinos TA, Ip S, Raman G, et al. Home diagnosis of obstructive sleep apnea-hypopnea syndrome. Technology Assessment. Rockville, MD: Agency for Healthcare Research and Quality. August 2007. Available at: http://www.cms.hhs.gov/determinationprocess/downloads/id48TA.pdf. Accessed July 9, 2010.

28. Becker DM, Philbrick JT, Bachhuber TL, et al. D-dimer testing and acute venous thromboembolism. A shortcut to accurate diagnosis? Arch Intern Med 1996;156(9):939-46.

29. Lijmer JG, Bossuyt PM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. Stat Med 2002;21(11):1525-37.

30. Balk EM, Ioannidis JP, Salem D, et al. Accuracy of biomarkers to diagnose acute cardiac ischemia in the emergency department: a meta-analysis. Ann Emerg Med 2001;37(5):478-94.

31. Lau J, Ioannidis J, Balk E, et al. Evaluation of Technologies for Identifying Acute Cardiac Ischemia in Emergency Departments. Evidence Report/Technology Assessment Number 26. (Prepared by The New England Medical Center Evidence-based Practice Center under Contract No. 290-97-0019) AHRQ Publication No. 01-E006, Rockville, MD: Agency for Healthcare Research and Quality. May 2001. Available at:

http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=hstat1.chapter.37233. Accessed July 9, 2010.

32. Gupta JK, Chien PF, Voit D, Clark TJ, Khan KS. Ultrasonographic endometrial thickness for diagnosing endometrial pathology in women with postmenopausal bleeding: a meta-analysis. Acta Obstet Gynecol Scand. 2002 Sep;81(9):799-816.

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. Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. J Clin Epidemiol 2004;57(9):925-32.

35. 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):1331-2.

Methods Guide for Medical Test reviews

Paper 9

Meta-analysis of Test Performance Evidence When There is an Imperfect Reference Standard

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

AHRQ Publication No. xx-EHCxxx-EF
<Month Year>

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, through a collaborative effort within the Evidence-based Practice Center Program, have developed a Methods Guide for Medical Test Reviews. We intend that these documents will serve as a resource for our EPCs as well as for other investigators interested in conducting systematic reviews on medical tests.

This Medical Test Methods guide is intended to be a practical guide for those who prepare and use systematic reviews on medical tests. This document complements the EPC Methods Guide on Comparative Effectiveness Reviews (http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318), which focuses on methods to assess the effectiveness of treatments and interventions. The guidance here applies the same principles for assessing treatments to the issues and challenges in assessing medical tests and 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. We provide guidance in stepwise fashion for those conducting a systematic review.

The *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 *Medical Test Methods Guide* and the Effective Health Care Program can be made at www.effectivehealthcare.ahrq.gov.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

Paper 9. Meta-analysis of Test Performance Evidence When There is an Imperfect Reference Standard

To evaluate the performance of a medical test (index test), we have to compare its results with the “true” status of every tested individual or specimen. Sometimes, this true status is directly observable; for example, when we use a test to predict short-term mortality after a procedure, or when we verify whether a suspect lesion is malignant with an excisional biopsy. However, in many cases the true status of the tested unit is judged based on another test as a reference method. Problems can arise when the reference standard test does not mirror truth adequately well: in such a case, we will be measuring the performance of the index test against a faulty standard, and we are bound to err. In fact, the further the reference standard test deviates from the truth, the poorer our estimate of the index test’s performance will be. This is otherwise known as “reference standard bias.”[1-4]

Common Challenges

Only rarely are we absolutely sure that a reference standard test is a perfect reflection of the truth. Most often, we are very comfortable with overlooking small or moderate misclassifications by the reference standard. In fact, this is exactly what we do, implicitly, when we calculate the index test’s sensitivity, specificity, and related quantities. But how should we approach the evaluation of a diagnostic or prognostic test when the reference standard itself performs (too) poorly? Table 9-1 lists some common situations where we might question the validity of the reference standard. We do not discuss the case of a “missing gold standard,” where the reference standard is guided by the results of the index test and not universally applied (also known as verification bias).

Table 9-1. Situations where the validity of the reference standard is in question

Situation	Example
The reference standard test 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 between-rater and between-laboratory variability.
The definition of the “disease” being tested for is idiosyncratic to some extent.	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, and imaging and laboratory measurements are used to identify it. Unavoidably, diagnostic criteria will be differentially applied across studies, and this is a potential explanation of the varying prevalence of the disease across geographic locations.[5]

Situation	Example
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.[6] In this case, 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.

Principles for Addressing the Challenges

Principle 1: Favor the Simplest Analysis That Properly Summarizes the Data

There are several ways to approach a systematic review of medical tests when the reference standard is a poor approximation of the truth, or when no test can be regarded as a reference standard. One can choose among the following options depending on the topic at hand:

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

(2) Adjust or correct the "naïve" estimates of sensitivity and specificity of the index test to account for the imperfect reference standard.

(3) Assess the concordance of difference tests instead of test accuracy.

Assess the test's ability to predict patient-relevant outcomes instead of the test's accuracy. Recast the review so that it does not aim to calculate estimates of test performance[a] but rather aims to assess whether test results predict relevant clinical data, such as history, future clinical events, and response to therapy.[2] Essentially, this implies ignoring all information that compares the index test with the imperfect reference standard. The rationale is that this information is not informative or interpretable. We will not discuss this option further. Paper 12 of the *Medical Test Guide* elaborates on the evaluation of prognostic tests.

Adjust or correct the naïve estimates of sensitivity and specificity. One can adjust or correct the naïve estimates of sensitivity and specificity of the index test to account for the imperfect reference standard. This implies that, apart from the index test, there is a test that can be regarded as a reference standard, albeit an imperfect one. In such case, one cannot calculate sensitivity (ability to maximize true positives) and specificity (ability to minimize false positives) of the test using only the observed data.

To overcome this problem, one must either use constraints on a subset of the parameters (i.e., assume that the sensitivity and specificity of, e.g., the reference standard to detect true disease status is known,[7] or that both specificities are known but the sensitivities are unknown[8]), or use an approach that treats the missing gold standard as latent data and estimate it by combining external information from other sources (prior distributions) with the available data using Bayesian inference.[9-11] These prior distributions provide a different type of constraint, which

[a] Test performance as meant in Paper 8: "accuracy" measures such as sensitivity and specificity of the index test or other metrics that can be derived from these quantities.

makes use of prior knowledge of the parameters and thus may be less arbitrary. The resulting posterior distribution provides information on the specificities and sensitivities of both the index test and the reference standard, and on the prevalence of people with the diseases in each study.

In addition, statistical adjustment requires meta-analysts either to assume "conditional independence" between the index and reference standard results (which often does not hold), or to supply data regarding the between-test correlation if conditional independence is not assumed. However, such correlation data are not commonly available.[3]

Assess the concordance of difference tests instead of test accuracy. Here, the reviewer is no longer interested in the sensitivity and specificity of the examined test but rather in how well it agrees with the other test(s) and, perhaps, whether one test can be used in place of the other. Assessing concordance may be the only realistic option if none of the compared tests is an obvious choice for a reference standard; for example, when both tests are alternative methodologies to measure the same quantity.

Depending on the topic at hand, one can summarize the extent of agreement between two tests using Cohen's kappa statistics (a measure of categorical agreement which takes into account the probability that some agreement will occur by chance), Bland-Altman plots (agreement of continuous measurements),[12-14] and interclass correlation (ICC) statistics. It is anticipated that, in many cases, quantitative summaries of concordance data will not be particularly meaningful or even possible; then, the review can be limited to a qualitative descriptive analysis of the diagnostic research available.

Principle 2: Qualify Findings To Avoid Misinterpretations

As mentioned above, when the reference standard is (grossly) misclassifying the "true" disease status, the usually calculated estimates of sensitivity and specificity (naïve estimates) are biased. The direction of the bias can be either upward or downward, and the magnitude will depend on the frequency of reference standard errors and the degree of correlation in errors between the index test and reference standard.

Conditionally independent tests. In the simplest case, the index test and the reference standard are independent conditional on disease status. In other words, the two tests do not tend to agree more (or less) than expected among people with the disease or people without the disease. Then the naïve estimates of sensitivity and specificity are underestimates.

Conditionally dependent tests. When the two tests are correlated conditional on disease status, the naïve estimates of sensitivity and specificity can be overestimates or underestimates. Overestimates can happen when the tests tend to agree more than expected by chance. Underestimates can happen when the correlation is relatively small or the tests disagree more than expected by chance.[b] It is common for researchers to assume conditional independence of the results of different tests. However, this assumption cannot be uniformly justified, particularly when the tests are based on a common mechanism (e.g., both tests are based on a particular

[b] We are hard pressed to find a good clinical example where tests disagree more than expected by chance; however, it is a theoretical possibility.

chemical reaction so that something that interferes with the reaction for one of the tests will likely interfere with the other as well).[15]

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.[16] One explanation for these misclassifications (false-negative results) is that obesity can reduce serum PSA levels. In this situation, 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 were compared to a non-PSA-based reference standard that was not prone to error due to obesity, such as a 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.

Therefore, in deciding how to approach the evaluation of test performance in the face of an imperfect reference standard, consider the following guidance:

(1) If not established at the outset that one test constitutes the reference standard and if there are multiple alternatives for a reference standard, decide which reference standard is most common (or otherwise acceptable) for the main analysis.[c] Consider exploring alternative reference standards for completeness.

(2) Decide which is more informative for the user of the systematic review:
   a. To have estimates of sensitivity and specificity for the index test?
   b. To learn whether the index test and the reference test are concordant or interchangeable?
   c. Both?

(3) Keep analyses and presentations simple. There is a relative paucity of empirical data on the merits and pitfalls of using the more complex methods that account for imperfect reference standards. We recommend that, in the majority of cases, reviewers should perform at most simple analyses, followed by clear discussion and illustration using relevant examples. We believe that it is transparent and instructive to:
   a. Present naïve estimates of sensitivity, specificity, and related analyses.
   b. Discuss the expected direction and magnitude of the bias of the naïve estimates.

Illustration

In this section, we expand on the sleep apnea technology assessment report to illustrate the issue of an imperfect reference standard.[17] This is a useful case as there is no accepted reference standard for the diagnosis of sleep apnea. In this example, we will indicate how to apply the principles discussed above to the estimation of performance of a proposed test for this condition.

[c] This should include consideration of cut points for continuous or ordinal test results, or clusters for tests with multiple unordered categories.

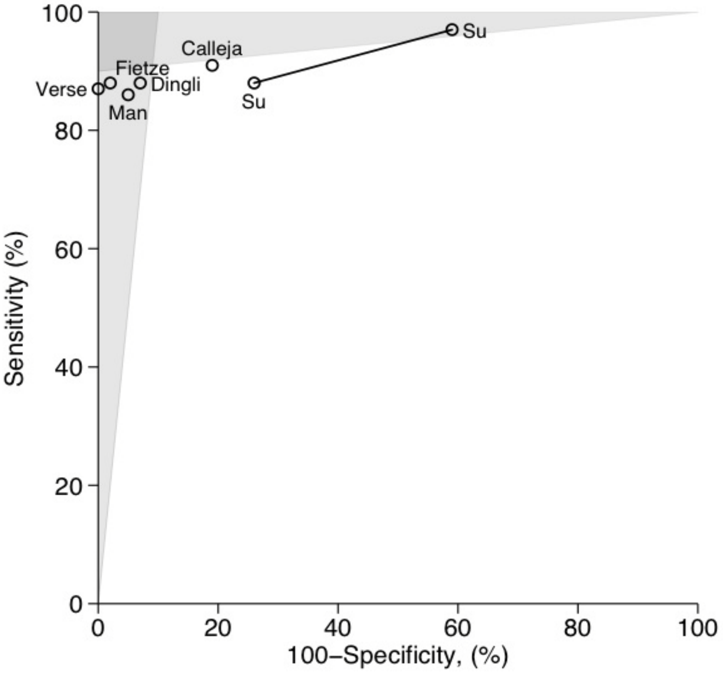
Sleep apnea technology assessment. As noted above, one consideration is what reference standard is most common or otherwise acceptable for the main analysis. In this case, the diagnosis of sleep apnea is usually (but imperfectly) diagnosed when the patient has a high Apnea-Hypopnea Index (AHI, an objective measurement), together with suggestive symptoms and signs. In all reviewed studies, patients were enrolled only if they had suggestive symptoms and signs, although it is likely that these were differentially ascertained across studies. Therefore, the definition of sleep apnea is reduced to whether people have a high enough AHI.

Defining the reference standard. Most studies and some guidelines define AHI ≥ 15 events per hour of sleep as being 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, 40 events per hour of sleep). As a sensitivity analysis, the reviewers decided to summarize studies also according to the 10 and 20 events per hour of sleep cutoffs; the other cutoffs were excluded because of data availability. It is worth noting that, in this case, the exploration of the alternative cutoffs did not affect the results or conclusions of the technology assessment but did require substantial time and effort.

Deciding how to summarize and present the findings of individual studies. Following the aforementioned principles, the reviewers decided to calculate and interpret naïve estimates of sensitivity and specificity of portable monitors and to describe the concordance of measurements with portable monitors (index test) and facility-based polysomnography (reference test).

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 9-1). These graphs suggest a high "sensitivity" and "specificity" of portable monitors to diagnose AHI ≥ 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-laboratory 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 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), and they identified this as a gap in the pertinent literature.

Figure 9-1. Diagnostic ability of type III monitors in specialized sleep units to identify AHI>15 events/hour in laboratory-based polysomnography: manual or combined manual and automated scoring

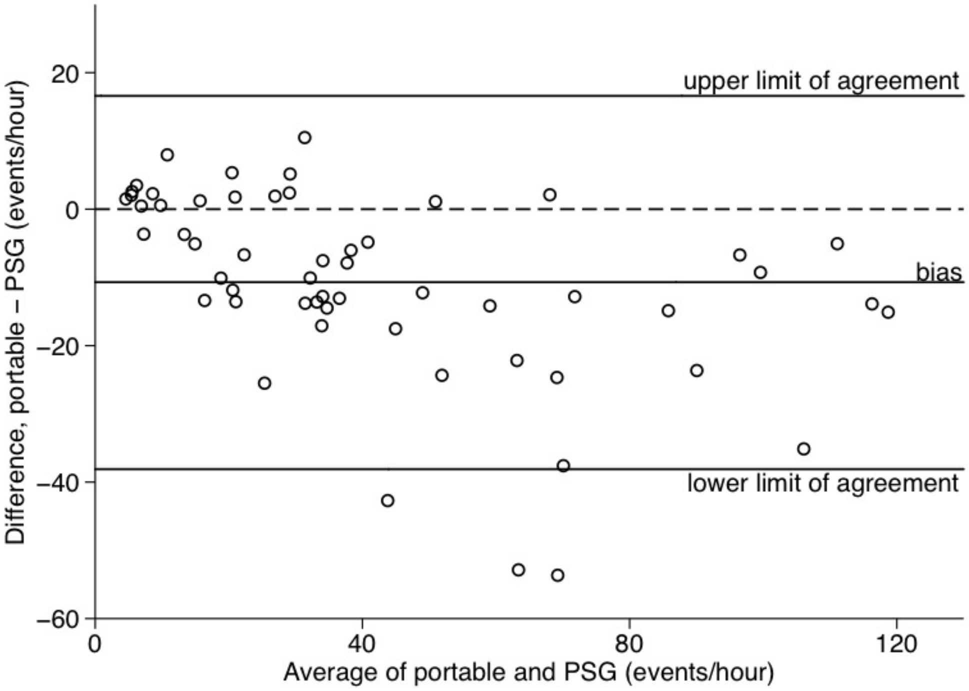


Sensitivity/specificity pairs from the same study (obtained with different cutoffs for the type III monitor) are connected with lines. These lines are not representative of the ROC curve of the pertinent studies. Studies lying on the left shaded area have a positive likelihood ratio (LR+) of 10 or more. Studies lying on the top shaded are have a negative likelihood ratio (LR-) of 0.1 or less. Studies lying on the intersection of the gray areas (darker gray polygon) have both LR+ > 10 and LR- < 0.1. The figure depicts studies that used manual scoring or combined manual and automated scoring for the type III monitor and a cutoff of 15 events/h as suggestive of sleep apnea in facility-based polysomnography.

Qualitative assessment of the concordance between measurement methods. Now, let us exemplify the alternative approach of summarizing analyses of agreement between the two methods. This approach essentially asks whether the two methods could be used interchangeably.

An obvious option is to record and summarize difference versus average analyses (Bland-Altman analyses) from the published studies. A Bland-Altman plot shows the differences between the two measurements against their average (which is the best estimate of the true unobserved value). An important concept in such analyses is the 95 percent limits of agreement. The 95 percent limits of agreement define 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 9-2).

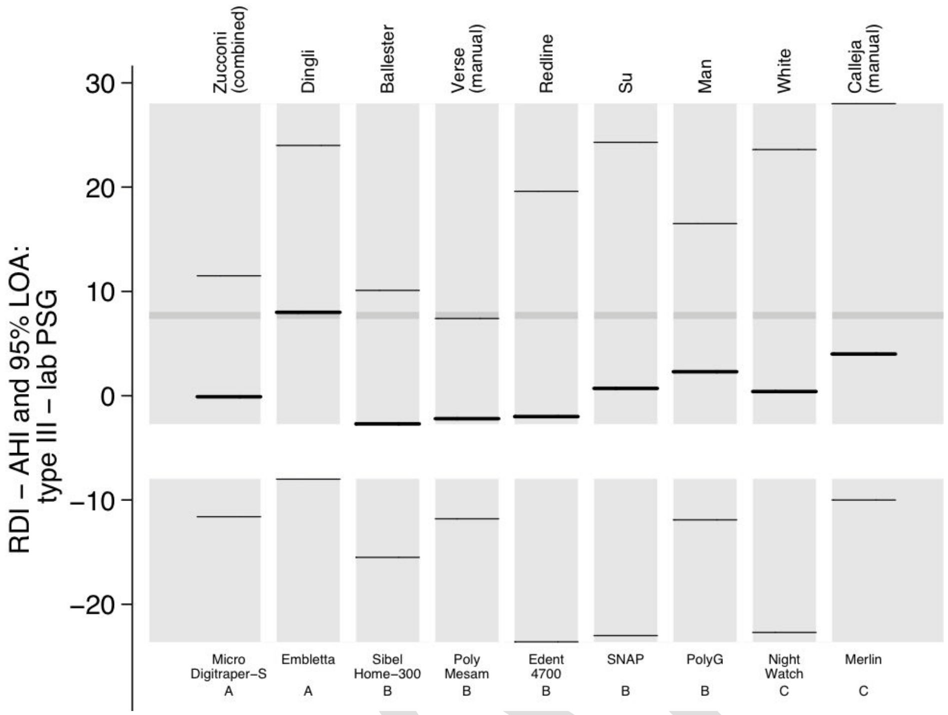
Figure 9-2. Example of a difference versus average analysis of measurements with facility-based polysomnography and type IV monitors



Digitized data from an actual study where type IV monitors (Pro-Tech® PTAF2 and Compumedics® P2) were compared with facility-based PSG.[17] 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 stand for the boundaries within which 95 percent of the differences lie. If these are very wide and encompass clinically important differences, one may concur 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 AHI or RDI levels and larger for larger AHI or RDI levels. In this example, 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 9-3 summarizes such plots across several studies. For each study, it shows the mean difference in the two measurements 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.

Figure 9-3. Schematic representation of the mean bias and limits of agreement between facility-based polysomnography and type III monitors in specialized sleep units: studies that used manual scoring for the portable monitor



Schematic representation of the agreement between portable monitors and facility-based polysomnography as conveyed by difference versus average analyses. Each study is represented by three lines; these stand for the mean bias and the 95 percent limits of agreement from the difference versus average analyses. The upper, middle, and lower gray areas group the upper 95 percent limits of agreement, the mean difference, and the lower 95 percent limits of agreement, respectively. Note that the upper and middle gray areas overlap slightly. The make of the monitor and the overall study quality are also depicted in the lower part of the graph. Only studies that used both apneas and hypopneas in the definition of respiratory events for both monitors are shown.

This is at odds with the findings of the previous analysis, which suggested the portable monitors are good in identifying people with sleep apnea. The explanation may be that the two measurements generally agree on who has 15 or more events per hour of sleep (which is a low number). They disagree on the exact measurement among people who have larger measurements on average: One method may calculate 20 events and the other 50 events per hour of sleep for the same person. Such differences can become important when the definition of sleep apnea is at a higher level (20, 30, or 40 events per hour). This means that the Bland-Altman analyses are not particularly helpful when considering the cutoff of 15 events per hour for defining sleep apnea.

Summary

Key points are:

- When dealing with the case of an imperfect reference standard, it may be preferable to use a simple description of study results.

- Because of the challenges in the interpretation of the naïve estimates for sensitivity and specificity, a typical meta-analysis may not be the best way to summarize the findings.
- There is a relative paucity of empirical data on the merits and pitfalls of statistical methods that try to adjust for the imperfect reference standard. Therefore, it does not seem prudent to recommend that EPCs implement and adopt the complicated methods for the analysis of this case.

References

1. Bossuyt PM. Interpreting diagnostic test accuracy studies. Semin Hematol 2008;45(3):189-95.

2. Reitsma JB, Rutjes AW, Khan KS, et al. 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, et al. 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, et al. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004;140(3):189-202.

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

6. Cantor T, Yang Z, Caraiani N, et al. 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.

7. 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.

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

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

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

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

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

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

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

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

16. Thompson IM, Pauler DK, Goodman PJ, 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.

17. Trikalinos TA, Ip S, Raman G, et al. Home diagnosis of obstructive sleep apnea-hypopnea syndrome. Technology Assessment. Rockville, MD: Agency for Healthcare Research and Quality. August 2007. Available at: http://www.cms.hhs.gov/determinationprocess/downloads/id48TA.pdf. Accessed July 9, 2010.

Methods Guide for Medical Test reviews

Paper 10

Decision Modeling

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

AHRQ Publication No. xx-EHCxxx-EF
<Month Year>

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, through a collaborative effort within the Evidence-based Practice Center Program, have developed a Methods Guide for Medical Test Reviews. We intend that these documents will serve as a resource for our EPCs as well as for other investigators interested in conducting systematic reviews on medical tests.

This Medical Test Methods guide is intended to be a practical guide for those who prepare and use systematic reviews on medical tests. This document complements the EPC Methods Guide on Comparative Effectiveness Reviews (http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318), which focuses on methods to assess the effectiveness of treatments and interventions. The guidance here applies the same principles for assessing treatments to the issues and challenges in assessing medical tests and 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. We provide guidance in stepwise fashion for those conducting a systematic review.

The *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 *Medical Test Methods Guide* and the Effective Health Care Program can be made at www.effectivehealthcare.ahrq.gov.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

Paper 10. Decision Modeling

Undertaking a modeling exercise requires technical expertise, good appreciation of clinical issues, and (sometimes extensive) resources, and should be pursued when it is maximally informative. In this paper, we provide practical suggestions on how to decide whether modeling is important for interpreting the findings of a systematic review of medical tests using specific examples. We do not discuss *how* to model medical testing and its downstream effects. Many excellent publications describe guidelines for good modeling practices, especially in the context of cost-effectiveness analyses.[1-7]

Common Challenges

Although many medical test evaluations are based on indirect evidence, developing a formal decision model is not always feasible. A commonly encountered challenge is to identify when it is important to perform formal decision or cost-effectiveness analyses to understand the effects of testing on patient-relevant outcomes. The remainder of this paper describes principles for determining when decision modeling is appropriate and feasible.

Principles for Addressing the Challenges

Basic Principle: Adopt a Stepwise Approach

Table 10-1 describes a stepwise approach for assessing whether models are appropriate for reviews of medical tests of different types (e.g., imaging, genetic tests, or strategies combining several tests) in various settings (screening, diagnosis, treatment guidance, prognosis, patient monitoring). Unavoidably, context- and topic-specific considerations will become relevant in practical applications.

Table 10-1. Stepwise approach to determining whether 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 test consequences and management strategy for each test result
3	Assess whether modeling will be useful
4	Evaluate prior modeling studies
5	Consider whether modeling is practically feasible

We expand on each of the above steps in the following sections.

Step 1: Define How the Test Will be Used

As described in the Introduction to this *Medical Test Methods Guide*, the PICOTS typology (Patient population, Intervention, Comparator, Outcomes, Timing, Setting) is a widely adopted formalism for establishing the context of a systematic review. 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 triages further diagnostic workup). 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 and Management Strategies

Medical tests exert most of their effects in an indirect way. Notwithstanding the emotional, cognitive, and behavioral changes conferred by testing and its results,[8] 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. In principle, test results will influence downstream clinical decisions that will eventually determine patient outcomes. From this point of view, test performance (as conveyed by sensitivity, specificity, positive and negative likelihood ratios, or other metrics) is only a surrogate endpoint.

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* Paper 2). This can form the basis for outlining a basic tree illustrating test consequences and management options that depend on test results. This exercise helps reviewers make explicit the clinical scenarios of interest, the alternate (comparator) strategies, and the assumptions they make.

Step 3. Assess Whether Modeling Will 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 reasonably be contrasted. Therefore, it may be easier to consider the opposite question; that is, whether formal modeling will *not* be 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, there is too much uncertainty in multiple central (influential) parameters.

The case of the "clear winner." For some medical testing evaluations, there may be a clearly superior strategy. In the ideal case, the test-and-treat strategies have been directly compared in well-designed and conducted randomized controlled trials (RCTs) or in properly analyzed non-randomized studies. Insofar as these studies match the PICOTS characteristics of interest (i.e., they are applicable to the clinical context of interest in the patient population of interest, evaluate all important test-and-treat strategies, inform outcomes of interest, are conclusive, and are adequately powered), a clear conclusion about the net benefit of alternate strategies could be reached without modeling. This situation is, however, exceedingly rare.

Alternatively, a “clear winner” strategy can be identified only on the basis of test performance when the link between test results and patient-relevant outcomes is very strong. The operating assumption here is that test results dictate the use of an effective treatment or avoidance of a harmful intervention. A complete discussion of this scenario is provided by Lord et al.[9-10] Briefly, a test-and-treat strategy will be dominant when it is better than all other alternate strategies (or at least as good) on the relevant decision criteria (e.g., performance characteristics), and at the same time better than all other strategies in at least some other criteria (e.g., cost or availability). For example, if two tests have identical performance characteristics (sensitivity and specificity, the “decision criteria”), but one of them is cheaper or poses less inconvenience or risk to the patient, the cheaper and safer test should be preferred. For example, Doppler ultrasonography and venography have similar sensitivity and specificity to detect the treatable condition of symptomatic distal deep venous thrombosis;[11] the Doppler test is, however, easier, faster, and non-invasive, and therefore preferable.

In a related scenario, a test may have better performance characteristics than its comparator. Following Lord et al.,[9-10] we distinguish two cases:

1. In the simplest case, the sensitivities are comparable but the specificities differ. In this case, one should pick the test with the better specificity because it avoids the harms of unnecessary further testing or treatment.
2. If the sensitivity of one test is better and the specifities are comparable, then one has to judge whether the additionally detected cases are likely to respond to treatment in the same way as the cases detected by the other test. If this is the case, one can prefer the test with the better sensitivity. If this is not the case, one cannot confidently prefer the test with the better sensitivity, because the additionally detected cases may represent a change in the spectrum of disease, and this in turn affects the effectiveness of downstream interventions.
    a. In the best case, RCTs document that the additionally detected cases respond to treatment. For example, RCTs have shown that estrogen receptor status predicts response to adjuvant tamixifen for breast cancer,[12] suggesting that a strategy that tests for estrogen receptor status to triage treatment is preferable to not testing.
    b. We can reasonably extrapolate that treatment effectiveness will be unaltered in the additional identified cases when the tests operate on the same principle, and the clinical and biological characteristics of the additional identified cases remain unaltered. An (extreme) example is computed tomography colonography for detection of large polyps. Dual positioning (prone and supine) of patients is more sensitive than supine-only positioning, without differences in specificity.[13] It is very reasonable to consider that the additional cases detected by dual positioning will respond to treatment in the same way as the cases detected by supine-only positioning.
3. In all other cases, one has to assess trade-offs in a decision analysis.[9-10]

The case of excessive uncertainty. 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 may not be helpful in illuminating answers to the key questions of the systematic review. Arguably, a modeling exercise that examines many plausible alternatives and conducts wide-ranging sensitivity analyses can still be useful. It can help explore the major factors that contribute to our uncertainty, should that be considered a goal of the project. In fact, modeling that includes value-of-information analyses may be particularly useful in such cases.[14-15]

Step 4. Evaluate Prior Modeling Studies

Prior to developing a model *de novo*, 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, one has 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.[1-7] 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.[16] MISCAN-COLON[17-18] and SimCRC[19] are two microsimulation models describing the natural history of colorectal cancer. Both assume an adenoma-carcinoma sequence for cancer development, but they 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;[18,20] and to approximately 30 years in SimCRC. Because of such esoteric differences, models can result in different conclusions. Ideally, simulation models should be validated against independent datasets that are comparable to the datasets on which the models were developed.[16] External validation is particularly important for simulation models in which the unobserved deep parameters are set without calibration (based on assumptions and analytical calculations).[16-17]

Second, once the systematic reviewers deem that good-quality models exist, they have to examine whether those models are applicable to the interventions and populations of the current evaluation; that is, whether they match the PICOTS items of the systematic review. In addition, the reviewers have to judge whether methodological and epidemiological challenges have been adequately addressed by the model developers.[21]

Third, reviewers have 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 from the systematic review. Also, consideration should be given to whether 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 existing modeling papers meet these three challenges, then synthesizing this literature may suffice. Alternatively, developing a new model may be considered, or one 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 have followed this practice for specific topics. For example, the USPSTF recommendations for colonoscopy screening[22] were informed by simulations based on the

MISCAN-COLON microsimulation model,[23] which was developed outside the EPC program.[17-18]

Step 5. Consider Whether Modeling is Practically Feasible

Even if a modeling has been determined to be useful, it may still not be feasible to develop a formal model within the context of a systematic review. Time and budgetary constraints, lack of experienced personnel, and other factors may all play a role in limiting the feasibility of developing or adapting a model to answer the relevant questions. 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 readily available or may not be available at all. An important point to note in this context is that the literature required for developing or adapting a model does not necessarily overlap with that used for an evidence report.

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 is not feasible within the context of the systematic review, consideration should be given to whether this effort could be undertaken sequentially as a related but distinct project. The systematic review could summerize available evidence, identify gaps, and and provide estimates of many necessary parameters for a model. The systematic review can also include a call for the development of a model in the future recommendations section. A subsequent report that uses modeling could then inform long-term outcomes.

Illustration

We illustrate the application of the first three steps in determining whether a model should be included in a systematic review with reference to an AHRQ-sponsored report on the ability of positron emission tomograhy (PET) to guide the management of suspected Alzheimer's dementia (AD), a progressive neurodegenerative disease for which current treatment options are at best modestly effective.[24] 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 summarized the diagnotic performance of PET to identify AD, but found no longitudinal studies or RCTs on the effects of PET testing on disease progression, mortality or other clinical outcomes. The reviewers deemed that decision modeling was appropriate to contextualize the information on test performance by linking test results to long-term patient-relevant outcomes. Modeling was also used to explore whether conclusions would differ if the available treatment options were more effective than the currently available.

In Step 1, reviewers defined how PET would be used. The complete PICOTS specification for the PET example is described in the evidence report[24] and is not reviewed in detail here. In brief, the focus was on *diagnosis* of the disease (AD) among the three scenarios of patients with suggestive symptoms.

AD is typically diagnosed with a clinical exam that includes complete history, physical and neuropsychiatric evaluation, and screening laboratory testing.[25] Reviewers were only interested in PET as a "confirmatory" test; that is, they were interested in adding PET to the usual diagnostic workup. They explicitly did not evaluate patient management strategies where PET is the only test (i.e., PET "replaces" the typical exam) or where PET triages who will receive the clinical exam (an unrealistic scenario). In this particular case, PET is used as an add-on to a clinical exam for *diagnosing* patients with different severities or types of AD (mild or moderate AD, mild cognitive impairment, family history of AD) and is compared against the clinical exam alone (*no PET* as an add-on test). Table 10-2 classifies the results of PET testing.

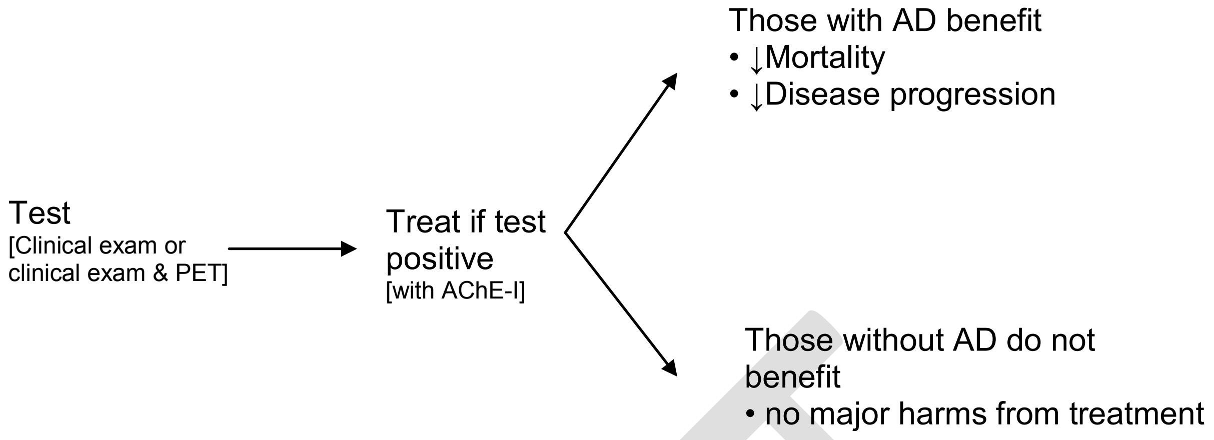
Table 10-2. Cross tabulation of PET results and actual clinical status among patients with initial clinical examination suggestive of Alzheimer's

	AD in long-term clinical evaluation	No AD in long-term clinical evaluation
PET suggestive of AD	"True positive"	"False negative"
PET not suggestive of AD	"False positive"	"True negative"

Abbreviations: AD = Alzheimer's disease; PET = positron emission tomography.
Entries 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.

In Step 2, reviewers further elaborated on the test-and-treat strategies of interest by constructing a simplified analytic framework and outlining patient management options conditional on test results. They assumed that there are no appreciable direct effects or direct complications of testing (with or without PET). A simplified ("straw man") version of the analytic framework is depicted in Figure 10-1. The "straw man" analytic framework conveys that the anticipated effects of PET testing on mortality and disease progression are only indirect and are exclusively conferred through the downstream clinical decision of whether to treat patients. In the clinical scenario of interest, patients with a positive test result (either by clinical exam or by the clinical exam-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.

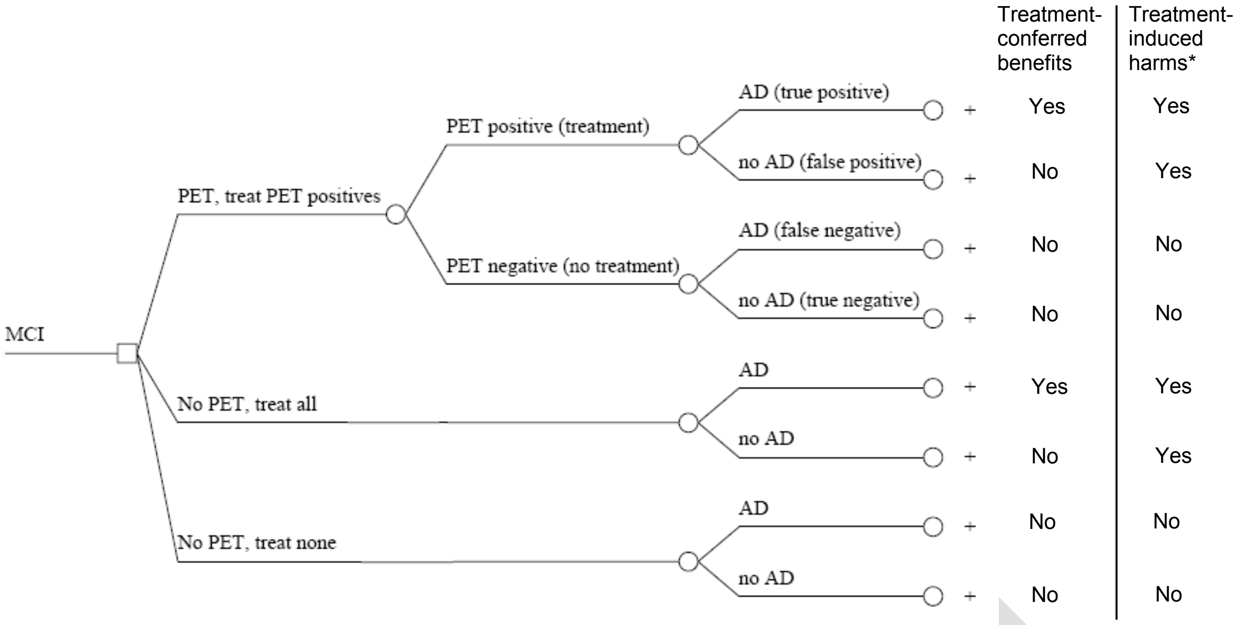
Figure 10-1. Simplified ("straw man") analytic framework


Those with AD benefit
• ↓Mortality
• ↓Disease progression
Test
[Clinical exam or clinical exam & PET]
Treat if test positive
[with AChE-I]
Those without AD do not benefit
• no major harms from treatment

Abbreviations: AD = Alzheimer's disease; AChE-I: acetylcholinestrase inhibitors (the treatment available at the time of the report); PET = positron emission tomography.
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, and for the clinical scenario of people with moderate cognitive impairment (MCI) in the initial clinical exam (scenario B). 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); describe the consequences of being classified as a true positive, true negative, false positive or false negative; and make explicit which test-and-treat strategies are being compared. This simplified outline provides a bird's-eye-view of a decision tree for the specific clinical test.

Figure 10-2. Management options for MCI



*When applicable (harms were not important in the main analyses of the evidence report).
Abbreviations: AD = Alzheimer's disease; MCI = moderate cognitive impairment; PET = positron emission tomography.
As per the evidence report, the then-available treatment options (achetylcholinesterase inhibitors) do not have important adverse effects. However, in other cases, harms can be induced both by the treament 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 are 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 involved assessing whether modeling could be useful. In the PET example, there were no data clearly linking one test-and-treat strategy to improved outcomes over the other strategies, but there are sufficient data to make reasonable estimates of the benefits and harms of pharmacologic therapy in those with and without AD. Conclusions on the superiority of either strategy could not be drawn based only on the performance characteristics of the tests, and therefore a model was developed and analyzed as part of the evidence report. The model offered insights not only on PET but on imaging for AD in general.

Steps 4 and 5 need not be illlustrated.

Summary

Key points are:

- In systematic review of medical tests, it can be illuminating to develop and evaluate a formal decision model that links the decision to use a particular test-and-treat strategy versus appropriate alternate strategies to the likely patient-relevant outcomes.
- In addition to helping identify preferred test strategies, modeling can provide insight into the dynamic interplay of various factors on decision-relevant effects, in turn leading to recommendations for further studies.

- A five-step algorithm can help EPCs evaluate whether modeling is appropriate for the interpretation of a systematic review of medical tests:
  Step 1. Define how the test will be used.
  Step 2. Use a framework to identify test consequences and management strategy for each test result.
  Step 3. Assess whether modeling will be useful.
  Step 4. Evaluate prior modeling studies.
  Step 5. Consider whether modeling is practically feasible.

References

1. 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):443-4.

2. Richardson WS, Detsky AS, Evidence-Based Working Group. 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? JAMA 1995;273(20):1610-3.

3. Richardson WS, Detsky AS, Evidence-Based Working Group. Users' guides to the medical literature. VII. How to use a clinical decision analysis. A. Are the results of the study valid? JAMA 1995;273(16):1292-5.

4. Philips Z, Ginnelly L, Sculpher M, 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.

5. Philips Z, Bojke L, Sculpher M, et al. Good practice guidelines for decision-analytic modelling in health technology assessment: a review and consolidation of quality assessment. Pharmacoeconomics 2006;24(4):355-71.

6. 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):461-77.

7. Weinstein MC, O'Brien B, Hornberger J, 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. Bossuyt PM, McCaffery K. Additional patient outcomes and pathways in evaluations of testing. Med Decis Making 2009;29(5):E30-8.

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. 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.

11. Gottlieb RH, Widjaja J, Tian L, et al. Calf sonography for detecting deep venous thrombosis in symptomatic patients: experience and review of the literature. J Clin Ultrasound 1999;27(8):415-20.

12. Early Breast Cancer Trialists' Collaborative Group. Tamoxifen for early breast cancer: an overview of the randomised trials Lancet 1998;351(9114):1451-67.

13. Fletcher JG, Johnson CD, Welch TJ, et al. Optimization of CT colonography technique: prospective trial in 180 patients. Radiology 2000;216(3):704-11.

14. Janssen MP, Koffijberg H. Enhancing value of information analyses. Value Health 2009.

15. Oostenbrink JB, Al MJ, Oppe M, et al. Expected value of perfect information: an empirical example of reducing decision uncertainty by conducting additional research. Value Health 2008;11(7):1070-80.

16. Karnon J, Goyder E, Tappenden P, et al. A review and critique of modelling in prioritising and designing screening programmes. Health Technol Assess 2007;11(52):iii-xi, 1.

17. Habbema JD, van Oortmarssen GJ, Lubbe JT, et al. The MISCAN simulation program for the evaluation of screening for disease. Comput Methods Programs Biomed 1985;20(1):79-93.

18. Loeve F, Boer R, van Oortmarssen GJ, et al. The MISCAN-COLON simulation model for the evaluation of colorectal cancer screening. Comput Biomed Res 1999;32(1):13-33.

19. National Cancer Institute. Cancer Intervention and Surveillance Modeling Network. Available at: http://cisnet.cancer.gov/. Accessed September 4, 2009.

20. Loeve F, Brown ML, Boer R, et al. Endoscopic colorectal cancer screening: a cost-saving analysis. J Natl Cancer Inst 2000;92(7):557-63.

21. 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):E22-9.

22. Screening for colorectal cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 2008;149(9):627-37.

23. Zauber AG, Lansdorp-Vogelaar I, Knudsen AB, et al. Evaluating test strategies for colorectal cancer screening: a decision analysis for the U.S. Preventive Services Task Force. Ann Intern Med 2008;149(9):659-69.

24. Matchar DB, Kulasingam SL, McCrory DC, et al. Use of Positron Emission Tomography and other neuroimaging techniques in the diagnosis and management of Altzheimer's disease and dementia. Technology Assessment. Prepared by the Duke Evidence-based Practice Center (under Contract No. 290-97-0014). December 2001. Rockville, MD: Agency for Healthcare Research and Quality. Available at: http://www.cms.hhs.gov/determinationprocess/downloads/id9TA.pdf. Accessed September 4, 2009.

25. Knopman DS, DeKosky ST, Cummings JL, 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):1143-53.

Methods Guide for Medical Test reviews

Paper 11

Genetic Tests as Predictive Indicators

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

AHRQ Publication No. xx-EHCxxx-EF
<Month Year>

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, through a collaborative effort within the Evidence-based Practice Center Program, have developed a Methods Guide for Medical Test Reviews. We intend that these documents will serve as a resource for our EPCs as well as for other investigators interested in conducting systematic reviews on medical tests.

This Medical Test Methods guide is intended to be a practical guide for those who prepare and use systematic reviews on medical tests. This document complements the EPC Methods Guide on Comparative Effectiveness Reviews (http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318), which focuses on methods to assess the effectiveness of treatments and interventions. The guidance here applies the same principles for assessing treatments to the issues and challenges in assessing medical tests and 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. We provide guidance in stepwise fashion for those conducting a systematic review.

The *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 *Medical Test Methods Guide* and the Effective Health Care Program can be made at www.effectivehealthcare.ahrq.gov.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

Paper 11. Genetic Tests as Predictive Indicators

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 in the degree to which certain issues are relevant, particularly when considering genetic test results that provide predictive rather than diagnostic information. This paper 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. In this paper, we reflect on genetic tests used to (1) determine risk or susceptibility in asymptomatic individuals (to identify individuals at risk for future health conditions, e.g., 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); or (3) predict response to treatments or environmental factors including diet (nutrigenomics), drugs (pharmacogenomics, e.g., 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.

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 paper, we discuss common challenges and principles for addressing those challenges related to (1) developing the topic and structuring a genetic test review (context and scoping), and (2) performing the review. 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.

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, which is related to time lag to outcomes, variable expressivity, and pleiotropy. These particular aspects of genetic tests result in specific actions at various stages of planning and performing the review.

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.”[2] 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.[3-4] This can be due to the lack of reliable prevalence data or a lack 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 be helpful to develop estimates.[3]

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 and followup is sufficient to characterize the relationship between a positive test 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 would be 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.[2] 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. Of note, penetrance, variable expressivity, and pleiotropy are terms generally used to describe autosomal dominant, single gene disorders.

Other Common Challenges

Another common challenge in evaluating predictive genetic tests is that direct evidence for the impact of the test results on health outcomes is often lacking. The evidence base may often be 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. Most notably, there may be insufficient published data related to the analytic validity of some genetic tests.

Genetic tests also have a number of technical issues that are particularly relevant to assessing 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.

Common challenges arise when attempting to use genetic tests to determine susceptibility or risk in asymptomatic individuals. The utility of such tests may depend on the ability of individuals (e.g., patients or health care providers) to report and identify certain clinical factors.

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.

Principles for Addressing the Challenges

Principle 1: Use an Organizing Framework Appropriate for Genetic Tests

Organizing frameworks for approaching the evaluation of genetic tests have been developed by the United States Preventive Services Task Force (USPSTF), the CDC, and EGAPP.[1,5-6] The model endorsed by the EGAPP initiative[1] was based on a previous Task Force report[7] 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.[5] 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 11-1). 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]

Table 11-1. ACCE model questions for reviews of genetic tests[5]

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?

Element	Questions
	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?

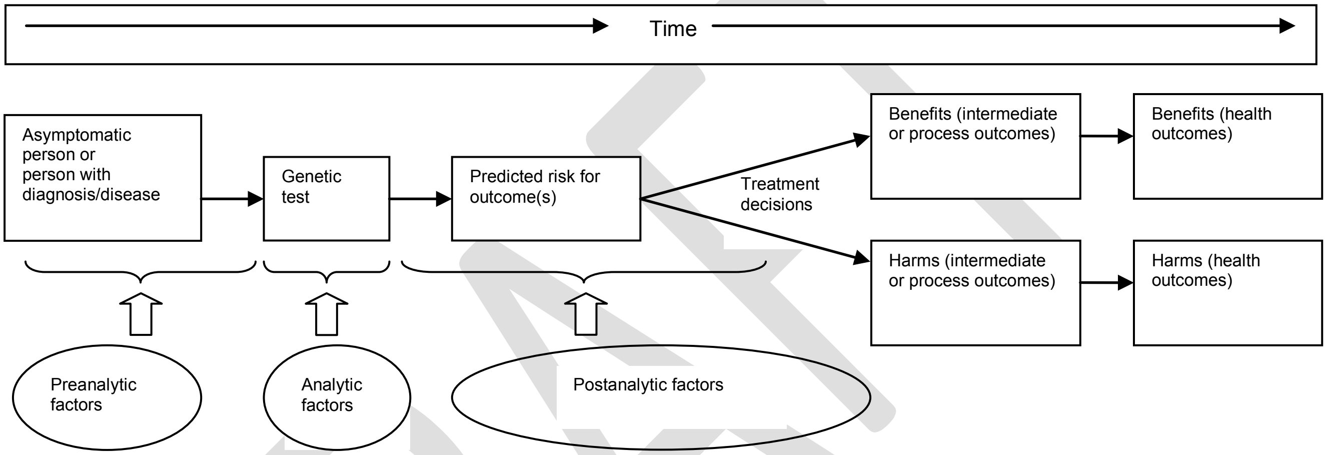
Abbreviations: ACCE = Analytic validity, Clinical validity, Clinical utility, and Ethical, legal and social implications; DNA = deoxyribonucleic acid.

Principle 2: Analytic Frameworks Should Reflect 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 used for some other purpose. 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.

In constructing an analytic framework for evaluating a genetic test, it may be useful for analysts to consider preanalytic, analytic, and postanalytic factors particularly applicable to genetic tests (described later in this paper), as well as the key outcomes of interest. Analytic frameworks should incorporate the factors and outcomes of greatest interest to decisionmakers. Figure 11-1 illustrates a generic analytic framework for evaluating predictive genetic tests. This framework can be modified as necessary for a variety of situations.

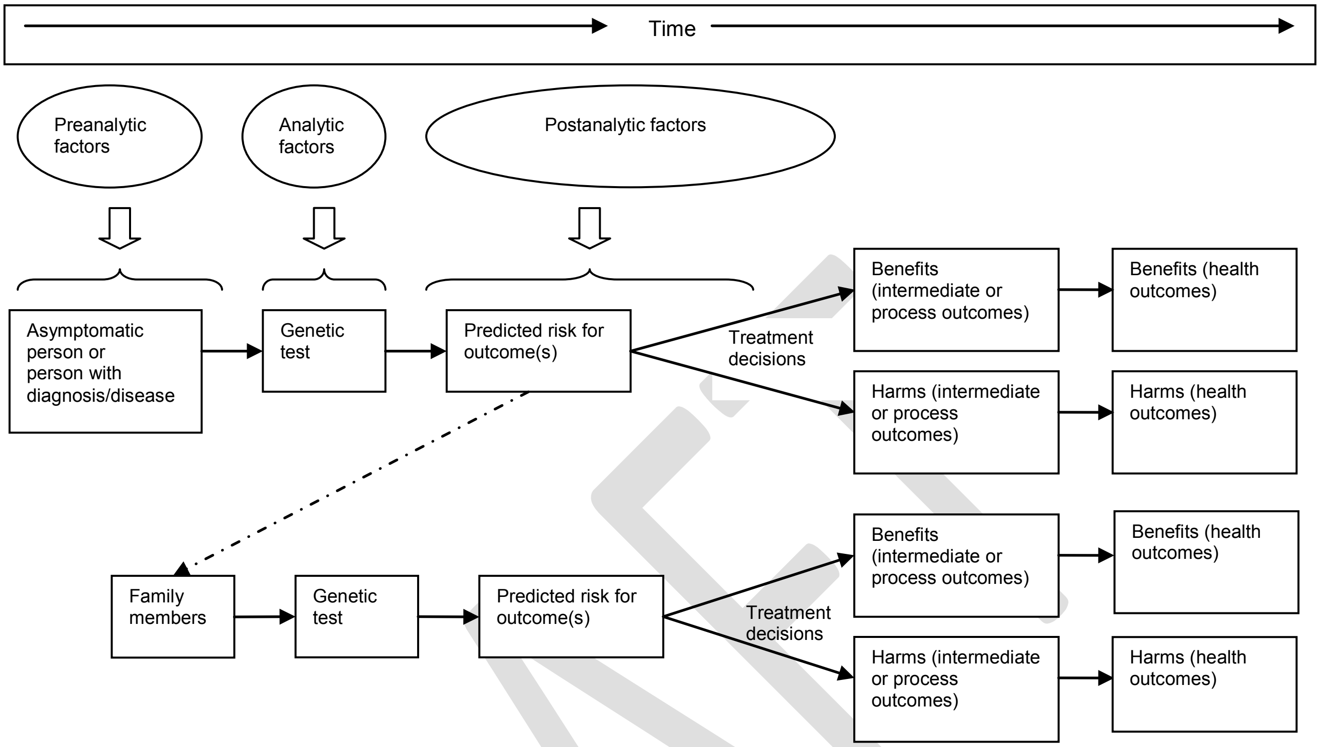
Figure 11-1. Generic analytic framework for evaluating predictive genetic tests


Time
Asymptomatic person or person with diagnosis/disease
Genetic test
Predicted risk for outcome(s)
Treatment decisions
Benefits (intermediate or process outcomes)
Benefits (health outcomes)
Harms (intermediate or process outcomes)
Harms (health outcomes)
Preanalytic factors
Analytic factors
Postanalytic factors

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 those test results that 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/variations 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.

Depending on the context, the impact of genetic testing on family members may be important. 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


Time
Preanalytic factors
Analytic factors
Postanalytic factors
Asymptomatic person or person with diagnosis/disease
Genetic test
Predicted risk for outcome(s)
Treatment decisions
Benefits (intermediate or process outcomes)
Benefits (health outcomes)
Harms (intermediate or process outcomes)
Harms (health outcomes)
Family members
Genetic test
Predicted risk for outcome(s)
Treatment decisions
Benefits (intermediate or process outcomes)
Benefits (health outcomes)
Harms (intermediate or process outcomes)
Harms (health outcomes)

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.[8]

When assessing the gray literature, U.S. Food and Drug Administration (FDA)-approved test package inserts contain data summaries of the analytic validity data that were instrumental in gaining FDA clearance for marketing or approval. 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.[9-11]

An AHRQ “horizon scan” found two databases—the LexisNexis® database (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.[12]

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.[13-14] In general, preanalytic steps are those involved in obtaining, fixing or preserving, and storing samples prior to staining and analysis. Preanalytic factors relevant to predictive genetic tests may also include a person's age, sex, ethnicity/race or ancestry, parental history of consanguinity, and family health history. Important analytic variables include the type of assay chosen and its reliability, types of samples, the specific analyte investigated (e.g., specification of which alleles, genes, or biochemical analytes were evaluated), 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.[13-14] Comparative effectiveness review teams should include or consult with molecular pathologists, geneticists, or others familiar with the issues related to the genetic tests being evaluated to determine which of these technical issues are pertinent for a given review. Table 11-2 summarizes some of the preanalytic, analytic, and postanalytic questions that should be addressed.

Table 11-2. 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 (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?

*Adapted from Burke et al., 2002[13] and Little et al., 2002.[14]

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 sensitivity and specificity. Functional assays may have the advantage of showing the functional significance of an underlying genetic polymorphism and, thus, may provide more important information. 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. Contrasting results between studies using DNA-based molecular methods and those using phenotypic assays have been reported.[14-16]

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.[14,17-21] 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).[14] Some studies utilize different sources of DNA for cases and controls, introducing potential measurement bias from differences in the 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).[14]

When evaluating DNA-based molecular tests, the complexity of the test method, laboratory-to-laboratory variability, and quality control should all be assessed. A number of methods are available for genotyping single nucleotide polymorphisms that vary in complexity and potential for polymorphism misclassification.[14,22-24] 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.[14,25]

Principle 6: Case-control Studies Should Be Carefully Evaluated 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 is important since the frequencies of DNA polymorphisms vary from population to population (i.e., population stratification). 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.[14-15,26-28] In general, only nested case-control studies could have low enough potential for selection bias to provide reliable information.

Principle 7: Determine Added Value of Genetic Test Over Existing Risk Models

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 (to a low-density lipoprotein of 160, 130, or 100) is based on these factors.[29] Assessment of newly identified polymorphisms—such as those described on chromosome 9p21[30]—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 compared with using clinical factors alone.

Principle 8: Studies of Genetic Tests Have Particular Statistical Issues

Hardy-Weinberg equilibrium. Most allele distributions follow a usual distribution, known as 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.[31] 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).[32] Although it is not specific and possibly not sensitive, deviation from HWE may be a clue to bias and genotyping error.[32] 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.[31-32]

Sample size calculations. When assessing the internal validity of studies of genetic tests, 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.[33] 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.[33]

Genetic association studies. 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 launch further understanding of the importance of genetics in relation to a variety of health conditions but should generally be considered hypothesis generating rather than hypothesis testing or confirming cause-effect relationships.[14] 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, it should be noted whether appropriate adjustments for multiple comparisons were used. Many recommend using a P value of less than 5 x 10^{-8} for the threshold of significance in large genome-wide studies.[32,34-35]

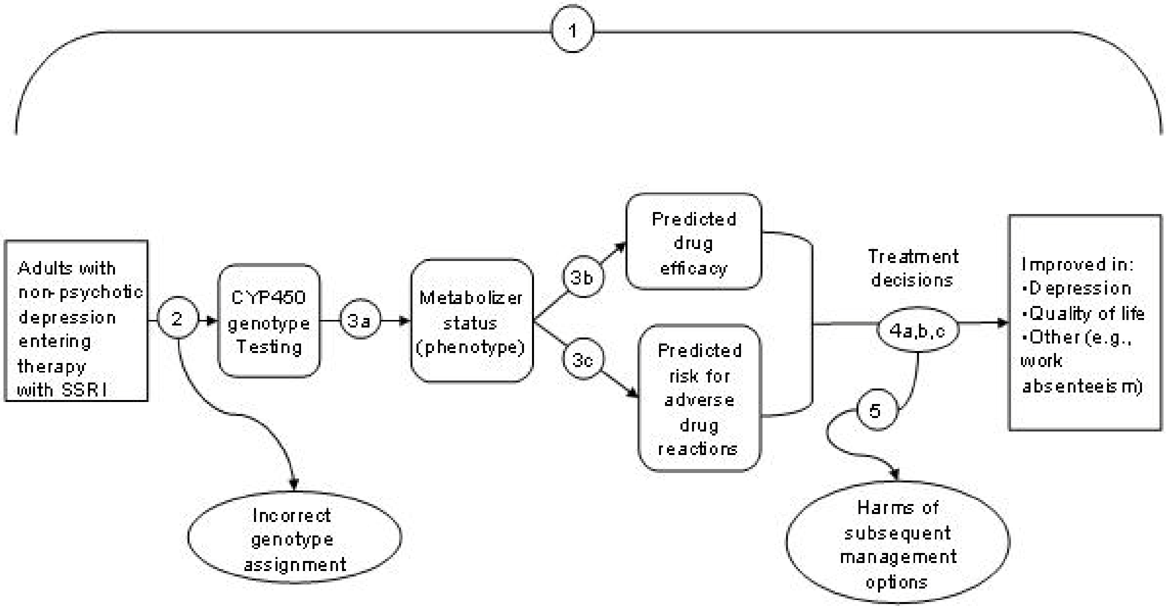
When a genetic mutation associated with increased risk is present, evaluating potential causality can be difficult as many other factors may influence associations. These include environmental exposures, behaviors, and other genes. Many genetic variants identified that are thought to influence susceptibility to diseases are associated with low relative and absolute risk.[14,36] 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.[14]

Overlapping data sets. Identifying studies for comparative effectiveness reviews may sometimes result in publications regarding prevalence of genetic variants that arose from overlapping data sets.[14] 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.[2] 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.[14]

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 many fewer studies of the clinical utility of genetic test results. 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).[37] The authors found no prospective studies of the use of genotyping to guide treatment that measured clinical outcomes. 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 genotyping test for SSRI treatment of depression.[37] Numbers refer to the key questions addressed.


1
Adults with non-psychotic depression entering therapy with SSRI
2
CYP450 genotype Testing
3a
Metabolizer status (phenotype)
3b
3c
Predicted drug efficacy
Predicted risk for adverse drug reactions
Treatment decisions
4a,b,c
5
Improved in:
•Depression
•Quality of life
•Other (e.g., work absenteeism)
Incorrect genotype assignment
Harms of subsequent management options

Abbreviation: SSRI = selective serotonin reuptake inhibitor.

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.[38]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 (KQ1. 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, 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).[3,39] 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 for hereditary susceptibility to breast and ovarian cancer.[3] Ideally, following this paradigm, genetic testing should begin in an affected family member suspected of having a hereditary susceptibility (due to an early age at onset, multifocal disease, or their position in the pedigree). If a deleterious mutation is identified, then testing of at-risk family members for this familial mutation will be highly informative; if the mutation is identified in family members, the cancer risk can be better defined and risk-appropriate interventions can be recommended. If a familial mutation is not found, then testing of at-risk relatives is not generally recommended. However, often an affected family member is not available for genetic testing. In these cases, testing of an at-risk relative first is an option, although interpreting results of such testing can be complex, particularly when a pathogenic mutation is not found or when a variant of uncertain significance is identified. In these instances, it is imperative that the test result interpretation and overall cancer risk assessment are considered within the context of an individual's personal and family history risk factors.

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. Sensitivity and specificity of self-reported family history are important in determining overall usefulness of predictive genetic testing.[3]

Summary

Key points are:

- 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.
- It is important to have a clear definition of the clinical scenario and an analytic framework when evaluating *any* test, including predictive genetic tests.
- Organizing frameworks, such as the ACCE model, and analytic frameworks are useful constructs for approaching the evaluation of genetic tests.
- In constructing an analytic framework for evaluating a genetic test, analysts should consider preanalytic, analytic, and postanalytic factors; such factors are useful when evaluating analytic validity.
- Predictive genetic tests are characterized by a delayed time between testing and clinically important events.
- It may be difficult to find published information on the analytic validity of some genetic tests. Web sites (FDA or diagnostic companies) and gray literature may be important sources.
- 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.
- Analysts should consider whether studies have tested for and reported Hardy-Weinberg equilibrium.
- 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, it should be noted 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 2009;11(1):3-14.

2. U.S. National Library of Medicine. What are reduced penetrance and variable expressivity? August 2009. In: Genetics Home Reference Handbook. Available at: http://ghr.nlm.nih.gov/handbook/inheritance/penetranceexpressivity. Accessed July 15, 2010.

3. Nelson HD, Huffman LH, Fu R, et al. Genetic risk assessment and BRCA mutation testing for breast and ovarian cancer susceptibility: systematic evidence review for the U.S. Preventive Services Task Force. Ann Intern Med 2005;143(5):362-79.

4. Whitlock EP, Garlitz BA, Harris EL, et al. Screening for hereditary hemochromatosis: a systematic review for the U.S. Preventive Services Task Force. Ann Intern Med 2006;145(3):209-23.

5. 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.

6. 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.

7. Task Force on Genetics in Disease Prevention. Promoting Safe and Effective Genetic Testing in the United States. Final Report of the Task Force on Genetic Testing created by the National Institutes of Health-Department of Energy Working Group on Ethical, Legal and Social Implications of Human Genome Research. Edited by Holtzman NA and Watson MS. September 1997. Available at: http://www.genome.gov/10001733. Accessed September 4, 2009.

8. Khoury MJ, Dorman JS. The Human Genome Epidemiology Network. Am J Epidemiol 1998;148(1):1-3.

9. Palomaki GE, Bradley LA, Richards CS, et al. Analytic validity of cystic fibrosis testing: a preliminary estimate. Genet Med 2003;5(1):15-20.

10. Palomaki GE, Haddow JE, Bradley LA, et al. Updated assessment of cystic fibrosis mutation frequencies in non-Hispanic Caucasians. Genet Med 2002;4(2):90-4.

11. Palomaki GE, Haddow JE, Bradley LA, et al. Estimated analytic validity of HFE C282Y mutation testing in population screening: the potential value of confirmatory testing. Genet Med 2003;5(6):440-3.

12. Agency for Healthcare Research and Quality. Technology Assessment: Genetic Tests for Cancer. Rockville, MD: Agency for Healthcare Research and Quality; January 2006. Available at: www.ahrq.gov/clinic/ta/gentests/gentests.pdf Accessed July 21, 2010.

13. Burke W, Atkins D, Gwinn M, et al. Genetic test evaluation: information needs of clinicians, policy makers, and the public. Am J Epidemiol 2002;156(4):311-8.

14. Little J, Bradley L, Bray MS, et al. Reporting, appraising, and integrating data on genotype prevalence and gene-disease associations. Am J Epidemiol 2002;156(4):300-10.

15. Brockton N, Little J, Sharp L, et al. N-acetyltransferase polymorphisms and colorectal cancer: a HuGE review. Am J Epidemiol 2000;151(9):846-61.

16. d'Errico A, Malats N, Vineis P, et al. Review of studies of selected metabolic polymorphisms and cancer. IARC Sci Publ 1999(148):323-93.

17. Yang M, Hendrie HC, Hall KS, et al. Improved procedure for eluting DNA from dried blood spots. Clin Chem 1996;42(7):1115-6.

18. Gale KB, Ford AM, Repp R, et al. Backtracking leukemia to birth: identification of clonotypic gene fusion sequences in neonatal blood spots. Proc Natl Acad Sci U S A 1997;94(25):13950-4.

19. Walker AH, Najarian D, White DL, et al. Collection of genomic DNA by buccal swabs for polymerase chain reaction-based biomarker assays. Environ Health Perspect 1999;107(7):517-20.

20. Harty LC, Shields PG, Winn DM, et al. Self-collection of oral epithelial cell DNA under instruction from epidemiologic interviewers. Am J Epidemiol 2000;151(2):199-205.

21. 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.

22. Hixson JE, Vernier DT. Restriction isotyping of human apolipoprotein E by gene amplification and cleavage with HhaI. J Lipid Res 1990;31(3):545-8.

23. Tobe VO, Taylor SL, Nickerson DA. Single-well genotyping of diallelic sequence variations by a two-color ELISA-based oligonucleotide ligation assay. Nucleic Acids Res 1996;24(19):3728-32.

24. Lee LG, Connell CR, Bloch W. Allelic discrimination by nick-translation PCR with fluorogenic probes. Nucleic Acids Res 1993;21(16):3761-6.

25. 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.

26. Botto LD, Yang Q. 5,10-Methylenetetrahydrofolate reductase gene variants and congenital anomalies: a HuGE review. Am J Epidemiol 2000;151(9):862-77.

27. Dorman JS, Bunker CH. HLA-DQ locus of the human leukocyte antigen complex and type 1 diabetes mellitus: a HuGE review. Epidemiol Rev 2000;22(2):218-27.

28. Cotton SC, Sharp L, Little J, et al. Glutathione S-transferase polymorphisms and colorectal cancer: a HuGE review. Am J Epidemiol 2000;151(1):7-32.

29. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. Circulation 2002;106(25):3143-421.

30. 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.

31. 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.

32. 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.

33. Williams JA, Johnson K, Paulauskis J, et al. So many studies, too few subjects: establishing functional relevance of genetic polymorphisms on pharmacokinetics. J Clin Pharmacol 2006;46(3):258-64.

34. Hoggart CJ, Clark TG, De Iorio M, et al. Genome-wide significance for dense SNP and resequencing data. Genet Epidemiol 2008;32(2):179-85.

35. McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 2008;9(5):356-69.

36. Caporaso N. Selection of candidate genes for population studies. In: Vineis P, Malats N, Lang M, et al., eds. *Metabolic polymorphisms and susceptibility to cancer*. Lyon, France: IARC Monogr Eval Carcinog Risks Hum; 1999:23-36.

37. 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. (Prepared by the Duke Evidence-based Practice Center under Contract No. 290-02-0025.) AHRQ Publication No. 07-E002. Rockville, MD: Agency for Healthcare Research and Quality. January 2007 Available at: http://www.ahrq.gov/downloads/pub/evidence/pdf/cyp450/cyp450.pdf. Accessed July 21, 2010.

38. 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 July 21, 2010.

39. Nelson HD, Huffman LH, Fu R, et al. Genetic Risk Assessment and BRCA Mutation Testing for Breast and Ovarian Cancer Susceptibility: Evidence Synthesis Number 37. (Prepared by the Oregon Evidence-based Practice Center under Contract No. 290-02-0024.) Rockville, MD: Agency for Healthcare Research and Quality. September 2005. Available at: http://www.ahrq.gov/downloads/pub/prevent/pdfser/brcagensyn.pdf. Accessed October 10, 2010.

Methods Guide for Medical Test reviews

Paper 12

Prognostic Tests

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

AHRQ Publication No. xx-EHCxxx-EF
<Month Year>

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, through a collaborative effort within the Evidence-based Practice Center Program, have developed a Methods Guide for Medical Test Reviews. We intend that these documents will serve as a resource for our EPCs as well as for other investigators interested in conducting systematic reviews on medical tests.

This Medical Test Methods guide is intended to be a practical guide for those who prepare and use systematic reviews on medical tests. This document complements the EPC Methods Guide on Comparative Effectiveness Reviews (http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318), which focuses on methods to assess the effectiveness of treatments and interventions. The guidance here applies the same principles for assessing treatments to the issues and challenges in assessing medical tests and 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. We provide guidance in stepwise fashion for those conducting a systematic review.

The *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 *Medical Test Methods Guide* and the Effective Health Care Program can be made at www.effectivehealthcare.ahrq.gov.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

Paper 12. Prognostic Tests

Methods specific to conducting a systematic review of a prognostic test are not well established. Generally speaking, many of the challenges confronted when reviewing evidence for prognostic tests—including use of the PICOTS typology and an analytic framework to develop the topic and structure the review, conducting a thorough literature search, assessing the quality of reported studies, extracting and summarizing varying types of statistics from clinical trials and observational studies, and modeling—have been discussed in relation to medical tests in general in other papers in this *Medical Test Methods Guide*. However, there are several important differences that should be considered before applying these methods to reviews of prognostic tests. This paper highlights some principal differences that should be considered when planning and conducting reviews of prognostic tests. It is organized by the general sequence of tasks required in such a review.

Steps Involved in Conducting a Prognostic Test Review

Step 1: Developing the Topic and Structuring the Review

Fundamental differences in perspectives and reference tests between diagnostic and prognostic tests should be considered when structuring the review. 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 measure of the true presence or absence the disease, based on a reference standard, and classify patients 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 estimate a patient's likelihood of developing a disease or experiencing a medical event over time, and the reference "test" is the observed probability of developing what is being predicted within prognostic groups defined by predicted probabilities that are estimated using the prognostic test. In some contexts it may be useful to categorize subjects as those who do or do not experience the outcome being predicted during a specfied time interval that is appropriate for the questions being addressed by a review. When this is done, key questions for systematic reviews of a prognostic test and comparisons thereof could assess the accuracy of the prognostic test in a way similar to assessments for diagnostic tests. However, the predictive accuracy of a prognostic test might be more appropriately evaluated as differences between the observed and predicted outcome probabilities within prognostic groups. These differences can vary over time, thus reviews must clearly define the period of interest.

Whatever the mode of presentation, the challenge in developing the topic and structuring a review of prognostic tests is to identify how prognistic information is proposed (or expected) to lead to different courses of action; this understanding should then be reflected in the key questions and analytic framework. The way prognostic information is proposed to relate to decisions, actions, and outcomes should guide the approach to the literature. For example, if there are three potential actions that might follow testing (e.g., no further evaluation, definitive testing with treatment based on positive findings at that stage, or treat based on the initial test findings alone), then it would be useful to consider up to three test categories (e.g., "low", "intermediate" or "high" risk). If a decision model is used as the framework for a systematic review of a prognostic test and meta-analysis, the estimated predicted probabilites, their precision, and the populations they represent may be the primary focus of a review. Unlike the case of diagnostic tests, when prognostic tests are being evaluated these probabilites (and probability categories such as “low”, “intermediate” or “high” risk) will often be expressed or at least derived from time-to-event analyses.

As noted, developing the topic and structuring the review can follow the general approach applicable to other medical tests. It is valuable to recognize that studies of prognostic tests will reflect a somewhat unique sequence of development.[1] Each phase of development is focused on different types of questions, research designs, and statistical methods, which could be used to formulate key questions for systematic reviews as follows:

1. Proof of concept: Is the test result associated with a clinically important outcome?
2. Prospective validation: How well does the measure predict outcomes in different patient cohorts?
3. Incremental predictive value: Does the new measure add predictive information to established prognostic methods?
4. Clinical utility: Does the measure change predicted probabilities enough to change medical decisions?
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?

The first three types of study tend to be unique to prognostic tests, while the last three types of studies are likely to be similar to those for other medical tests.

A separate category of prognostic test are tests that predict responsiveness to treatment (i.e., success or adverse effects). Generally evidence about predictive tests from randomized controlled trials (RCTs) or observational studies is reported as subgroup analyses using a statistical test for an interaction between the treatment groups and subgroups defined by the baseline indicator (predictor). Systematic reviews of treatment interactions are not specifically discussed here.

Step 2: Searching for Studies

Reliable and validated methods to search the literature for information about prognostic tests have not been established. Some search strategies have been based on words in titles or abstracts and Medical Subject Headings (MeSH) that appeared consistently in publications deemed most pertinent to a review.[2] Others have used search terms such as “cohort,” “incidence,” “mortality,” “followup studies,” “course,” or the word roots “prognos*” and “predict*” to identify relevant studies.[3] Obviously, search terms describing the prognostic test itself and the clinical condition or medical event to be predicted should also be used to focus the search.

Unlike most diagnostic tests, many prognostic tests are based on multivariable regression models. Reports in which a multivariable prognostic test was developed but not prospectively tested should not be the primary interest of systematic reviews. There are a plethora of reports that provide only proof of the concept that a given variable is independently associated with the development of disease or medical event and therefore might be useful for a prognostic test.[4-5] Reports in which the prognostic indicator of interest did not add significantly to a multivariable regression model may be particularly difficult to find via an electronic search.[6] However, these types of negative findings would indicate situations in which a potential prognostic test does not provide independent prognostic information. Therefore, if a review is going to focus on proof-of-concept questions, all studies that tested the prognostic indicator should be sought out, reviewed, and discussed even when the study merely mentions that the outcome was not independently related to the proposed prognostic test or components of a multivariable prediction model.[7]

Whenever a systematic review focuses on key questions about prognostic groups that are defined by predicted probabilities derived from a prognostic test, 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, those that would be treated differently in practice. RCTs of medical interventions in specific prognostic groups would help establish the rationale for using the prognostic test to classify patients into the prognostic group.

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, methods of analysis (including definition of prognostic groups), adjustment for covariates, and presentation of results.[8-10] 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 their inclusion or exclusion in a review.[3,11-12]

Table 12-1 lists questions that should be considered in this context depending on the type of research question being addressed by the 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 these or other criteria to score the quality of studies of prognostic tests have not been established. Reviewers need to decide which of these general criteria or others are appropriate for judging studies for their particular review.

Table 12-1. Questions to help reviewers judge the quality of individual studies of prognostic tests

1. Was the study designed to test the new prognostic test, or was it a secondary analysis of data collected for other purposes? Did the study employ a prospective cohort or nested case-control design?
2. Were the population of interest and clinical application clearly described, and were the subjects in the study representative as judged by the sampling plan, inclusion or exclusion criteria, subject participation, and characteristics of the sample?
3. Did everyone in the samples have a common starting point for followup with respect to the condition of interest, including any treatments?
4. Were the prognostic tests clearly described and conducted using a standardized, reliable, and valid method? a. Was the test used the same way in multiple sites/studies? b. Were the test results ascertained without knowledge of the outcome? c. What was the extent of and reasons for interdeteminant test results or missing values? Was any data imputation method used? d. Were any previously established prognostic indicators included in the analysis?
5. 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 study outcome? b. Was the duration of followup adequate?
6. Were the data used to develop the prognostic test? a. Were any data-driven variable or cut-point selection procedures used? b. Were regression model assumptions checked, including the functional form of the relationship between the test components and outcomes? c. Were previously established prognostic indicators or prediction models being used for comparison fit to the sample data in the same manner as the potential new prognostic test? d. How many potential prognostic inidcators were tested, and were there enough outcome events? e. Were the prognostic groups pre-defined based on clinically meaningful decision thresholds for predicted outcome probabilities? f. Was the comparison of prognostic tests adjusted for any other factors? Which ones? How? g. Were the results externally validated using an independent sample or internally validated via boot strap or cross-validation methods?
7. Did the statistical analysis address the question under review? a. Was the analysis a proof-of-concept study to indicate whether an independent association exists between the test or its components and the outcome as evidenced by a statistically significant hazard ratio, odds ratio, or relative risk in a multivariable regression model? b. Did the analysis indicate an improved discrimination of those who did or did not experience the outcome, as evidenced by a statistically significant increase sensitivity, specificity, or area under the ROC curve or Harrell's C index for time-to-event analyses when the prognostic test was compared to an established prognostic method? c. Did the analysis indicate a more accurate prognostic classification of subjects as evidenced by smaller differences between observed and predicted outcomes over time or in a reclassification table for a specific point in time? d. Was the analysis based on a controlled clinical trial of using the prognostic test?

Comparisons of prognostic tests should use data from the same cohort of subjects to minimize confounding the comparison.[13] 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 subjects' health state. Reviewers should note the starting point of each study reviewed. All of the prognostic indicators should be ascertained without knowledge of the outcome to avoid ascertainment bias.

Reviewers need to be aware of any previously established prognostic indicators that should be included in the analysis when evaluating the incremental predictive value or clinical utility of potential new prognostic tests. Comparison studies that ignore established or routinely employed predictors are certainly questionable. Reviewers need to pay close attention to what a new prognostic test was compared.

If the investigators fit the components of a prognostic test to the sample data by using the data to define cut-off points or functions and estimate regression coefficient(s), the estimated predictive performance might be overly optimistic. In addition, the fitting might bias the comparison to an established prognostic method that was not fit the sample. The number of observed outcome events needs to be adequate for the number of variables in the comparison (at least 10 to 20 outcome events per variable). Any adjustments for covariates that could make studies more or less comparable need to be noted.

Step 4: Performing Statistical Analysis

The summary statistics reported in the selected articles need to be appropriate for the 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 or test components 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 potential predictive relationships in different patient cohorts, and to a very limited extent incremental predictive value. Potential predictors that exhibit statistically significant associations in the form of an odds or hazard ratios often do not discriminate between subjects who eventually do or do not experience the outcome event.[14-15] Statistically significant associations (hazard or odds ratios, relative risks) merely indicate that more definitive evaluation of a new predictor is warranted.[16-17] Reviewers who are interested in summarizing estimates of hazard or odds ratios or relative risks are referred to other recent methodological literature.[18-25] However, the questions a systematic review could answer about a prognostic test by summarizing its association with the outcome are quite limited.

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 as indices of discrimination. These indices of discrimination can be calculated 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 experienced clinicians.[26-29] Reviewers should note whether any point estimates of sensitivity and specificity correspond with clinically meaningful prognostic groups defined by estimated outcome probabilities. If reviewers believe these indices are helpful for evaluating a prognostic test, time-dependent measures of sensitivity, specificity, and the ROC curve have been developed.[30] Harrell's C-statistic, which is conceptually similar to area under an ROC curve, can also be derived in a time-to-event analysis.[22,31-32] Examples of reviews and meta-analyses of prognostic tests that used these time-dependent measures of outcome group discrimination are not readily available. However, discrimination statistics such as sensitivity, specificity, and the area under a ROC curve do not directly address questions about the clinical utility of a new prognostic indicator and its potential impact on patient outcomes.[33-35]

Reclassification tables. The clinical utility of a prognostic test depends largely on its effect on predicted outcome probabilities. For example, expert guidelines use prognostic groups defined by the estimated 10-year risk of developing cardiovascular disease based on the Framingham cardiovascular risk score to recommend interventions to prevent future adverse medical events.[36] In this case, reviewers should ask, "How does a new prognostic test affect the estimated 10-year risks (the predicted outcome probabilities)?" Analyses of reclassification tables are now being reported to help answer this question.[37-43] Table 12-2 shows a fictional example of a reclassification table. Ideally, the classification of outcome probabilities into prognostic groups should be based on outcome probabilities that will lead to different courses of action (> 0.10 in the example). If not, the reviewer needs to take note, because the observed reclassifications could be 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 subjects might have no clinical utility.

Table 12-2. Example reclassification table based on predicted outcome probabilities

Grouped estimated mortality probabilities from prediction model 1	Grouped estimated mortality probabilities from prediction model 1 + new predictor		
	0 to 0.10	**> 0.10**	**Total**
0 to 0.10			
Number	900	100 (10%)	1000
Group mortality prediction model 1	4.0%	8.0%	4.40%
Group mortality prediction model 1 + new	3.8%	11.0%	-
Observed mortality	3.9%	12.0%	4.7%
> 0.10			
Number	100(25%)	300	400
Group mortality prediction model 1	15.0%	17.0%	16.5%
Group mortality prediction model 1 + new	9.0%	19.0%	-
Observed mortality	10.0%	19.0%	16.8%
Total			
Number	1000	400	1400
Group mortality prediction model 1+ new	4.3%	17.0%	-
Observed mortality	4.5%	17.2%	8.2%

Reclassification tables that might be found in reviewed articles typically summarize the number of subjects who were placed in each prognostic group by a prognostic test (prediction model) and the number (percentage) reclassified into a different prognostic group when a new prognostic variable was added to a prediction model or a different prognostic test was used. The table also should summarize each prognostic group's individual predicted probabilities and the observed percentages (cumulative proportion from a Kaplan-Meier curve or a cumulative incidence curve) that experienced the outcome. If a systematic review finds several articles that report a reclassification table comparing the same prognostic test using the same clinically meaningful prognostic groups and follow-up time, then tests of homogeneity and perhaps pooling of the observed outcome probabilities should be considered..

Reclassification tables typically also provide information about the differences between the predicted and observed group outcome probabilities. Closer agreement between the predicted and observed outcome probabilities (percentages) in each prognostic group would indicate an overall improvement when using a new prognostic test. The differences between estimates from each prognostic test and the observed outcomes might be analyzed by a chi-square goodness-of-fit test separately for each prognostic test.[44] However, these results will not help the reviewer determine if the differences in predicted and observed probabilities would be substantially better when the new prognostic test is used. In the example depicted in Table 12-2, the differences between predicted and observed values for each prediction model or prognostic test are small, as expected in prognostic groups with a narrow range of predicted probabilities.

Reviewers might also encounter articles that report separate reclassification tables for patients who did or did not experience the outcome event along with a summary statistic known as the net reclassification improvement (NRI).[45] This method of analysis requires the same single time-point dichotomy of outcomes needed to calculate sensitivity and specificity. In the group that developed the outcome event with a 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 2-by-2 reclassification table of only subjects who experienced the outcome event (e.g., those who died), this net difference is essentially the average change in 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 2-by-2 reclassification table of only subjects who did not experience the event (e.g., those who survived), this net difference is essentially the average change in specificity. The NRI is the difference between the net differences in improvement in the groups that did or did not experience the outcome. If the calculation used the means of individual predicted probabilities rather than a classification into prognostic groups, the result is known as the integrated discrimination index (IDI). Estimates of the NRI or IDI from different studies could be summarized, but the result would be prone to the same previously mentioned limitations as using sensitivity and specificity, or combinations thereof, to compare predictive performance of prognostic tests.

Predictive values. Methods to analyze the predictive values of prognostic tests have been proposed but most likely will not be used extensively in the current literature. Published examples of meta-analyses of predictive values of prognostic effects were not found.

Treatment decisions based on outcome probabilities are often dichotomous—for example, “treat ‘high-risk’” and “don’t treat ‘low-risk’” groups. The observed proportions of those who would be treated or untreated who do or do not experience the predicted outcome have direct clinical relevance as predictive values. If patients in the lower risk group are predicted not to develop the outcome (or would be treated as if they were not going to), then one minus the observed outcome probability is the negative predictive value (the percentage of those who were predicted not to develop the outcome and did not) analogous to the negative predictive value of a diagnostic test. If patients in the higher risk group were predicted to develop the outcome (or at least would be treated because they are at “high risk”), then the observed percentage is similar to the positive predictive value (the percentage of those who were predicted to develop the outcome and did). If investigators simplify analysis of prognostic indicators by dichotomizing the stochastic outcome, the positive and negative predictive values of prognostic categories can be compared in a way that is similar to the way predictive values of diagnostic tests can be compared. Most likely the ratio of positive and negative predictive values of the two prognostic variables will be summarized in a report, along with a confidence interval.[46] The regression model proposed by Leisenring and colleagues might be used to determine how patient characteristics relate to the relative predictive values.[47] When two categorical prognostic tests are compared using time-to-event analyses, Kaplan-Meier curves could be graphed for each prognostic group defined by the comparator alone and then by the new prognostic test within each of the original prognostic groups. Having information about the percentages of the sample in each category and the overall cumulative outcome probabilities would help show how much the new prognostic test changed the observed outcome probabilities in each of the original prognostic groups. Measures of the degree of separation have been proposed.[48-49]

If the outcome probabilities increase with scores derived from the time-to-event predictive models , Kaplan-Meier curves could graphed for percentiles—for example, quartiles of scores estimated from each prognostic test to summarize the magnitude of the differences in outcome probabilities.[50-52] Alternatively, a modification of the Brier score of predictive accuracy that is based on the difference between predicted probabilities at specific time points during followup and actual outcome expressed as a 0 (no event occurred) or 1 event (event did occur) has been developed to compare the predictive accuracy in different stratifications (prognostic groups).[53] However, the reviewer would have to be sure the percentiles or stratification used in both of these methods corresponded with the clinically meaningful prognostics groups defined by the predicted probabilities. When the results of prognostic tests being compared are continuous and prognostic groups have not been established, investigators might compare “predictiveness” curves and test whether one variable has significantly different predictive values over a range of cut-points using a regression model.[50,52,54-55]

Step 5: Using Decision Models

The most definitive level of evidence to answer the most important questions about a prognostic test would come from RCTs of its use that demonstrate a net improvement in patient outcomes and cost-effectiveness. As with most studies of medical tests, there are few trials to provide this level of evidence. Decision modeling can provide insight into the possible value of prognostic tests.[56-57]

Modeling the variation in outcome probabilities from several studies of a prognostic group may provide some insights into the stability of the estimates and whether variation in the estimates is related to characteristics of the prognostic groups from different studies. Methods have been developed to model outcome probabilities from different studies.[23] Dear proposed a generalized least squares linear regression model similar to a meta-regression for cumulative probabilities of being event-free that can incorporate covariates including characteristics of the prognostic groups.[58] Dear’s method takes the estimated covariance between multiple time points in the same prognostic group into account, but requires estimates for the same time points—for example, 6 and 12 months of followup—for each prognostic group. This linear model of untransformed outcome probabilities does not restrict the results to the appropriate 0 to 1 interval for a probability estimate, and fixed rather than random effects are modeled.

Arends and colleagues have proposed a multivariate mixed-effects model for the joint analysis of cumulative probabilities reported at one or more time (possibly different times) in different studies that incorporate time (or some transformation thereof) as a continuous rather than discrete variable.[59] This model reduces to the commonly used DerSimonian-Laird random-effects model when there is only one common followup time for all prognostic groups/studies in the analysis. Arends' model can provide estimates of cumulative probabilities that are confined to the appropriate 0 to 1 interval. However, the estimates of the cumulative probability of being event-free are not forced to decrease with followup time or have an intercept equal to 1. In addition, the fit of the model is difficult to judge.

Summary

Key points from the above discussion are:

- In contrast to a medical test conducted to help determine the presence or absence of a disease at the time a test is performed, prognostic tests are used to make probabilistic predictions about clinically important outcomes that might or might not occur in the future. The time dependency inherent in the evaluation of prognostic tests should be taken into consideration in every step of a systematic review.
- A large number of published reports focus on the associations between prognostic indicators and patient outcomes as the first stage of development of prognostic tests. These articles can be difficult to find through electronic searches when an association was not found. Other studies focus on the statistical development of multivariable prediction models as prognostic tests. A systematic review of comparative effectiveness need not focus on these types of articles.
- 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 with some modifcations for differences inherent in studies of prognostic tests. Suggestions are listed in Table 12-1.
- 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 that would be treated differently. The resultant prognostic groups need to be described in detail including their outcome probabilities.
- Given fundamental differences in how diagnostic and prognostic tests are used in practice, and the time dimension inherent in prognostic tests, some of the most commonly used methods for evaluating and reviewing diagnostic tests, such as point estimates of test sensitivity and specificity, are not as informative for prognostic tests. The most applicable summary of the predictive accuracy of a prognostic test is the difference between predicted and observed outcome probabilites within the prognostic groups. Methods to summarize and compare these differences need further development and more wide spread use to facilitate systematic reviews.

References

1. 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.

2. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. J Am Med Inform Assoc 2001;8(4):391-7.

3. Hayden JA, Cote P, Bombardier C. Evaluation of the quality of prognosis studies in systematic reviews. Ann Intern Med 2006;144(6):427-37.

4. McShane LM, Altman DG, Sauerbrei W, et al. Reporting recommendations for tumor marker prognostic studies (REMARK). J Natl Cancer Inst 2005;97(16):1180-4.

5. 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.

6. 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.

7. Kyzas PA, Loizou KT, Ioannidis JP. Selective reporting biases in cancer prognostic factor studies. J Natl Cancer Inst 2005;97(14):1043-55.

8. Altman DG. Systematic reviews of evaluations of prognostic variables. BMJ 2001;323(7306):224-8.

9. Hall PA, Going JJ. Predicting the future: a critical appraisal of cancer prognosis studies. Histopathology 1999;35(6):489-94.

10. Altman DG, Riley RD. Primer: an evidence-based approach to prognostic markers. Nat Clin Pract Oncol 2005;2(9):466-72.

11. 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.

12. 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.

13. 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.

14. Ware JH. The limitations of risk factors as prognostic tools. N Engl J Med 2006;355(25):2615-7.

15. 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.

16. 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.

17. 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.

18. 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.

19. Hunink MG, Wong JB. Meta-analysis of failure-time data with adjustment for covariates. Med Decis Making 1994;14(1):59-70.

20. Williamson PR, Smith CT, Hutton JL, et al. Aggregate data meta-analysis with time-to-event outcomes. Stat Med 2002;21(22):3337-51.

21. Moodie PF, Nelson NA, Koch GG. A non-parametric procedure for evaluating treatment effect in the meta-analysis of survival data. Stat Med 2004;23(7):1075-93.

22. The Fibrinogen Studies Collaboration. Measures to assess the prognostic ability of the stratified Cox proportional hazards model. Stat Med 2009;28(3):389-411.

23. Earle CC, Pham B, Wells GA. An assessment of methods to combine published survival curves. Med Decis Making 2000;20(1):104-11.

24. Peters J, Mengersen K. Selective reporting of adjusted estimates in observational epidemiology studies: reasons and implications for meta-analyses. Eval Health Prof 2008;31(4):370-89.

25. 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.

26. 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.

27. 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.

28. Mackillop WJ, Quirt CF. Measuring the accuracy of prognostic judgments in oncology. J Clin Epidemiol 1997;50(1):21-29.

29. 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.

30. 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.

31. 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.

32. 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.

33. 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.

34. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007;115(7):928-35.

35. 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.

36. 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.

37. 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.

38. Ridker PM, Buring JE, Rifai N, et al. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. JAMA 2007;297(6):611-9.

39. 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.

40. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clin Chem 2008;54(1):17-23.

41. Ridker PM, Paynter NP, Rifai N, et al. C-reactive protein and parental history improve global cardiovascular risk prediction: the Reynolds Risk Score for men. Circulation 2008;118(22):2243-51.

42. 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.

43. 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.

44. Pigeon JG, Heyse JF. An improved goodness of fit statistic for probability prediction models. Biom J 1999;41(1):71-82.

45. 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.

46. 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.

47. Leisenring W, Alonzo T, Pepe MS. Comparisons of predictive values of binary medical diagnostic tests for paired designs. Biometrics 2000;56(2):345-51.

48. 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.

49. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. Stat Med 2004;23(5):723-48.

50. Bura E, Gastwirth JL. The binary regression quantile plot: assessing the importance of predictors in binary regression visually. Biom J 2001;43(1):5-21.

51. Folsom AR, Chambless LE, Ballantyne CM, et al. An assessment of incremental coronary risk prediction using C-reactive protein and other novel risk markers: the atherosclerosis risk in communities study. Arch Intern Med 2006;166(13):1368-73.

52. 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.

53. Schemper M, Henderson R. Predictive accuracy and explained variation in Cox regression. Biometrics 2000;56(1):249-55.

54. Huang Y, Sullivan Pepe M, Feng Z. Evaluating the predictiveness of a continuous marker. Biometrics 2007;63(4):1181-8.

55. Moskowitz CS, Pepe MS. Quantifying and comparing the predictive accuracy of continuous prognostic factors for binary outcomes. Biostatistics 2004;5(1):113-27.

56. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26(6):565-74.

57. Greenland S. The need for reorientation toward cost-effective prediction: comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M. J. Pencina et al., Statistics in Medicine (DOI: 10.1002/sim.2929). Stat Med 2008;27(2):199-206.

58. Dear KB. Iterative generalized least squares for meta-analysis of survival data at multiple times. Biometrics 1994;50(4):989-1002.

59. Arends LR, Hunink MG, Stijnen T. Meta-analysis of summary survival curve data. Stat Med 2008;27(22):4381-96.