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

DRAFT

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.

DRAFT

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

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

DRAFT

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

DRAFT

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.

DRAFT

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.

DRAFT

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.

DRAFT