Chapter 5: Assessing Risk of Bias as a Domain of Quality in Medical Test Studies

P. Lina Santaguida, PhD, MSc[1], Crystal M. Riley, MA[2,4], and David B. Matchar, MD[3,4]

[1]Department of Clinical Epidemiology & Biostatistics, McMaster University, Hamilton, ON, Canada; [2]Division of Psychology at Nanyang Technological University, Nanyang Technological University, Singapore, Singapore; [3]Duke University, Durham, NC, USA; [4]Health Services and Systems Research, Duke-NUS Graduate Medical School, Singapore, Singapore.

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

KEY WORDS: medical test; review; systematic error; risk of bias.
J Gen Intern Med 27(Suppl 1):S33–8
DOI: 10.1007/s11606-012-2030-8

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

The evaluation of study features that might influence the relative importance of a particular study has often been framed as an assessment of quality. Quality assessment—a broad term used to encompass the examination of factors such as systematic error, random error, adequacy of reporting, aspects of data analysis, applicability, specifying ethics approval and detailing sample size estimates—has been conceptualized in a variety of ways.[2, 3] In addition, some schemes for quality assessment apply to individual studies and others to a body of literature. As a result, many different tools have been developed to formally evaluate the quality of studies of medical tests; however, there is no empirical evidence that any sort of score based on quantitative weights of individual study features can predict the degree to which a study is more or less "true." In this context, systematic reviewers have not yet achieved consensus on the optimal criteria to assess study quality.

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

This chapter is part of the *Methods Guide for Medical Test Reviews* produced by the Agency for Healthcare Research and Quality (AHRQ) Evidence-Based Practice Centers (EPC) for AHRQ and the Journal of General Internal Medicine. Similar to the *General Methods Guide*,[1] assessment of the major features that influence the importance of a study to key review questions are assessed separately. Chapter 6 of this *Guide* considers the evaluation of the applicability of a particular study to a key review question. Chapter 7 details the assessment of the quality of a body of evidence, and Chapter 8 covers the issue of random error, which can be addressed when considering all relevant studies through the use, if appropriate, of a summary

measure combining study results. Thus, this chapter highlights key issues when assessing risk of bias in studies evaluating medical tests—systematic error resulting from design, conduct, or reporting that can lead to over- or underestimation of test performance.

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

EVIDENCE FOR BIASES AFFECTING MEDICAL TEST STUDIES

Before considering risk of systematic bias, it is useful to consider the range of limitations in medical test studies. In a series of studies of bias in the context of medical test literature, Whiting et al. reviewed studies of the impact of a range of specific sources of error in diagnostic test studies conducted from 1966 to 2000.[3–5] In the review, the term "test" was defined broadly to include traditional laboratory tests, clinical examinations, imaging tests, questionnaires, pathology, and measures of health status (e.g., the presence of disease or different stages/severity of a disease).[6] Each test included in the analysis was compared to a reference standard, defined as the best comparator test to diagnose the disease or health condition in question. The results of this analysis indicated that no conclusions could be drawn about the direction or relative magnitude of effects for these specific biases. Although not definitive, the reviews showed that bias does occur and that some sources of bias—including spectrum bias, partial verification bias, clinical review bias, and observer or instrument variation—are particularly common in studies of diagnostic accuracy.[3] As a guide to further work, the authors summarized the range of quality issues arising in the reviewed articles (Table 1).

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

COMMON CHALLENGES

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

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

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

PRINCIPLES FOR ADDRESSING THE CHALLENGES

Principle 1: Use Validated Criteria to Address Relevant Sources of Bias

In selecting criteria for assessing risk of bias, multiple instruments are available, and reviewers must choose the one most appropriate to the task. Two systematic reviews have evaluated quality assessment instruments specifically in the context of diagnostic accuracy. West et al.[9] evaluated 18 tools (six scales, nine guides, and three EPC rating systems). All of the tools were intended for use in conjunction with other tools relevant for judging the design-specific attributes of the study (for example, quality of RCTs or observational studies). Three scales met all six criteria considered important: 1) the Cochrane Working group checklist,[10] 2) the tool of Lijmer et al.,[11] and 3) the National Health and Medical Research Council checklist.[12]

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

Table 1. Commonly Reported Sources of Systematic Bias in Studies of Medical Test Performance

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

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

The QUADAS checklist attempted to incorporate the sources of bias and error that had some empirical basis and validity.[6–8] This tool contains elements of study limitations beyond those concerned with risk of systematic bias; it also includes questions related to reporting. An updated version of this scale, called QUADAS-2, identifies four key domains (patient selection, index test(s), reference standard, and flow and timing), which are each rated in terms of risk of bias.[13] The updated checklist is shown in Table 2.

We recommend that reviewers use criteria that assess the risk of systematic error that have been validated to some degree from an instrument like QUADAS-2. Chapters 6 and 8 discuss applicability and random error, which are other important aspects of quality assessment. In

Table 2. QUADAS-2 Questions for Assessing Risk of Bias in Diagnostic Accuracy Studies*

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

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

addition to disregarding irrelevant items, systematic reviewers may also need to add additional criteria from other standardized checklists such as Standards for Reporting of Diagnostic Accuracy (STARD)[14] or Strengthening the Reporting of Genetic Association Studies (STREGA),[15] (an extension of the Strengthening the Reporting of Observational Studies in Epidemiology [STROBE]).[16]

Principle 2: Standardize the Application of Criteria

In order to maintain objectivity in an otherwise subjective process, it is useful to standardize the application of criteria. There is little empirical evidence to inform decisions about this process. Thus, we recommend that the review team establish clear definitions for each criterion. This approach is demonstrated in the Illustration section below. In addition, it can be useful to pilot the criteria definitions with at least two reviewers. In this way, reviewers can revise unreliable terms and measure the reliability of the ultimate criteria.

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

Principle 3: Decide When Inadequate Reporting Constitutes a Fatal Flaw

Reviewers must also carefully consider how to handle inadequate reporting. Inadequate reporting, in and of itself, does not introduce systematic bias, but it does limit the

Table 3. Categorizing Individual Studies into General Quality Classes*

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

**Adapted from AHRQ's General Methods Guide* [1]

reviewers' ability to assess the risk of bias. Some systematic reviewers may take a conservative approach by assuming the worst, while others may be more liberal by giving the benefit of the doubt.

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

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

ILLUSTRATION

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

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

SUMMARY

An assessment of methodological quality is a necessary activity for authors of systematic reviews; this should include an evaluation of the evidence for studies of medical test

Table 4. Interpretation of Partial Verification Bias: the Example of Family History[17, 18]*

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

* *See text*
Abbreviation: QUADAS = Quality Assessment of Diagnostic Accuracy Studies

performance. Judging the overall quality of an individual study involves examining the size of the study, the direction and degree of findings, the relevance of the study, and the risk of bias in the form of systematic error, internal validity, and other study limitations. In this chapter of the *Methods Guide for Medical Test Reviews*, we focus on the evaluation of systematic bias in an individual study as a distinctly important component of quality in studies of medical test performance.

KEY POINTS

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

ACKNOWLEDGEMENTS: *The AHRQ has funded the preparation of the Methods Guide for Medical Test Reviews, including this chapter. Sean R. Love assisted in the editing and preparation of this manuscript.*

Conflict of Interest: *The authors declare that they do not have a conflict of interest.*

Corresponding Author: *David B. Matchar, MD; Health Services and Systems Research, Duke-NUS Graduate Medical School, 8 College Road, Singapore, Singapore 169857 (e-mail: David.matchar@duke-nus.edu.sg).*

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. **Higgins JPT, Altman DG, Sterne JAC** on behalf of the Cochrane Statistical Methods Group and the Cochrane Bias Methods Group. Chapter 8: Assessing risk of bias in included studies. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 (updated March 2011). The Cochrane Collaboration, 2011. Available at: http://www.cochrane-handbook.org. Accessed September 19, 2011.
3. **Whiting P, Rutjes AWS, Reitsma JB, et al.** Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med. 2004;140(3):189–202.
4. **Whiting P, Rutjes AWS, Dinnes J, et al.** A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. J Clin Epidemiol. 2005;58:1–12.
5. **Whiting P, Rutjes AWS, Dinnes J, et al.** Development and validation of methods for assessing the quality of diagnostic accuracy studies. Health Technol Assess. 2004;8(25):iii, 1-234.
6. **Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J.** The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. 2003;3:25.
7. **Leeflang MMG, Deeks JJ, Gatsonis C, Bossuyt PMM.** on behalf of the Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. Ann Intern Med. 2008;149(12):889–97.
8. Centre for Reviews and Dissemination. Systematic Reviews: CRD's Guidance for Undertaking Reviews in Health Care. Centre for Reviews and Dissemination: York, UK; 2009. Available at: http://www.york.ac.uk/inst/crd/pdf/Systematic_Reviews.pdf. Accessed September 19, 2011.
9. **West S, King V, Carey TS, et al.** Systems to rate the strength of scientific evidence. (Prepared by the Research Triangle Institute – University of North Carolina Evidence-based Practice Center under Contract No. 290-97-0011.) AHRQ Publication No. 02-E016. Rockville, MD: Agency for Healthcare Research and Quality. April 2002. Available at: http://www.thecre.com/pdf/ahrq-system-strength.pdf. Accessed September 19, 2011.
10. Cochrane Methods Working Group on Systematic Review of Screening and Diagnostic Tests. Recommended Methods; 1996.
11. **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.
12. National Health and Medical Research Council (NHMRC). How to Review the Evidence: Systematic Identification and Review of the Scientific Literature. Canberra: NHMRC; 2000.
13. **Whiting P, Rutjes A, Sterne J, et al.** QUADAS-2. (Prepared by the QUADAS-2 Steering Group and Advisory Group). Available at: http://www.bris.ac.uk/quadas/resources/quadas2.pdf. Accessed September 12, 2011.
14. **Bossuyt PM, Reitsma JB, Bruns DE, et al.** Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med. 2003;138(1):40–4.
15. **Little J, Higgins JPT, Ioannidis JPA, et al.** STrengthening the REporting of Genetic Association studies (STREGA) - an extension of the STROBE statement. Eur J Clin Invest. 2009;39:247–66.
16. **von Elm E, Altman DG, Egger M, et al.** The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007;370:1453–7.
17. **Qureshi N, Wilson B, Santaguida P, et al.** Family History and Improving Health. Evidence Report/Technology Assessment No. 186. (Prepared by the McMaster University Evidence-based Practice Center, under Contract No. HHSA 290-2007-10060-I.) AHRQ Publication No. 09-E016. Rockville, MD: Agency for Healthcare Research and Quality. August 2009. Available at: http://www.ahrq.gov/downloads/pub/evidence/pdf/famhistory/famhimp.pdf. Accessed February 28, 2011.
18. **Wilson BJ, Qureshi N, Santaguida P, et al.** Systematic review: family history in risk assessment for common diseases. Ann Intern Med. 2009;151(12):878–85.