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

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 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.

DRAFT

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.

DRAFT

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]

DRAFT

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.

DRAFT

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

DRAFT

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.

DRAFT