Published in final edited form as:
Ann Intern Med. 2008 December 16; 149(12): 889–897.

Systematic Reviews of Diagnostic Test Accuracy

Mariska.M.G. Leeflang, PhD[1,2], **Jonathan J. Deeks, PhD**[*,3], **Constantine Gatsonis, PhD**[4], and **Patrick M.M. Bossuyt, PhD**[2]

[1]The Dutch Cochrane Centre, Academic Medical Center, University of Amsterdam, The Netherlands [2]Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, The Netherlands [3]Department of Public Health and Epidemiology, University of Birmingham, United Kingdom [4]Center for Statistical Sciences, Brown University, Providence, USA

Abstract

Systematic reviews of diagnostic test accuracy studies are increasingly being published, but they can be methodologically challenging. In this paper we 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 the paired nature of the estimates and their dependence on threshold into account, we therefore advice authors of these reviews to use the hierarchical summary ROC or the bivariate model for the analysis of the data for the analysis. Challenges that remain are the poor reporting of original diagnostic test accuracy research, and difficulties with the interpretation of the results of diagnostic test accuracy research.

Introduction

Diagnostic tests are a critical component of health care, and clinicians, policy makers and patients routinely face a range of questions regarding diagnostic tests. They want to know if testing improves outcome, would like to know what test to use, to purchase, or to 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 Standards for Reporting of Diagnostic Accuracy (STARD) statement (1). That 25 item checklist was

[*]Author for correspondence Prof Jonathan J Deeks, Department of Public Health and Epidemiology, University of Birmingham, Edgbaston, Birmingham B15 2TT UK, Tel: +44-(0)-121-414-5328, J.Deeks@bham.ac.uk.

Contributors to the Cochrane Diagnostic Test Accuracy Working Group include (in alphabetical order):
Bert Aertgeerts, Doug Altman, Gerd Antes, Lucas Bachmann, Patrick Bossuyt, Heiner Buchner, Peter Bunting, Frank Buntinx, Jonathan Craig, Roberto D'Amico, 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, Riekie de Vet, Vasivy Vlassov, Joseph Watine, Danielle van der Windt, Penny Whiting.

published in this and many other journals, and is now adopted by more than 200 scientific journals.

As elsewhere in 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 summary estimates with a stronger generalizability than estimates from a single study. Systematic reviews may help identify the risk of 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). They decided in 2003 to make preparations for including systematic reviews of diagnostic test accuracy in their Cochrane Database of Systematic Reviews (CDSR). To enable this, a working group was constituted to develop methodology, software and a handbook (see Appendix). The first diagnostic test accuracy reviews will be published in the CDSR 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 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 et al in this journal (3).

Definition of the objectives of the review

Diagnostic test accuracy refers to the ability of that test to distinguish between patients with disease (or more generally, a specified target condition) and those without. In such a test accuracy study, the results of the test under evaluation, or ‘index test’, are compared with those of the reference standard determined in the same patients. The reference standard is the best available method for identifying patients that have the target condition. Test accuracy is most often expressed as the test’s sensitivity (the proportion of those positive to the reference standard who are also positive to the index test) and specificity (the proportion of those negative to the reference standard who are also negative to the index test), but many alternative measures have been proposed and are in use (4,5).

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, with the test interpreters, and may depend on the results of prior testing. For this reason, it is essential to include these elements in the study question. Review authors should at least consider whether the test of interest will be mainly used in general practice or in a secondary or even tertiary setting. If the index test is physical examination for example, a test more important for family practice than for specialized care, then the review authors must realize that their review may be of limited value if the included studies are all done in a tertiary setting.

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 has equivalent accuracy but offers other advantages (6–8). As with the evaluation of interventions, systematic reviews need to include comparative analyses between alternative testing strategies, and not focus solely on evaluating the performance of a test in isolation.

In relation to the existing situation, three possible roles for a new test can be defined: replacement, triage, and add-on (6). 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 existing 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, one will be looking for a test that gives a minimal proportion of false negatives and thus a relatively high sensitivity. 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 positives or false negatives 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 (9). Clinicians may use cytology to triage patients before they undergo invasive cystoscopy, the reference standard for bladder cancer. As cytology combines a high specificity with a low sensitivity (10), the goal of the review was to identify a tumor marker with sufficient accuracy to either replace cytology or to be used in addition to cytology. For a marker to replace cytology, it has to achieve equally high specificity with improved sensitivity. New markers which 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 one or more of the markers, cytology only, or a combination of markers and cytology.

Although information on accuracy can help clinicians in making decisions about tests, review authors and readers should realize that good diagnostic accuracy is a desirable but not a sufficient condition for the effectiveness of a test (7). To show 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. In most cases, such randomized trials are rare and systematic reviews of test accuracy may provide the most useful evidence to guide decision making, and provide key evidence to incorporate into decision models.

Identification and selection of studies

Identifying test accuracy studies is more difficult than searching for randomized trials (11). There is not a clear, unequivocal key word or indexing term for an accuracy study in literature databases, comparable to 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 additional manual searches, such as screening reference lists, may be necessary.

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 it may be necessary to add the patient description, 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, possibly, also haematuria.

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 (11–14). 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, so they are not recommended for systematic reviews (15,16). The incremental value of searching in languages other than English and in the so called grey literature has not yet been fully investigated.

In systematic reviews of intervention studies, publication bias is an important and well-studied form of bias, where 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 (17). A consistent observation has been that studies with statistically significant results are more likely to be published than studies with non-significant findings (17). Investigating publication bias for diagnostic tests is problematic, as many studies are undertaken without ethical review or study registration, so follow-up of cohorts of studies is not well possible (18). 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 (19). Also, as results of test accuracy studies do not routinely report P-values and dichotomize findings as significant or not significant, the determinants for publication of diagnostic studies are unlikely to be the same as the determinants for publication of intervention studies.

Assessment of methodological quality

More variability among diagnostic accuracy study results is to be expected than with randomized trials. Some of this variability is due to chance, as many diagnostic studies have small sample sizes (20). 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 (21). Test accuracy studies with design deficiencies can produce biased results (22–24). Table 1 lists some of the more important forms of bias. Sources of bias for which there is unambiguous evidence that these can overestimate diagnostic accuracy are the inclusion of healthy controls and the incomplete or differential use of reference standards (22,24).

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 Quality Assessment of Diagnostic Accuracy Studies (QUADAS) checklist to assess the quality of diagnostic test accuracy studies (25). 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 cut-off value may bias diagnostic accuracy (26,27). Review authors should therefore think carefully 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, as different shortcomings may generate different magnitudes of bias, even in opposing directions, making it very hard to attach sensible weights to each quality item (28). A way to summarize the quality assessment is shown in Figure 1, where stacked bars are 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 single study. In the analysis phase, the results of the quality appraisal may guide explorations of the sources of heterogeneity (30,31). 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 low for meaningful investigations. Also, incomplete reporting hampers any evaluation of study quality (32). The effects of the STARD guidelines for complete and transparent reporting (1) are only gradually becoming visible in the literature (33).

Analyzing the data and presenting the results

Whereas the results of a randomized trial are often reported 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 two 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 (34,35).

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 a ROC space (see 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 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 the diseased and the non-diseased. It is expected that most studies will lie above this line. The best diagnostic tests will be positioned in the upper right corner of the ROC space, where both sensitivity and specificity are close to 1. As confidence limits are not displayed on these plots it is not possible to discern the cause of scatter across studies – it can be caused either due to small sample sizes or between study heterogeneity. Paired forest plots (see Figure 3) display sensitivity and specificity separately (but on the same row) for each study together with confidence intervals 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 when plotted in an ROC plot. A major contributor to this appearance is the trade-off between sensitivity and specificity when the threshold for defining test positivity varies. When high test results are labelled as 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, a 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 is said to have 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 where test results are sensitive to machine calibration.

Because threshold effects cause sensitivity and specificity estimates to appear 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 ‘operating point’ on this curve indicates where the centre of the study results lie. Separate pooling of sensitivity and specificity to identify this point has been discredited, because such an approach may identify a summary point which is not representative of the paired data, for example a point which does not lie on the summary ROC curve.

Meta-analyses of studies reporting pairs of sensitivity and specificity estimates often used the linear regression model for the construction of summary ROC curves proposed by Moses et al, which is based on regressing the log diagnostic odds ratio against a measure of the proportion

reported as test positive (36). To examine differences between tests and to relate them to study or sample characteristics, the regression model can be extended by adding covariates (37). 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 confidence intervals and *P*-values unsuitable for formal inference (35,38).

Two newly developed approaches to fitting random effects in hierarchical models overcome these limitations: the hierarchical summary ROC model (35,39–41) and the bivariate random effects model (38,42). The hierarchical summary ROC model focuses on identifying the underlying ROC curve, estimating the average accuracy (as a diagnostic odds ratio) and average threshold (and unexplained variation in these parameters across studies), together with a shape parameter that describes the asymmetry in the curve. 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 two basic models are mathematically equivalent in the absence of covariates (43). Both models give a valid estimation of the underlying summary ROC curve and the average operating point (38,43). 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 (35,38,40,42).

Estimates of summary likelihood ratios can best be derived from summary estimates of sensitivity and specificity obtained using the methods described above. Whilst some authors have advocated pooling likelihood ratios rather than pooling sensitivity and specificity or ROC curves (44–46), these methods do not account for the correlated bivariate nature of likelihood ratios, and may yield impossible summary estimates and confidence intervals, with positive and negative likelihood ratios either both above or both below 1 (47).

Curves or summary estimates?

The ability to estimate underlying summary ROC curves and average operating points 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. Figure 2a shows 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 estimate a summary estimate of sensitivity and specificity. The certainty associated with the estimate can be described by confidence regions marked on the summary ROC plot around the average point. Figure 2b shows an example of 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 will describe the variability and uncertainty in the estimates which may lead to difficulties in drawing firm conclusions about the accuracy of a particular test.

Comparative analyses

Systematic reviews of diagnostic test accuracy may evaluate more than one tes,t 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. An theoretically preferable approach is to only use studies that have directly compared the tests in the same patients, or have randomized patients to one of the tests. Such direct comparisons do not suffer from confounding. Paired analyses can be displayed in an ROC plot, by linking the sensitivity-specificity pairs from each study with a dashed or dotted 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 addresses the objectives of the review. It may involve considerations about whether the study sample was representative, whether the included studies indeed investigated the intended future role of the test under evaluation, and whether the results are unlikely to be biased. The potential effects of quality differences on the results, or the lack of high quality studies should be considered. The interpretation of the findings should furthermore consider the consequences of the false positive and false negative test results and whether the estimates of accuracy that were found 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, or because the authors only investigated the comparative accuracy by comparing summary ROC curves. A decision model could be used to structure the interpretation of the findings. Such a model would incorporate important factors as the disease prevalence, likely outcomes, and the available diagnostic and therapeutic interventions that may follow the test (48). Additional information, such as costs or important trade-offs between harms and benefits can be included.

Conclusion

The development of the methodology for systematic reviews of diagnostic test accuracy studies has progressed importantly in recent years. We now know more about searching, about sources of bias in study design, and about quality appraisal, and about 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 of summary estimates of sensitivity and specificity are now available. All these advances will be described in detail in the Cochrane Handbook for Diagnostic Test Accuracy Reviews (49). Table 2 provides a summary of the key issues that both readers and reviews authors should think of.

Diagnostic test accuracy reviews face two major challenges. Firstly, they are limited by the quality and availability of primary test accuracy studies that address important relevant questions. More studies are needed which recruit suitable spectrums of participants, make direct comparisons between tests, use rigorous methodology, and clearly report their methods and findings. Secondly, more development is needed in the area of interpretation and presentation of the results of diagnostic test accuracy reviews. It has been shown that many clinicians struggle with the definitions of sensitivity, specificity and likelihood ratios (50,51). We have to explore how well the concept of diagnostic accuracy applies to other forms of testing, such as prognosis, prediction and monitoring, and to new test modalities, such as microarrays and genotyping. Policy makers and guideline developers may be interested in the comparative accuracy only, as well as in additional information, such as the costs and burden of testing, or in new test modalities. Developing systematic reviews that are relevant for policy

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

References

1. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D. de Vet HC; Standards for Reporting of Diagnostic Accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med 2003;138 (1):40–44. [PubMed: 12513043]
2. The Cochrane Collaboration. The Cochrane Manual Issue 3. 2008 [accessed 18th July 2008]. [updated 15 May 2008] http://www.cochrane.org/admin/manual.htm
3. Irwig L, Tosteson AN, Gatsonis CA, Lau J, Colditz G, Chalmers TC, Mosteller F. Guidelines for meta-analyses evaluating diagnostic tests. Ann. Int. Med 1994;120:667–676. [PubMed: 8135452]
4. Knottnerus, JA., editor. The Evidence Base of Clinical Diagnosis. London: BMJ Books; 2002.
5. Bossuyt PM. Interpreting diagnostic test accuracy studies. Semin Hematol 2008;45(3):189–195. [PubMed: 18582626]
6. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006;332(7549):1089–1092. [PubMed: 16675820]
7. 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(11):850–855. [PubMed: 16754927]
8. Thornbury JR. Clinical efficacy of diagnostic imaging: love it or leave it. AJR 1994;162:1–8. [PubMed: 8273645]
9. 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(6):1975–1982. [PubMed: 12771702]
10. Lokeshwar VB, Slezer MB. Urinary bladder tumour markers. Urol Oncol 2006;24(6):528–537. [PubMed: 17138134]
11. Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. J Am Med Inform Assoc 1994;1(6): 447–458. [PubMed: 7850570]
12. Deville WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. J Clin Epidemiol 2000;53:65–69. [PubMed: 10693905]
13. Bachmann LM, Coray R, Estermann P, Ter Riet G. Identifying diagnostic studies in MEDLINE: reducing the number needed to read. JAMA 2002;9:653–658.
14. Haynes RB, Wilczynski NL. Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. BMJ 2004;328(7447):1040. [PubMed: 15073027]
15. 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(5):444–449. [PubMed: 15845330]
16. 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(3):234–240. [PubMed: 16488353]
17. Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ. Publication and related biases. Health Technol Assess 2000;4(10):1–115.
18. 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(1):88–95. [PubMed: 11914301]
19. 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(9): 882–893. [PubMed: 16085191]
20. Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: literature survey. BMJ 2006 May 13;332(7550):1127–1129. [PubMed: 16627488]

21. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. BMJ 2002;324(7338):669–671. [PubMed: 11895830]
22. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, Bossuyt PM. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999 Sep 15;282(11): 1061–1066. [PubMed: 10493205]
23. 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(3):189–202. [PubMed: 14757617]
24. 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 Feb 14;174(4):469–476. [PubMed: 16477057]
25. 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. [PubMed: 14606960]
26. 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 [Epub ahead of print].
27. Ewald B. Post hoc choice of cut points introduced bias to diagnostic research. J Clin Epidemiol 2006;59(8):798–801. [PubMed: 16828672]
28. 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. [PubMed: 15918898]
29. Whiting P, Harbord R, Main C, Deeks JJ, Filippini G, Egger M, Sterne JA. Accuracy of magnetic resonance imaging for the diagnosis of multiple sclerosis: systematic review. BMJ 2006;332(7546): 875–884. [PubMed: 16565096]
30. Westwood ME, Whiting PF, Kleijnen J. How does study quality affect the results of a diagnostic meta-analysis? BMC Med Res Methodol 2005 Jun 8;5(1):20. [PubMed: 15943861]
31. Leeflang M, Reitsma J, Scholten R, Rutjes A, Di Nisio M, Deeks J, Bossuyt P. Impact of adjustment for quality on results of metaanalyses of diagnostic accuracy. Clin Chem 2007;53(2):164–172. [PubMed: 17185365]
32. Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Reitsma JB, Bossuyt PM, Bouter LM, de Vet HC. Quality of reporting of diagnostic accuracy studies. Radiology 2005;235(2):347–353. [PubMed: 15770041]
33. Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Bossuyt PM, Reitsma JB, Bouter LM, de Vet HC. The quality of diagnostic accuracy studies since the STARD statement: has it improved? Neurology 2006 Sep 12;67(5):792–797. [PubMed: 16966539]
34. Zhou, X-H.; Obuchowski, N.; McClish, D. Statistical methods in diagnostic medicine. Wiley; 2002.
35. Gatsonis C, Paliwal P. Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer. AJR Am J Roentgenol 2006;187(2):271–281. [PubMed: 16861527]
36. Moses LE, Littenberg B, Shapiro D. Combining Independent Studies of a Diagnostic Test Into a Summary ROC Curve: Data--Analytic Approaches and Some Additional Considerations. Stat Med 1993;12:1293–1316. [PubMed: 8210827]
37. Lijmer JG, Bossuyt PM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. Stat Med 2002;21(11):1125–1137.
38. Arends, LR. Multivariate meta-analysis: modeling the heterogeneity Mixing apples and oranges: dangerous or delicious?. Haveka BV: Alblasserdam; 2006.
39. Rutter C, Gatsonis C. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med 2001;20:2865–2884. [PubMed: 11568945]
40. 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(9):925–932. [PubMed: 15504635]
41. Dukic V, Gatsonis C. Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. Biometrics 2003;59(4):936–946. [PubMed: 14969472]
42. 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(10):982–990. [PubMed: 16168343]

43. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics 2007;8(2):239–251. [PubMed: 16698768]
44. Stengel D, Bauwens K, Sehouli J, Ekkernkamp A, Porzsolt F. A likelihood ratio approach to meta-analysis of diagnostic studies. J Med Screening 2003;10(1):47–51.
45. Khan KS. Systematic reviews of diagnostic tests: a guide to methods and application. Best Pract Res Clin Obstet Gynaecol 2005;19(1):37–46. [PubMed: 15749064]
46. Khan KS, Dinnes J, Kleijnen J. Systematic reviews to evaluate diagnostic tests. Eur J Obstet Gynecol Reprod Biol 2001;95(1):6–11. [PubMed: 11267714]
47. Zwinderman AH, Bossuyt PM. We should not pool diagnostic likelihood ratios in systematic reviews. Stat Med 2008;27(5):687–697. [PubMed: 17611957]
48. …
49. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.0. The Cochrane Collaboration. 2008. in press
50. Steurer J, Fischer JE, Bachmann LM, Koller M, ter Riet G. Communicating accuracy of tests to general practitioners: a controlled study. BMJ 2002;324(7341):824–826. [PubMed: 11934776]
51. 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(3):184–189. [PubMed: 16061916]

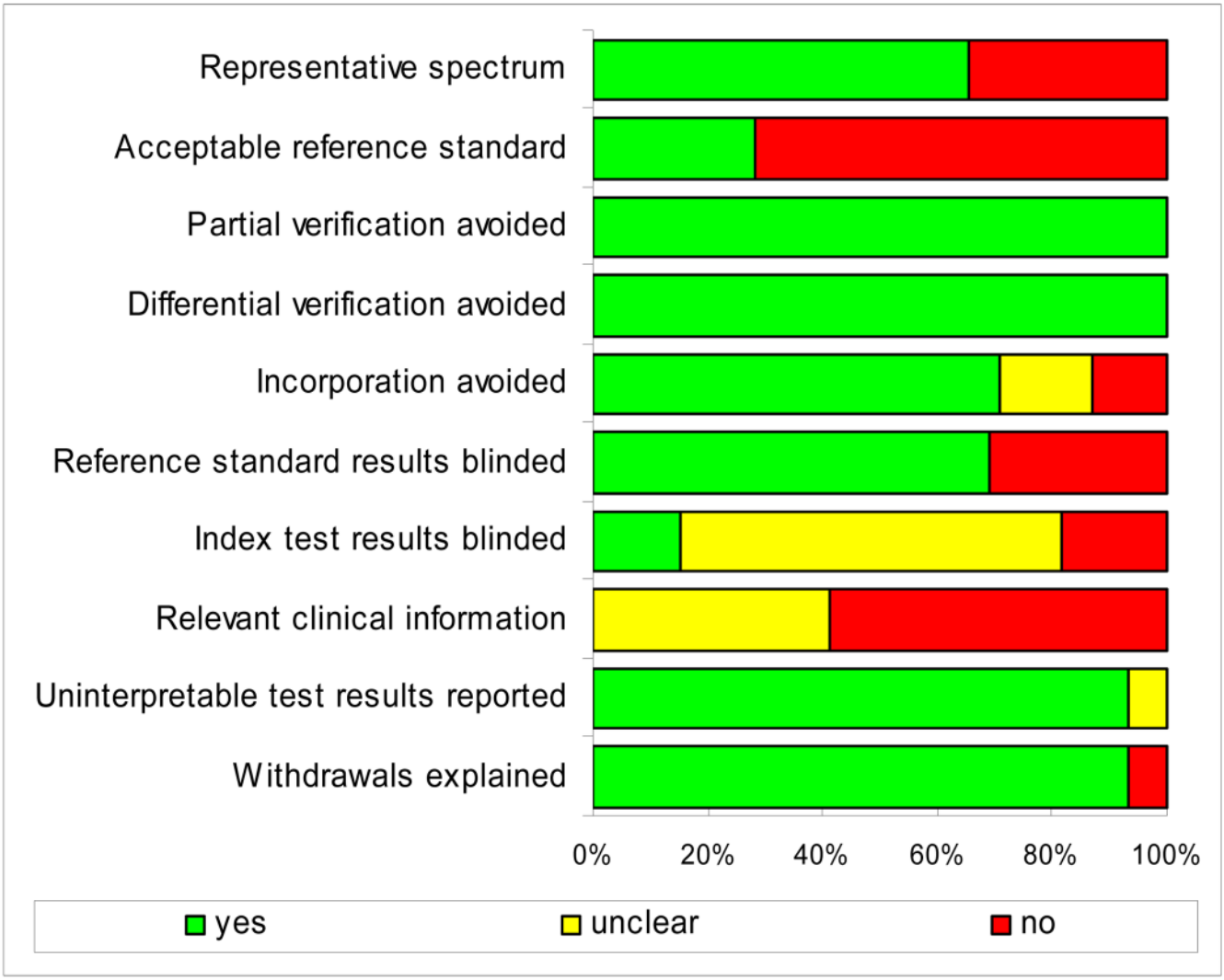

Figure 1.
Review authors' judgments about quality items presented as percentages across all included studies. Based on a re-analysis of data from a systematic review on magnetic resonance imaging for multiple sclerosis[29]. The item "acceptable delay between tests" did not apply in this review. The authors considered the relative lack of acceptable reference standard as the main weakness of the review.

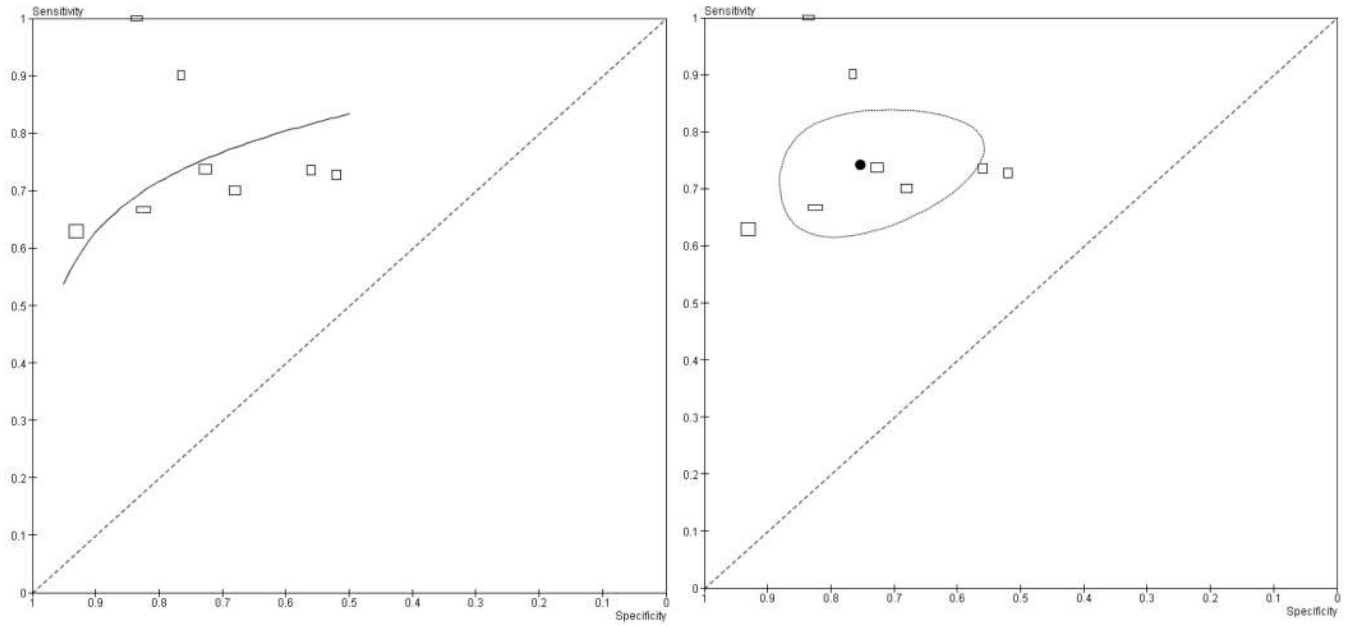

Figure 2.

a and b: ROC showing pairs of sensitivity and specificity values for the included studies. The height of the rectangles is proportional to the number of patients with bladder cancer across studies, the width of the rectangles corresponds to the number of patients without bladder cancer. Figure 1.3a shows the summary ROC curve that can be drawn through these values. Figure 1.3b shows the summary point estimate (black spot) and its 95% confidence region around it. Based on a re-analysis of the data from Glas et al.[10].

Study	TP	FP	FN	TN	Cutoff	Sensitivity	Specificity
Abbate 1998	59	4	50	69	12.0	0.54 [0.44, 0.64]	0.95 [0.87, 0.98]
Casella 2000	67	17	63	88	10.0	0.52 [0.43, 0.60]	0.84 [0.75, 0.90]
Chahal 2001	7	7	9	73	10.0	0.44 [0.20, 0.70]	0.91 [0.83, 0.96]
Giannopoulos 2001	47	16	21	34	8.0	0.69 [0.57, 0.80]	0.68 [0.53, 0.80]
Lahme 2001	25	31	15	98	10.0	0.63 [0.46, 0.77]	0.76 [0.68, 0.83]
Landman 1998	38	7	9	23	7.0	0.81 [0.67, 0.91]	0.77 [0.58, 0.90]
Lee 2001	53	10	17	26	7.7	0.76 [0.64, 0.85]	0.72 [0.55, 0.86]
Miyanga 1999	20	68	2	219	12.0	0.91 [0.71, 0.99]	0.76 [0.71, 0.81]
Oge 2001	20	4	7	6	10.0	0.74 [0.54, 0.89]	0.60 [0.26, 0.88]
Paoluzzi 1999	27	22	5	36	10.0	0.84 [0.67, 0.95]	0.62 [0.48, 0.74]
Ramakumar 1999	30	56	27	83	3.6	0.53 [0.39, 0.66]	0.60 [0.51, 0.68]
Sharma 1999	4	33	2	166	10.0	0.67 [0.22, 0.96]	0.83 [0.78, 0.88]
Sozen 1999	29	19	11	81	10.0	0.72 [0.56, 0.85]	0.81 [0.72, 0.88]
Zippe 1999	18	45	0	267	10.0	1.00 [0.85, 1.00]	0.86 [0.81, 0.89]

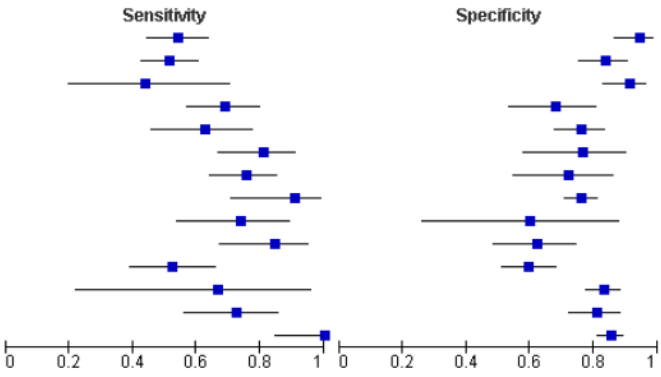

Figure 3.
Forest plots of sensitivity and specificity of a tumor marker for bladder cancer. Based on a re-analysis of the data from Glas et al.[10].

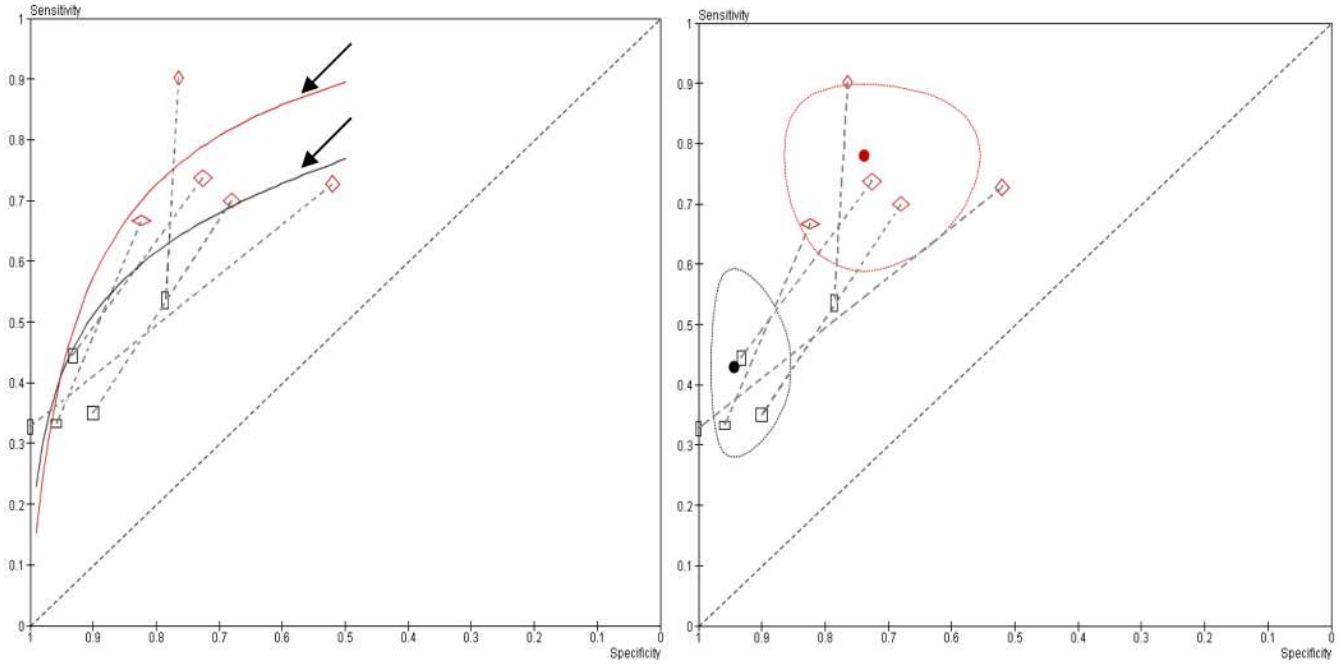

Figure 4.

Direct comparison of two index tests for bladder cancer: cytology (squares) and bladder tumor antigen (diamonds). Figure 1.4a shows the summary ROC curve that can be drawn through these values. Figure 1.4b shows the summary point estimate of sensitivity and specificity (black spot) and its 95% confidence region around it. The two tests clearly show a trade-off between sensitivity and specificity: cytology has a significantly higher specificity (ellipse closest to Y-axis lower arrow points at ROC curve) and BTA has a significantly higher sensitivity (higher ellipse and arrow points at highest ROC curve). It will depend on the role of the test in practice which test is considered 'best'. Based on a re-analysis of the data from Glas et al.[10].

Table 1

Bias in diagnostic test accuracy studies

Element	Type of Bias	When does it occur?	Under- or overestimation of diagnostic accuracy?
Patients	**Spectrum bias**	When patient inclusion does not represent the intended segment of target condition severity spectrum.	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	**Verification 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.	Variable.
	Incorporation bias	When the index test is incorporated in a (composite) reference standard.	Usually leads to overestimation.
	Time lag bias	When the target 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	**Result elimination bias**	When uninterpretable or intermediate test results and withdrawals are not included in the analysis.	Usually leads to overestimation.

Table 2

Essential elements in a systematic review of diagnostic test accuracy

	Phase in review process	Key issues
	Goal	To describe and, if possible, summarize diagnostic tests accuracy. Not meant to summarize other test features or downstream consequences of testing.
1.	**Objectives**	Include target condition, reference standard, intended patient group, the tests under evaluation, and their intended role in practice.
2.	**Study identification and selection**	Do not use restrictive electronic search filters Use index test and target condition as key terms.
3.	**Quality assessment**	Use the QUADAS checklist, adapted if necessary. Do not use summary quality scores.
4.	**Data-analysis and presentation**	Use plots in ROC space to present data graphically. Use the hierarchical summary ROC or the bivariate model for the analysis of the data for the analysis. Statistical tests for heterogeneity are not generally useful, nor are tests for Investigating publication bias.
5.	**Interpretation**	Report more than just summary estimates of accuracy. Include risk of bias and lack of applicability.

QUADAS: Quality Assessment of Diagnostic Accuracy Studies

ROC: Receiver operating characteristic