

Systematic Reviews of Diagnostic Test Accuracy

Mariska M.G. Leeflang, PhD; Jonathan J. Deeks, PhD; Constantine Gatsonis, PhD; and Patrick M.M. Bossuyt, PhD, on behalf of the Cochrane Diagnostic Test Accuracy Working Group

More and more systematic reviews of diagnostic test accuracy studies are being published, but they can be methodologically challenging. In this paper, the authors present some of the recent developments in the methodology for conducting systematic reviews of diagnostic test accuracy studies. Restrictive electronic search filters are discouraged, as is the use of summary quality scores. Methods for meta-analysis should take into account the paired nature of the estimates and their dependence on threshold.

Authors of these reviews are advised to use the hierarchical summary receiver-operating characteristic or the bivariate model for the data analysis. Challenges that remain are the poor reporting of original diagnostic test accuracy studies and difficulties with the interpretation of the results of diagnostic test accuracy research.

Ann Intern Med. 2008;149:889-897.

www.annals.org

For author affiliations, see end of text.

Diagnosis is a critical component of health care, and clinicians, policymakers, and patients routinely face a range of questions regarding diagnostic tests. They want to know whether testing improves outcome; what test to use, purchase, or recommend in practice guidelines; and how to interpret test results. Well-designed diagnostic test accuracy studies can help in making these decisions, provided that they transparently and fully report their participants, tests, methods, and results as facilitated, for example, by the STARD (Standards for Reporting of Diagnostic Accuracy) statement (1). That 25-item checklist was published in many journals and is now adopted by more than 200 scientific journals worldwide.

As in other areas of science, systematic reviews and meta-analysis of accuracy studies can be used to obtain more precise estimates when small studies addressing the same test and patients in the same setting are available. Reviews can also be useful to establish whether and how scientific findings vary by particular subgroups, and may provide summary estimates with a stronger generalizability than estimates from a single study. Systematic reviews may help identify the risk for bias that may be present in the original studies and can be used to address questions that were not directly considered in the primary studies, such as comparisons between tests. The Cochrane Collaboration is the largest international organization preparing, maintaining, and promoting systematic reviews to help people make well-informed decisions about health care (2). The Collaboration decided in 2003 to make preparations for including systematic reviews of diagnostic test accuracy in their Cochrane Database of Systematic Reviews. To enable this, a working group (Appendix, available at www.annals.org) was formed to develop methodology, software, and a handbook. The first diagnostic test accuracy review was published in the Cochrane Database in October 2008.

In this paper, we review recent methodological developments concerning problem formulation, location of literature, quality assessment, and meta-analysis of diagnostic accuracy studies by using our experience from the work on the Cochrane Handbook. The information presented here is based on the recent literature and updates previously published guidelines by Irwig and colleagues (3).

DEFINITION OF THE OBJECTIVES OF THE REVIEW

Diagnostic test accuracy refers to the ability of a test to distinguish between patients with disease (or more generally, a specified target condition) and those without. In a study of test accuracy, the results of the test under evaluation, the index test, are compared with those of the reference standard determined in the same patients. The reference standard is an agreed-on and accurate method for identifying patients who have the target condition. Test results are typically categorized as positive or negative for the target condition. By using such binary test outcomes, the accuracy is most often expressed as the test's sensitivity (the proportion of patients with positive results on the reference standard that are also positive on the index test) and specificity (the proportion of patients with negative results on the reference standard that are also negative on the index test). Other measures have been proposed and are in use (4–6).

It has long been recognized that test accuracy is not a fixed property of a test. It can vary between patient subgroups, with their spectrum of disease, with the clinical setting, or with the test interpreters and may depend on the results of previous testing. For this reason, inclusion of these elements in the study question is essential. In order to make a policy decision to promote use of a new index test, evidence is required that using the new test increases test accuracy over other testing options, including current practice, or that the new test has equivalent accuracy but offers other advantages (7–9). As with the evaluation of interventions, systematic reviews need to include comparative analyses between alternative testing strategies and should not

See also:

Print

Editorial comment. 904

Web-Only

Appendix
Appendix Table
Conversion of graphics into slides

focus solely on evaluating the performance of a test in isolation.

In relation to the existing situation, 3 possible roles for a new test can be defined: replacement, triage, and add-on (7). If a new test is to replace an existing test, then comparing the accuracy of both tests on the same population and with the same reference standard provides the most direct evidence. In triage, the new test is used before the existing test or testing pathway, and only patients with a particular result on the triage test continue the testing pathway. When a test is needed to rule out disease in patients who then need no further testing, a test that gives a minimal proportion of false-negative results and thus a relatively high sensitivity should be used. Triage tests may be less accurate than existing ones, but they have other advantages, such as simplicity or low cost. A third possible role of a new test is add-on. The new test is then positioned after the existing testing pathway to identify false-positive or false-negative results after the existing pathway. The review should provide data to assess the incremental change in accuracy made by adding the new test.

An example of a replacement question can be found in a systematic review of the diagnostic accuracy of urinary markers for primary bladder cancer (10). Clinicians may use cytology to triage patients before they undergo invasive cystoscopy, the reference standard for bladder cancer. Because cytology combines high specificity with low sensitivity (11), the goal of the review was to identify a tumor marker with sufficient accuracy to either replace cytology or be used in addition to cytology. For a marker to replace cytology, it has to achieve equally high specificity with improved sensitivity. New markers that are sensitive but not specific may have roles as adjuncts to conventional testing. The review included studies in which the test under evaluation (several different tumor markers and cytology) was evaluated against cystoscopy or histopathology. Included studies compared 1 or more of the markers, cytology only, or a combination of markers and cytology.

Although information on accuracy can help clinicians make decisions about tests, good diagnostic accuracy is a desirable but not sufficient condition for the effectiveness of a test (8). To demonstrate that using a new test does more good than harm to patients tested, randomized trials of test-and-treatment strategies and reviews of such trials may be necessary. However, with the possible exception of screening, in most cases, such randomized trials are not available and systematic reviews of test accuracy may provide the most useful evidence available to guide clinical and health policy decision making and use as input for decision and cost-effectiveness analysis (12).

IDENTIFICATION AND SELECTION OF STUDIES

Identifying test accuracy studies is more difficult than searching for randomized trials (13). There is not a clear, unequivocal keyword or indexing term for an accuracy

study in literature databases comparable with the term “randomized, controlled trial.” The Medical Subject Heading “sensitivity and specificity” may look suitable but is inconsistently applied in most electronic bibliographic databases. Furthermore, data on diagnostic test accuracy may be hidden in studies that did not have test accuracy estimation as their primary objective. This complicates the efficient identification of diagnostic test accuracy studies in electronic databases, such as MEDLINE. Until indexing systems properly code studies of test accuracy, searching for them will remain challenging and may require additional manual searches, such as screening reference lists.

In the development of a comprehensive search strategy, review authors can use search strings that refer to the test(s) under evaluation, the target condition, and the patient description or a subset of these. For tests with a clear name that are used for a single purpose, searching for publications in which those tests are mentioned may suffice. For other reviews, adding the patient description may be necessary, although this is also often poorly indexed. A search strategy in MEDLINE should contain both Medical Subject Headings and free text words. A search strategy for articles about tests for bladder cancer, for example, should include as many synonyms for bladder cancer as possible in the search strategy, including neoplasm, carcinoma, transitional cell, and hematuria.

Several methodological electronic search filters for diagnostic test accuracy studies have been developed, each attempting to restrict the search to articles that are most likely to be test accuracy studies (13–16). These filters rely on indexing terms for research methodology and text words used in reporting results, but they often miss relevant studies and are unlikely to decrease the number of articles one needs to screen. Therefore, they are not recommended for systematic reviews (17, 18). The incremental value of searching in languages other than English and in the gray literature has not yet been fully investigated.

In systematic reviews of intervention studies, publication bias is an important and well-studied form of bias in which the decision to report and publish studies is linked to their findings. For clinical trials, the magnitude and determinants of publication bias have been identified by tracing the publication history of cohorts of trials reviewed by ethics committees and research boards (19). A consistent observation has been that studies with significant results are more likely to be published than studies with nonsignificant findings (19). Investigating publication bias for diagnostic tests is problematic, because many studies are done without ethical review or study registration; therefore, identification of cohorts of studies from registration to final publication status is not possible (20). Funnel plot-based tests used to detect publication bias in reviews of randomized, controlled trials have proven to be seriously misleading for diagnostic studies, and alternatives have poor power (21). Also, because diagnostic accuracy studies frequently do not compare tests, they tend not to routinely

report *P* values that dichotomize comparisons as significant or not significant. Without the same emphasis being given to statistical significance, the determinants for publication of diagnostic studies are unlikely to be the same as those of intervention studies.

ASSESSMENT OF METHODOLOGICAL QUALITY

Variability among diagnostic accuracy study results is to be expected. Some of this variability is due to chance, because many diagnostic studies have small sample sizes (22). The remaining heterogeneity may be due to differences in study populations, but differences in study methods are also likely to result in differences in accuracy estimates (23). Test accuracy studies with design deficiencies can produce biased results (24–26). The **Table** lists some of the more important forms of bias. Sources of bias for which unambiguous evidence indicates that they lead to overestimation of diagnostic accuracy are the inclusion of healthy control participants and the differential use of reference standards (24, 26).

Quality assessment of individual studies in systematic reviews is therefore necessary to identify potential sources of bias and to limit the effects of these biases on the estimates and the conclusions of the review. We recommend the QUADAS (Quality Assessment of Diagnostic Accuracy

Studies) checklist to assess the quality of diagnostic test accuracy studies (27). In addition, specific sources of bias may exist for different types of diagnostic tests. For example, in studies assessing the accuracy of biochemical serum markers, data-driven selection of the cutoff value may bias diagnostic accuracy (28, 29). Review authors should therefore think carefully about whether specific items need to be added to the QUADAS list.

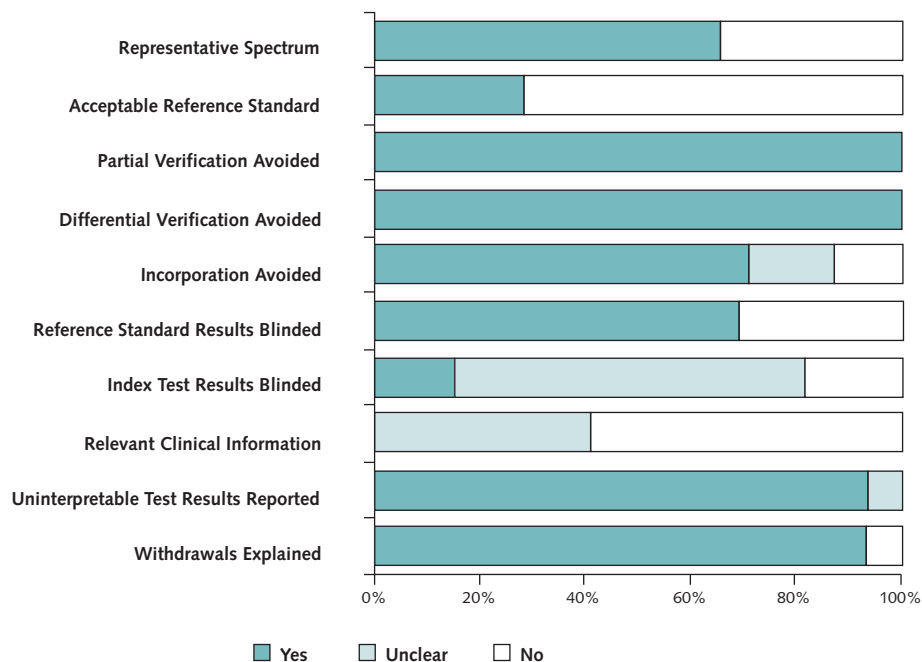
The results of quality appraisal can be summarized to offer a general impression of the validity of the available evidence. Review authors should not use an overall quality score, because different shortcomings may generate different magnitudes of bias, even in opposing directions, which makes it very hard to attach sensible weights to each quality item (30). **Figure 1** shows a way to summarize the quality assessment, with stacked bars used for each QUADAS item. Another way of presenting the quality assessment results is by tabulating the results of the individual QUADAS items for each study. In the analysis phase, the results of the quality appraisal may guide explorations of the sources of heterogeneity (32, 33). Possible methods to address quality differences are sensitivity analysis, subgroup analysis, or meta-regression analysis, although the number of included studies may often be too small for meaningful investigations. Also, incomplete reporting hampers any evaluation of study quality (34). The

Table. Sources of Bias in Diagnostic Test Accuracy Studies

Type of Bias	When Does It Occur?	Under- or Overestimation of Diagnostic Accuracy?*
Patients		
Spectrum bias	When included patients do not represent the intended spectrum of severity for the target condition or alternative conditions	Depends on difference between targeted and included part of spectrum
Selection bias	When eligible patients are not enrolled consecutively or randomly	Usually leads to overestimation
Index test		
Information bias	When the index test results are interpreted with knowledge of the results of the reference standard, or with more (or less) information than in practice	Usually leads to overestimation, unless less clinical information is provided than in practice, which may result in underestimation
Reference standard		
Misclassification bias	When the reference standard does not correctly classify patients with the target condition	Depends on whether both tests make the same mistakes
Partial verification bias	When a nonrandom set of patients does not undergo the reference standard	Usually leads to overestimation of sensitivity; effect on specificity varies
Differential verification bias	When a set of patients is verified with a second or third reference standard, especially when this selection depends on the index test result	Usually leads to overestimation
Incorporation bias	When the index test is incorporated in a (composite) reference standard	Usually leads to overestimation
Disease progression bias	When the patients' condition changes between administering the index test and the reference standard	Under- or overestimation, depending on change in patients' condition
Information bias	When the reference standard is interpreted knowing the index test results	Usually leads to overestimation
Data analysis		
Excluded data	When uninterpretable or intermediate test results and withdrawals are not included in the analysis	Usually leads to overestimation

* From references 24–26.

Figure 1. Review authors' judgments about quality items in a systematic review of magnetic resonance imaging for multiple sclerosis.



Data from reference 31. Data are presented as the proportion of included studies. Criteria that are unclear or not met introduce a risk for bias. The authors considered the relative lack of an acceptable reference standard as the main weakness of the review.

effects of the STARD guidelines for complete and transparent reporting (1) are only gradually becoming visible in the literature (35).

ANALYZING THE DATA AND PRESENTING THE RESULTS

Whereas the results of a randomized trial are often reported by using a single measure of effect, such as a difference in means, a risk difference, or a risk ratio, most diagnostic test accuracy studies report 2 or more statistics: the sensitivity and the specificity, the positive and negative predictive value, the likelihood ratios for the respective test results, or the receiver-operating characteristic (ROC) curve and quantities based on it (6, 36).

The first step in the meta-analysis of diagnostic test accuracy is to graph the results of the individual studies. The paired results for sensitivity and specificity in the included studies should be plotted as points in ROC space (Figure 2), which can highlight the covariation between sensitivity and specificity. In Figure 2, the x-axis of the ROC plot displays the specificity obtained in the studies in the review. The y-axis shows the corresponding sensitivity. The rising diagonal line indicates values of sensitivity and specificity that could be obtained by guessing and refers to a noninformative test: The chances of a positive test result are identical for patients with disease and those without. It is expected that most studies will be above this line. The best diagnostic tests will be positioned in the upper-left

corner of the ROC space, where both sensitivity and specificity are close to 1. Because CIs are not typically displayed on these plots, it is not possible to discern the cause of scatter across studies—it can be caused by either small sample sizes or heterogeneity between studies. Paired forest plots (Figure 3) display sensitivity and specificity separately (but on the same row) for each study together with CIs and tabular data. A disadvantage is that forest plots do not display the covariation between sensitivity and specificity.

The estimated sensitivity and specificity of a test often display a pattern of negative correlation. A major contributor to this appearance is the tradeoff between sensitivity and specificity when the threshold for defining test positivity varies. When high test results are positive, decreasing the threshold value that defines a test result as positive increases sensitivity and lowers specificity, and vice versa. When studies included in a review differ in positivity thresholds, an ROC-curve-like pattern may be discerned in the ROC plot. There may be explicit variation in thresholds if different studies use different numerical thresholds to define a test result as positive (for example, variation in the blood glucose level, above which a patient has diabetes). In other situations, unquantifiable or implicit variation in threshold may occur when test results depend on interpretation or judgment (for example, between radiographers classifying images as normal or abnormal) or when test results are sensitive to machine calibration.

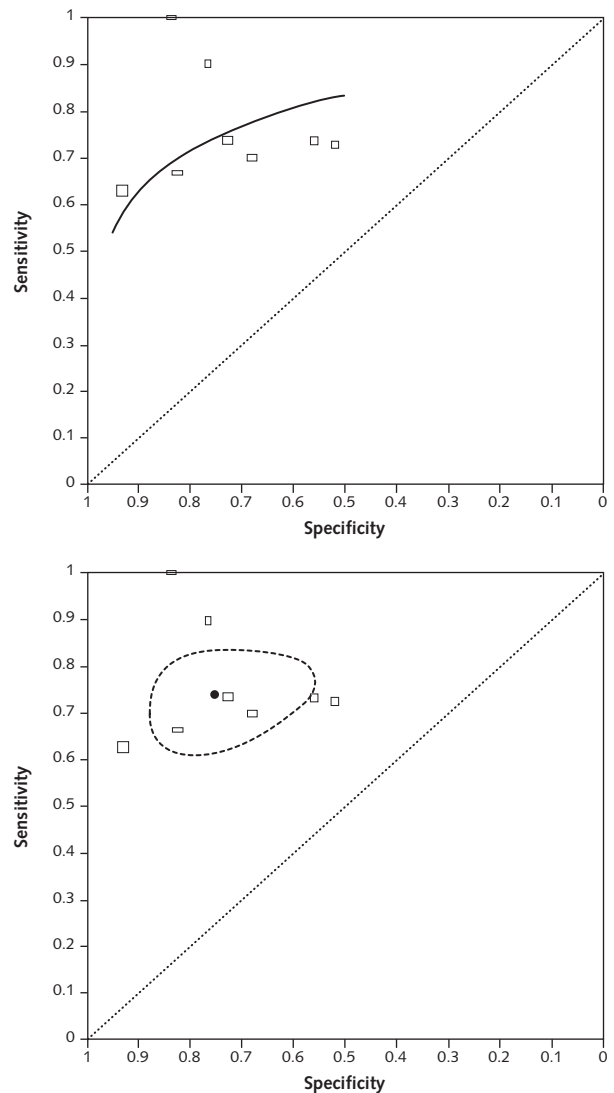
Because threshold effects cause sensitivity and specificity estimates to seem negatively correlated, and because threshold variation can be expected in many situations, robust approaches to meta-analysis take the underlying relationship between sensitivity and specificity into account. One way of doing so is by constructing a summary ROC curve. An average sensitivity and specificity point on this curve indicates where the center of the study results is. Separate pooling of sensitivity and specificity to identify this point has been discredited, because such an approach may identify a summary point that is not representative of the paired data (for example, a point that does not lie on the summary ROC curve).

Meta-analyses of studies reporting pairs of sensitivity and specificity estimates have often used the linear regression model for the construction of summary ROC curves proposed by Moses and colleagues (51), which is based on regressing the log diagnostic odds ratio against a measure of the proportion reported as positive. To examine differences between tests and to relate them to study or sample characteristics, the regression model can be extended by adding covariates (52). However, we now know that the formulation of the Moses model has its limitations. It fails to consider the precision of the study estimates, does not estimate between-study heterogeneity, and the explanatory variable in the regression is measured with error. These problems render estimates of CIs and *P* values unsuitable for formal inference (36, 53).

Two newly developed approaches to fitting random effects in hierarchical models overcome these limitations: the hierarchical summary ROC model (36, 54–56) and the bivariate random-effects model (53, 57). Both approaches model the distribution of the observed pairs of sensitivity and specificity values from each study. The hierarchical summary ROC model assumes an explicit formula linking sensitivity and specificity through a threshold; accounts for the variability across studies; and can be used to estimate summaries of the data, including a summary ROC curve and average values of accuracy measures, such as sensitivity and specificity. The bivariate random-effects model focuses on estimating the average sensitivity and specificity, but also estimates the unexplained variation in these parameters and the correlation between them. These 2 basic models are mathematically equivalent in the absence of covariates (58). Both models give a valid estimation of the underlying summary ROC curve and the average sensitivity and specificity (53, 58). Addition of covariates to the models, or application of separate models to different subgroups, enables exploration of heterogeneity. Both models can be fitted with statistical software for fitting mixed models (36, 53, 55, 57).

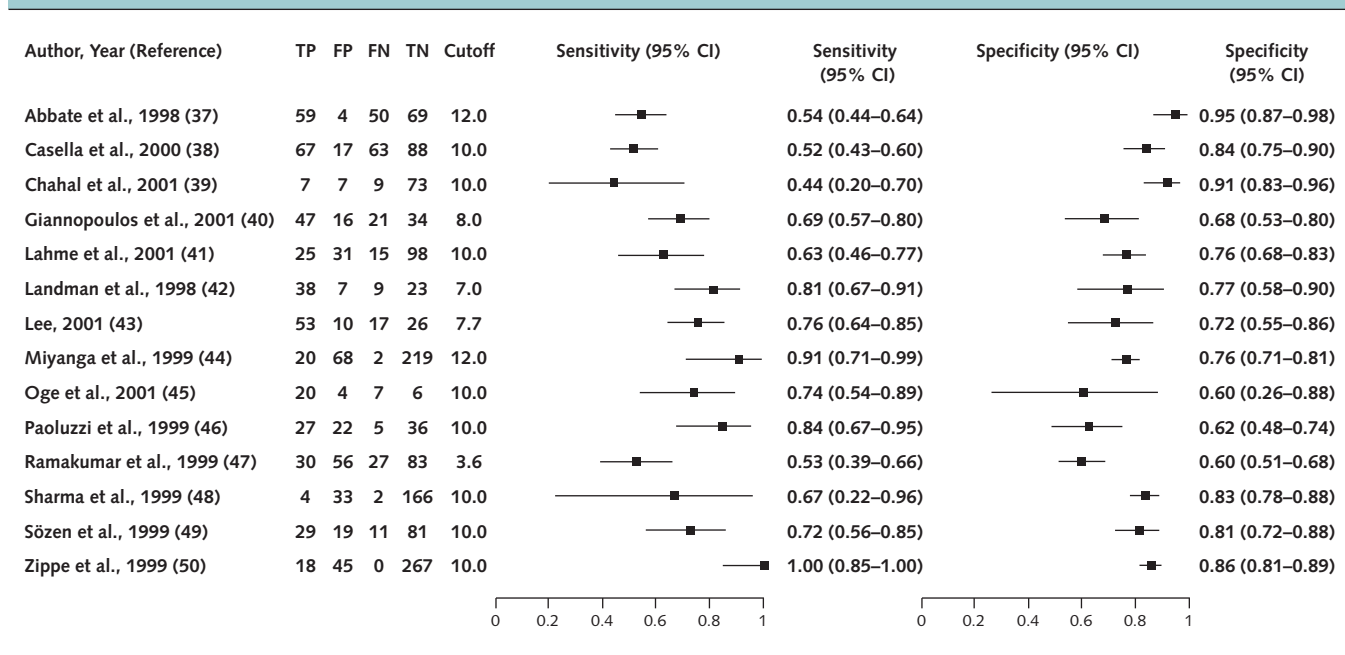
Estimates of summary likelihood ratios can best be derived from summary estimates of sensitivity and specificity obtained by using the methods described previously. Although some authors have advocated pooling likelihood ratios rather than sensitivity and specificity or ROC curves

Figure 2. Summary receiver-operating characteristic (ROC) curve plots showing test accuracy of a tumor marker for bladder cancer from 8 studies included in a systematic review.



Data from reference 10. Each study is represented by a small box positioned at the estimated sensitivity and specificity. The height and width of each box are proportional to the numbers of patients with and without bladder cancer, respectively, in each study. **Top.** This panel shows the summary ROC curve that can be drawn through these values. The scatter of the points fit, to a degree, with the existence of a threshold-type relationship between sensitivity and specificity. The curve is an estimate of the underlying relationship between sensitivity and specificity for the test used across varying thresholds. **Bottom.** This panel shows the average sensitivity and specificity estimate of the study results (solid circle) and a 95% confidence region around it. Estimation of a summary point only makes sense when the included studies have used a common threshold. The curves, points, and confidence regions can be estimated by using either the hierarchical summary ROC curve model (36, 54–56) or the bivariate random-effects model (53, 57).

Figure 3. Paired forest plot of the sensitivity and specificity of a tumor marker for bladder cancer.



FN = false-negative; FP = false-positive; TN = true-negative; TP = true-positive. Data are from reference 10. Forest plots document the extracted data for each study (numbers of TP, FP, FN, and TN results) together with estimates of sensitivity and specificity accompanied by 95% CIs. The scatter of the estimates and CIs indicates that the variability in sensitivity and specificity is unlikely to be explained by chance only, but it is not possible to ascertain whether a threshold-type relationship is evident.

(59–61), these methods do not account for the correlated bivariate nature of likelihood ratios and may yield impossible summary estimates and CIs, with positive and negative likelihood ratios either both above or both below 1.0 (62).

ROC Curves and Summary Estimates

The ability to estimate underlying summary ROC curves and average sensitivities and specificities allows flexibility in testing hypotheses and estimating diagnostic accuracy. Analyses based on all included studies facilitate well-powered comparisons between different tests or between subgroups of studies, which are not restricted to investigating accuracy at a particular threshold. The top panel of Figure 2 shows such a summary ROC curve for the diagnostic accuracy of a tumor antigen test for diagnosing bladder cancer. In contrast, when a test is being used at the same threshold in all included studies, review authors may make a summary estimate of sensitivity and specificity. The uncertainty associated with the estimate can be described by confidence regions marked on the summary ROC plot around the average point. The bottom panel of Figure 2 illustrates this approach.

Judgments about the validity of pooling data should be informed by considering the quality of the studies, the similarity of patients and tests being pooled, and whether the results may consequently be misleading. Where there is statistical heterogeneity in results, random-effects models are used to account for the variability and to derive suitably conservative assessments of the uncertainty in the estimates. Naturally, increased uncertainty about the estimates

may make it more difficult to draw firm conclusions about the accuracy of a particular test.

Comparative Analyses

Systematic reviews of diagnostic test accuracy may evaluate more than 1 test to determine which test or combination of tests can better serve the intended purpose. Indirect comparisons can be made by calculating separate summary estimates of the sensitivity and specificity for each test, including all studies that have evaluated that test regardless of whether they evaluated the other tests. The substantial variability that can be expected between tests means that such comparisons are prone to confounding. Restricting inclusion to studies of similar design and patient characteristics may limit confounding. A theoretically preferable approach is to use only studies that have directly compared the tests in the same patients or have randomly assigned patients to 1 of the tests. Such direct comparisons do not suffer from confounding. Paired analyses can be displayed in a ROC plot, by linking the sensitivity–specificity pairs from each study with a dashed line, as in Figure 4. Unfortunately, fully paired studies are not always available.

INTERPRETATION OF THE RESULTS

The interpretation of the results offered in the systematic review should help readers to understand the implications for practice. This interpretation should consider whether evidence derived from the review suitably ad-

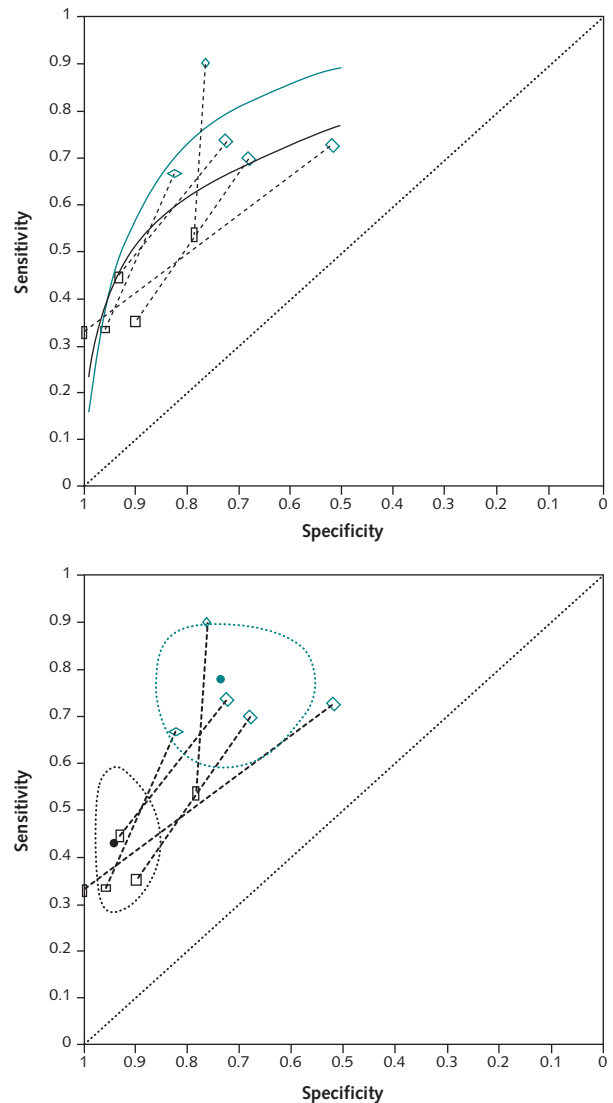
dresses the objectives of the review. It may involve considerations about whether the study sample was representative, the included studies indeed investigated the intended future role of the test under evaluation, and the results are unlikely to be biased. Review authors should consider the potential effects of quality differences on the results or the lack of high-quality studies. The interpretation of the findings should also consider the consequences of the false-positive and false-negative results and whether the estimates of accuracy are sufficiently high for the foreseen role that the test will have in practice. Some reviews may not result in useful summary estimates of sensitivity and specificity, for example, because of large variability in the individual study estimates. A decision model could be used to structure the interpretation of the findings. Such a model would incorporate important factors, such as the disease prevalence, probable outcomes, and available diagnostic and therapeutic interventions that may follow the test. Additional information, such as costs or important tradeoffs between harms and benefits, can be included (12).

CONCLUSION

The development of the methodology for systematic reviews of diagnostic test accuracy studies has made important progress in recent years. We now know more about searching, sources of bias in study design, quality appraisal, and data analysis. In meta-analysis, new hierarchical random-effects models have been developed with sound statistical properties that allow robust inferences. Methods for the estimation of summary ROC curves and summary estimates of sensitivity and specificity are now available. All these advances will be described in detail in the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy (63). The **Appendix Table** (available at www.annals.org) provides a summary of the key issues that both readers and review authors should consider.

Diagnostic test accuracy reviews face 2 major challenges. First, they are limited by the quality and availability of primary test accuracy studies that address important relevant questions. More studies are needed that recruit suitable spectrums of participants, make direct comparisons between tests, use rigorous methodology, and clearly report their methods and findings. Second, more development is needed in the area of interpretation and presentation of the results of diagnostic test accuracy reviews. Clinicians struggle with the definitions of sensitivity, specificity, and likelihood ratios (64, 65) possibly because, in the clinical context, the predictive value of tests is more immediately relevant. The results of systematic reviews of diagnostic accuracy can, of course, be used to assess the predictive value. Policymakers and guideline developers may be particularly interested in comparative accuracy, the costs and burden of testing, or new test methods. Developing systematic reviews that are relevant for policymakers and clin-

Figure 4. Meta-analysis of the diagnostic test accuracy of 2 index tests for bladder cancer: cytology (black squares) and bladder tumor antigen (green diamonds).



Data are from reference 10. The meta-analysis is restricted to studies that made a direct comparison between the tests by using both tests in each patient and comparing them with the reference standard of invasive cystoscopy. Restriction of the meta-analysis to direct test comparisons reduces concerns of confounding and allows stronger inferences to be drawn from the comparison of tests. The dashed lines link together the cytology and bladder tumor antigen results from each study and give the impression that bladder tumor antigen is much more sensitive but less specific than cytology. **Top.** Summary receiver-operating characteristic curves fitted to the data indicate that the bladder tumor antigen curve dominates the cytology curve as specificity decreases. Thus, bladder tumor antigen has the potential to be a more sensitive test than cytology, but only at specificities below 90%. **Bottom.** Cytology has an average sensitivity of 0.43 and an average specificity of 0.94 (black circle); bladder tumor antigen has an average sensitivity of 0.78 and an average specificity of 0.74 (green circle). The nonoverlapping 95% confidence regions indicate that the differences between the tests are unlikely to have occurred by chance alone.

ical practice poses a major challenge and requires clear thinking about the scope and purpose of the review.

From the Dutch Cochrane Centre and Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands; Unit of Public Health, Epidemiology and Biostatistics, University of Birmingham, Birmingham, United Kingdom; and Center for Statistical Sciences, Brown University, Providence, Rhode Island.

Grant Support: By the UK National Institute for Health Research (grant RNC/018/0003), the National Cancer Institute (grant 2U01CA079778), and the Cochrane Collaboration.

Potential Financial Conflicts of Interest: None disclosed.

Requests for Single Reprints: Jonathan J. Deeks, PhD, Unit of Public Health, Epidemiology and Biostatistics, School of Health and Population Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom; e-mail, j.deeks@bham.ac.uk.

Current author addresses are available at www.annals.org.

References

- Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Standards for Reporting of Diagnostic Accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Ann Intern Med.* 2003;138:40-4. [PMID: 12513043]
- The Cochrane Collaboration. The Cochrane Manual Issue 3, 2008 [updated 15 May 2008]. Oxford, UK: Cochrane Collaboration; 2008. Accessed at www.cochrane.org/admin/manual.htm on 18 July 2008.
- Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, et al. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med.* 1994;120:667-76. [PMID: 8135452]
- Knottnerus JA, ed. The Evidence Base of Clinical Diagnosis. London: BMJ Books; 2002.
- Bossuyt PM. Interpreting diagnostic test accuracy studies. *Semin Hematol.* 2008;45:189-95. [PMID: 18582626]
- Zhou X-H, Obuchowski N, McClish D. Statistical Methods in Diagnostic Medicine. Hoboken, NJ: J Wiley; 2002.
- Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ.* 2006;332:1089-92. [PMID: 16675820]
- Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? *Ann Intern Med.* 2006;144:850-5. [PMID: 16754927]
- Thornbury JR. Eugene W. Caldwell Lecture. Clinical efficacy of diagnostic imaging: love it or leave it. *AJR Am J Roentgenol.* 1994;162:1-8. [PMID: 8273645]
- Glas AS, Roos D, Deutekom M, Zwinderman AH, Bossuyt PM, Kurth KH. Tumor markers in the diagnosis of primary bladder cancer. A systematic review. *J Urol.* 2003;169:1975-82. [PMID: 12771702]
- Lokeshwar VB, Selzer MG. Urinary bladder tumor markers. *Urol Oncol.* 2006;24:528-37. [PMID: 17138134]
- Hunink M, Glasziou P, Siegel J, Weeks J, Pliskin J, Elstein A, Weinstein M. Decision Making in Health and Medicine: Integrating Evidence and Values. Cambridge, UK: Cambridge Univ Pr; 2001.
- Haynes RB, Wilczynski N, McKibbin KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. *J Am Med Assoc.* 1994;1:447-58. [PMID: 7850570]
- Deville WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. *J Clin Epidemiol.* 2000;53:65-9. [PMID: 10693905]
- Bachmann LM, Coray R, Estermann P, Ter Riet G. Identifying diagnostic studies in MEDLINE: reducing the number needed to read. *J Am Med Assoc.* 2002;9:653-8. [PMID: 12386115]
- Haynes RB, Wilczynski NL. Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. *BMJ.* 2004;328:1040. [PMID: 15073027]
- Doust JA, Pietrzak E, Sanders S, Glasziou PP. Identifying studies for systematic reviews of diagnostic tests was difficult due to the poor sensitivity and precision of methodologic filters and the lack of information in the abstract. *J Clin Epidemiol.* 2005;58:444-9. [PMID: 15845330]
- Leeflang MM, Scholten RJ, Rutjes AW, Reitsma JB, Bossuyt PM. Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. *J Clin Epidemiol.* 2006;59:234-40. [PMID: 16488353]
- Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ. Publication and related biases. *Health Technol Assess.* 2000;4:1-115. [PMID: 10932019]
- Song F, Khan KS, Dinnes J, Sutton AJ. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. *Int J Epidemiol.* 2002;31:88-95. [PMID: 11914301]
- Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol.* 2005;58:882-93. [PMID: 16085191]
- Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: literature survey. *BMJ.* 2006;332:1127-9. [PMID: 16627488]
- Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. *BMJ.* 2002;324:669-71. [PMID: 11895830]
- Lijmer JG, Mol BW, Heisterkamp S, Bossuyt PM, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA.* 1999;282:1061-6. [PMID: 10493205]
- Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med.* 2004;140:189-202. [PMID: 14757617]
- Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ.* 2006;174:469-76. [PMID: 16477057]
- Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, 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. [PMID: 14606960]
- Leeflang MM, Moons KG, Reitsma JB, Zwinderman AH. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. *Clin Chem.* 2008;54:729-37. [PMID: 18258670]
- Ewald B. Post hoc choice of cut points introduced bias to diagnostic research. *J Clin Epidemiol.* 2006;59:798-801. [PMID: 16828672]
- Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. *BMC Med Res Methodol.* 2005;5:19. [PMID: 15918898]
- Whiting P, Harbord R, Main C, Deeks JJ, Filippini G, Egger M, et al. Accuracy of magnetic resonance imaging for the diagnosis of multiple sclerosis: systematic review. *BMJ.* 2006;332:875-84. [PMID: 16565096]
- Westwood ME, Whiting PF, Kleijnen J. How does study quality affect the results of a diagnostic meta-analysis? *BMC Med Res Methodol.* 2005;5:20. [PMID: 15943861]
- Leeflang M, Reitsma J, Scholten R, Rutjes A, Di Nisio M, Deeks J, et al. Impact of adjustment for quality on results of metaanalyses of diagnostic accuracy. *Clin Chem.* 2007;53:164-72. [PMID: 17185365]
- Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Reitsma JB, Bossuyt PM, et al. Quality of reporting of diagnostic accuracy studies. *Radiology.* 2005;235:347-53. [PMID: 15770041]
- Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Bossuyt PM, Reitsma JB, et al. The quality of diagnostic accuracy studies since the STARD statement: has it improved? *Neurology.* 2006;67:792-7. [PMID: 16966539]
- Gatsonis C, Paliwal P. Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer. *AJR Am J Roentgenol.* 2006;187:271-81. [PMID: 16861527]
- Abbate I, D'Introno A, Cardo G, Marano A, Addabbo L, Musci MD, et al. Comparison of nuclear matrix protein 22 and bladder tumor antigen in urine of patients with bladder cancer. *Anticancer Res.* 1998;18:3803-5. [PMID: 9854500]
- Casella R, Huber P, Blöchliger A, Stoffel F, Dalquen P, Gasser TC, et al. Urinary level of nuclear matrix protein 22 in the diagnosis of bladder cancer:

- experience with 130 patients with biopsy confirmed tumor. *J Urol*. 2000;164:1926-8. [PMID: 11061883]
39. Chahal R, Darshane A, Browning AJ, Sundaram SK. Evaluation of the clinical value of urinary NMP22 as a marker in the screening and surveillance of transitional cell carcinoma of the urinary bladder. *Eur Urol*. 2001;40:415-20; discussion 421. [PMID: 11713396]
40. Giannopoulos A, Manousakas T, Gounari A, Constantinides C, Choremi-Papadopoulou H, Dimopoulos C. Comparative evaluation of the diagnostic performance of the BTA stat test, NMP22 and urinary bladder cancer antigen for primary and recurrent bladder tumors. *J Urol*. 2001;166:470-5. [PMID: 11458049]
41. Lahme S, Bichler KH, Feil G, Krause S. Comparison of cytology and nuclear matrix protein 22 for the detection and follow-up of bladder cancer. *Urol Int*. 2001;66:72-7. [PMID: 11223747]
42. Landman J, Chang Y, Kavalier E, Droller MJ, Liu BC. Sensitivity and specificity of NMP-22, telomerase, and BTA in the detection of human bladder cancer. *Urology*. 1998;52:398-402. [PMID: 9730450]
43. Lee KH. Evaluation of the NMP22 test and comparison with voided urine cytology in the detection of bladder cancer. *Yonsei Med J*. 2001;42:14-8. [PMID: 11293494]
44. Miyayaga N, Akaza H, Tsukamoto T, Ishikawa S, Noguchi R, Ohtani M, et al. Urinary nuclear matrix protein 22 as a new marker for the screening of urothelial cancer in patients with microscopic hematuria. *Int J Urol*. 1999;6:173-7. [PMID: 10226833]
45. Oge O, Atsü N, Kendi S, Ozen H. Evaluation of nuclear matrix protein 22 (NMP22) as a tumor marker in the detection of bladder cancer. *Int Urol Nephrol*. 2001;32:367-70. [PMID: 11583354]
46. Paoluzzi M, Cuttano MG, Mugnaini P, Salsano F, Giannotti P. Urinary dosage of nuclear matrix protein 22 (NMP22) like biologic marker of transitional cell carcinoma (TCC): a study on patients with hematuria. *Arch Ital Urol Androl*. 1999;71:13-8. [PMID: 10193018]
47. Ramakumar S, Bhuiyan J, Besse JA, Roberts SG, Wollan PC, Blute ML, et al. Comparison of screening methods in the detection of bladder cancer. *J Urol*. 1999;161:388-94. [PMID: 9915409]
48. Sharma S, Zippe CD, Pandrangi L, Nelson D, Agarwal A. Exclusion criteria enhance the specificity and positive predictive value of NMP22 and BTA stat. *J Urol*. 1999;162:53-7. [PMID: 10379739]
49. Sözen S, Biri H, Sinik Z, Küpeli B, Alkibay T, Bozkirli I. Comparison of the nuclear matrix protein 22 with voided urine cytology and BTA stat test in the diagnosis of transitional cell carcinoma of the bladder. *Eur Urol*. 1999;36:225-9. [PMID: 10450007]
50. Zippe C, Pandrangi L, Agarwal A. NMP22 is a sensitive, cost-effective test in patients at risk for bladder cancer. *J Urol*. 1999;161:62-5. [PMID: 10037369]
51. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med*. 1993;12:1293-316. [PMID: 8210827]
52. Lijmer JG, Bossuyt PM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Stat Med*. 2002;21:1525-37. [PMID: 12111918]
53. Arends LR, Hamza TH, van Houwelingen JC, Heijnenbrok-Kal MH, Hunink MGM, Stijnen T. Bivariate random effects meta-analysis of ROC curves. *Med Decis Making* 2008;28:621-638. [PMID: 18591542]
54. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med*. 2001;20:2865-84. [PMID: 11568945]
55. Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *J Clin Epidemiol*. 2004;57:925-32. [PMID: 15504635]
56. Dukic V, Gatsonis C. Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. *Biometrics*. 2003;59:936-46. [PMID: 14969472]
57. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol*. 2005;58:982-90. [PMID: 16168343]
58. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics*. 2007;8:239-51. [PMID: 16698768]
59. Stengel D, Bauwens K, Sehouli J, Ekkernkamp A, Porzsolt F. A likelihood ratio approach to meta-analysis of diagnostic studies. *J Med Screen*. 2003;10:47-51. [PMID: 12790315]
60. Khan KS. Systematic reviews of diagnostic tests: a guide to methods and application. *Best Pract Res Clin Obstet Gynaecol*. 2005;19:37-46. [PMID: 15749064]
61. Khan KS, Dinnes J, Kleijnen J. Systematic reviews to evaluate diagnostic tests. *Eur J Obstet Gynecol Reprod Biol*. 2001;95:6-11. [PMID: 11267714]
62. Zwinderman AH, Bossuyt PM. We should not pool diagnostic likelihood ratios in systematic reviews. *Stat Med*. 2008;27:687-97. [PMID: 17611957]
63. Deeks JJ, Bossuyt PM, Gatsonis C (editors). *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.0*. Oxford, UK: The Cochrane Collaboration, 2009. [Forthcoming].
64. Steurer J, Fischer JE, Bachmann LM, Koller M, ter Riet G. Communicating accuracy of tests to general practitioners: a controlled study. *BMJ*. 2002;324:824-6. [PMID: 11934776]
65. Puhan MA, Steurer J, Bachmann LM, ter Riet G. A randomized trial of ways to describe test accuracy: the effect on physicians' post-test probability estimates. *Ann Intern Med*. 2005;143:184-9. [PMID: 16061916]

Current Author Addresses: Drs. Leeflang and Bossuyt: Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, PO Box 22700, 1100 DE Amsterdam, the Netherlands.

Dr. Deeks: Unit of Public Health, Epidemiology and Biostatistics, School of Health and Population Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom.

Dr. Gatsonis: Center for Statistical Sciences, Brown University, Box G-S121, 121 South Main Street, 7th Floor, Providence, RI 02912.

APPENDIX: CONTRIBUTORS TO THE COCHRANE DIAGNOSTIC TEST ACCURACY WORKING GROUP

Contributors are listed in alphabetical order.

Bert Aertgeerts, Doug Altman, Gerd Antes, Lucas Bachmann, Patrick Bossuyt, Heiner Buchner, Peter Bunting, Frank Buntinx, Jonathan Craig, Roberto D'Amico, Riekje de Vet, Jon Deeks, Jenny Doust, Matthias Egger, Anne Eisinga, Graziella Fillipini, Yngve Flack-Ytter, Constantine Gatsonis, Afina Glas, Paul Glasziou, Fritz Grossenbacher, Roger Harbord, Jorgen Hilden, Lotty Hooft, Andrea Horvath, Chris Hyde, Les Irwig, Monica Kjeldstrøm, Petra Macaskill, Susan Mallett, Ruth Mitchell, Tess Moore, Rasmus Moustgaard, Wytze Oosterhuis, Madhukar Pai, Prashni Paliwal, Daniel Pewsner, Hans Reitsma, Jacob Riis, Ingrid Riphagen, Anne Rutjes, Rob Scholten, Nynke Smidt, Jonathan Sterne, Yemisi Takwoingi, Danielle van der Windt, Vasivy Vlassov, Joseph Watine, and Penny Whiting.

Appendix Table. Essential Elements in a Systematic Review of Diagnostic Test Accuracy

Phase in Review Process	Key Issues
1. Definition of the review objectives	To identify the review question: State the patient group and define presenting condition(s), previous test results, and health care setting. Describe the tests (or test strategies) under evaluation, specifying their intended roles. Identify tests and test strategies currently used in practice for comparison, if available. Define the target condition to be diagnosed and reference standards to be used.
2. Study identification and selection	Search several electronic databases. Use a search strategy built around terms for the index test, target condition, and possibly patient characteristics. Do not use restrictive methodological search filters.
3. Quality assessment	Identify biases for which the included studies are at risk. Use the QUADAS checklist as a tool for identifying many common deficiencies. Comment on the adequacy of each aspect of study design. Do not use summary quality scores.
4. Data extraction, analysis, and presentation	Extract paired estimates of test sensitivity and specificity from each study overall and, if available, for patient subgroups. Plot studies in ROC space to identify the location, variability, and correlations. The hierarchical summary ROC and bivariate random-effects models provide a sound statistical framework for analysis, accounting for sampling variability, unexplained heterogeneity, and covariation between sensitivity and specificity. Compute average values of sensitivity and specificity when the data combined share a common threshold. Use summary ROC curves to describe test performance and to compare tests without restricting to particular thresholds. Obtain estimates of summary likelihood ratios from average values of sensitivity and specificity and not through separate pooling of likelihood ratios. Global tests for heterogeneity before data synthesis or tests for publication bias are typically not useful. Meta-analyze and present studies that compare tests by using randomized or within-patient designs separately from the results of indirect comparisons.
5. Interpretation	Consider the consequences of using the test, in terms of (changes in) the numbers of true-positive, false-positive, true-negative, and false-negative test results with the expected prevalence of the target disorder. Address the applicability of the results in terms of whether the patients in the primary studies were similar to those outlined in the objective, and whether tests and test strategies evaluated and compared were representative of test strategies that are used in practice. Address to what extent the original studies were biased and how these biases could influence the results and the degree to which comparisons between tests may be confounded. Consider complementing the interpretation with decision modeling by using results of the review.

QUADAS = Quality Assessment of Diagnostic Accuracy Studies; ROC = receiver-operating characteristic.