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,

DRAFT

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

DRAFT

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]

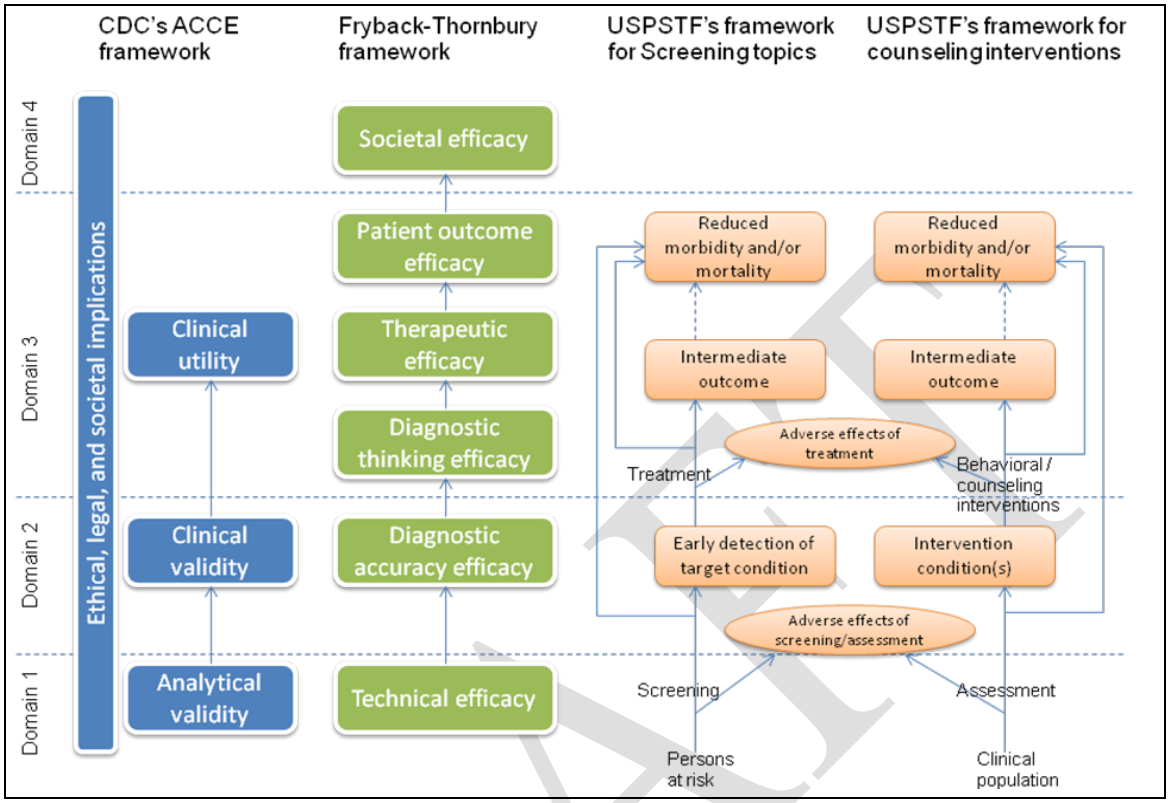
A Note on Terminology

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



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.

DRAFT

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.

DRAFT

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

DRAFT

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.

DRAFT

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.

DRAFT