Systematic Reviews of Diagnostic Test Accuracy
– New Developments within The Cochrane Collaboration

Mariska M.G. Leeflang, Jon J. Deeks,
Constantine Gatsonis, Patrick M.M. Bossuyt

Submitted for publication
Chapter 1

Abstract

During the last two decades, the number of systematic reviews and meta-analyses of diagnostic test accuracy has grown considerably and substantial progress has been made in developing and agreeing on methodological standards.

The Cochrane Collaboration now considers it timely to register systematic reviews of diagnostic test accuracy studies, with the first Cochrane Diagnostic Test Accuracy Reviews scheduled to be published in the Cochrane Library in October 2008. Adding such reviews to the Cochrane Library may increase its appeal as the best single source of reliable evidence about the effects of health care.

Systematic review of diagnostic test accuracy studies can be methodologically challenging. Diagnostic accuracy studies can be difficult to identify. They are likely to show substantial variability, because of small sample sizes, clinical diversity, due to differences in setting or spectrum, and because of differences in design. Unlike randomized trials, which report a single measure such as the relative risk, diagnostic test accuracy studies usually report a pair of measures of test performance, such as the test's sensitivity and specificity, either at a point or along a ROC curve. Methods for meta-analysis have to take this bivariate nature of the data into account.

In this paper we present some of the most recent developments in the methodology for conducting systematic reviews and meta-analyses of diagnostic test accuracy studies that will be incorporated in the Cochrane review process.
1.1 Introduction

Diagnostic tests are a critical component of health care. 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 the results of testing.

Systematic reviews can help practitioners and decision-makers in answering these questions, by summarizing the available evidence and helping to explain differences among studies on the same question. The number of systematic reviews and meta-analyses of diagnostic test accuracy has grown remarkably in recent years. A search in MEDLINE (see Appendix 1.1) identified 77 published diagnostic reviews in 1996, a number that increased to 591 in 2006.

The Cochrane Collaboration is the largest international organization preparing, maintaining and promoting systematic reviews to help people make well-informed decisions about health care. In 2008 the 1st Issue of the Cochrane Database of Systematic Reviews (CDSR) included 3,384 reviews. Up until now, CDSR has been restricted to reviews of interventions, but the growing interest and the methodological advances in the synthesis of studies of diagnostic tests has lead to a change of policy, and from October 2008 CDSR will also include systematic reviews of diagnostic test accuracy.

The Cochrane Diagnostic Test Accuracy Working Group was launched in 2003 to systematize the approach and develop the software for these new systematic reviews. A meeting of more than 40 methodologists and expert reviewers from around the world was held in 2004 which reached consensus on appropriate methods and a reporting structure for protocols and reviews. Smaller working groups were subsequently formed to address each of the stages involved in the systematic review process. In the following years, these smaller groups reviewed methods and developed detailed guidance for review authors and review groups, which will be made available in the Cochrane Handbook for Diagnostic Test Accuracy Reviews. The methods in the Handbook are based on empirical evidence where available, making it a valuable resource for all authors of systematic reviews and meta-analyses of diagnostic test accuracy, including those preparing such reviews outside the scope of The Cochrane Collaboration.

In 1994, Irwig and colleagues presented guidelines for meta-analyses evaluating diagnostic tests in this journal. We review the key methodological developments concerning problem formulation, location of literature, quality assessment and analysis that have occurred since then, using our experience from the work on the Handbook.
1.1.1 Diagnostic Test Accuracy Reviews
A study of the diagnostic accuracy of a test is undertaken to estimate the ability of that test to distinguish between patients with disease (or more generally, a specified target condition) and those without. In such a study, the results of the test under evaluation, or ‘index test’, are compared with those of the clinical reference standard determined in the same patients. The clinical reference standard is the best available method for classifying patients as having the target condition or not. 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. The diagnostic accuracy of several tests may be evaluated in parallel in a single study.

Accuracy measures estimate the ability of a test to distinguish between persons with and without the target condition. Transformed to likelihood ratios, they can also be used to convert estimates of pre-test probabilities of disease to post-test probabilities, using Bayes’ theorem. When a new test is supposed to replace an existing one, one has to find out how the accuracy of that test compares to the existing one4-6. More generally, accuracy can help clinicians to make decisions about tests and their future role7. Good accuracy is a desirable but not a sufficient condition for the effectiveness of that test. To show that using a new test does more good than harm in terms of patient outcomes, one may require randomized trials of test-and-treatment strategies.

As elsewhere in science, systematic reviews and meta-analyses can be used to obtain more precise estimates, when small studies addressing the same test and patients in the same setting are summarized. Systematic reviews can also be useful to establish whether and how scientific findings vary significantly by particular subsets, providing summary or subgroup estimates of diagnostic test accuracy that may be more applicable than estimates from a single study. They may help in identifying studies with the lowest risk of bias and they can be used to explore the between study heterogeneity in results. Such heterogeneity is to be expected, and probably even more so with diagnostic accuracy studies. Some of the variability is due to chance, as many diagnostic studies have small sample sizes8. Some will be due to differences in study methods, but study populations are also likely to differ between studies, resulting in differences in accuracy estimates9. Systematic reviews may also be used to address questions that were not directly considered in the primary studies, such as comparisons between tests.

In what follows, we briefly discuss the steps for conducting a systematic review of test accuracy studies (see Table 1.1). The account is our summary of the methods profiled in the Handbook for Cochrane diagnostic test accuracy reviews.
1.2 Definition of the objectives of the review

Any diagnostic research question should start with a precise description of the test or tests of interest, the disease or condition which they have to help identify, and a definition of the clinical context in which they will be used. From these statements inclusion criteria can be developed that define the studies of relevance to include in the review. A typical question is whether the test of interest has sufficient accuracy, in a well defined patient population, setting and testing strategy, to fulfil a particular role. Many such questions will be comparative, contrasting the accuracy of two or more tests or testing strategies.

The role of the test under evaluation relative to the current best practice needs to be specified, including its relative position to other tests used for the same target condition. Possible questions for a new test are: (1) can this test replace another test; (2) can it serve as a triage instrument, guiding further testing, and (3) can the test be used in addition to current best practice to pick up additional cases of the target disease, or to identify and eliminate false positives? 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 the case of triage, one will be looking for a test that gives a minimal proportion of false negatives, so that the test can rule out disease in patients who will need no further testing. If the new test is to be used in addition to existing strategies, its aim will mainly be to reduce the number of false negatives, or, alternatively, the number of false positives. The review should provide data to assess the incremental change in accuracy made by adding the new test.

Test accuracy is not a fixed property of a test. It varies with the group of patients tested, with their spectrum of disease, with the clinical setting, with the test interpreters, and depends on the level of prior testing. For this reason, it is essential to include these elements in the study question.

1.2.1 Framing the research question

In a systematic review of the diagnostic accuracy of urinary markers for bladder cancer, the following issues were considered while defining the research question and objectives of the review. In clinical practice, cytology was used to triage pa-
patients before they underwent invasive cystoscopy. As cytology combines a high specificity with a low sensitivity, the goal of the review was to identify a tumour 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 combine an equally high specificity with a sufficiently high sensitivity, around 100%. From these objectives followed the inclusion criteria for the review. To include a study, markers and cytology had to be evaluated against the same reference standard, cystoscopy or histopathology; data to calculate sensitivity and specificity had to be available. Bladder tumours secondary to a cancer already identified and other target conditions were not allowed, as the diagnostic accuracy obtained in these cases cannot be translated directly to primary bladder tumours.

1.3 Identification of studies

Searching for and identifying test accuracy studies is now known to be more difficult than searching for randomized trials. There is not a clear, unequivocal key word or indexing term for an accuracy study, comparable to the term “randomized controlled trial”. The term “sensitivity and specificity” may look suitable, but is inconsistently applied in most databases. Data on diagnostic test accuracy may also 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. So until indexing systems are changed to properly code studies of test accuracy, searching for them will remain challenging.

In the development of a comprehensive search strategy, search strings that refer to the (1) test(s) under evaluation, (2) the target condition, (3) the patient description, or a subset of these can be used. For tests with a clear name that are used for a single purpose, just 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 poorly indexed. A search strategy in MEDLINE should contain both MeSH headings and text words. If one searches for articles about tests for bladder cancer, for example, it will be necessary to 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, which attempt to restrict a topic search to articles most likely to be test accuracy studies. These filters rely on indexing terms for research methodology and text words used in reporting results. However, they often miss relevant studies and are unlikely to decrease the number of articles one needs to screen, so are not recommended for use in systematic reviews. The incremental
Systematic reviews of diagnostic test accuracy

effects of searching in languages other than English and in the so called grey literature have not yet been fully investigated.

In systematic reviews of intervention studies, publication bias is an important and well-studied form of bias. 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. A consistent observation has been that studies with statistically significant results are more likely to be published than studies with non-significant findings. 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 possible. Tests used in reviews of randomized controlled trials to detect publication bias have proven to be seriously misleading for diagnostic studies, and alternatives have poor power. The determinants for publication of diagnostic studies may not be the same as the determinants for publication of intervention studies.

1.4 Assessment of methodological quality

Test accuracy studies with design deficiencies can produce biased results. Sources of bias in test accuracy studies for which there is unambiguous evidence that diagnostic accuracy can be overestimated are the inclusion of healthy controls and the incomplete or differential use of reference standards. 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.

Based on the available evidence, Whiting and colleagues have used a Delphi procedure to develop the QUADAS checklist, now the recommended tool for quality assessment in diagnostic test accuracy studies. The items that are listed in QUADAS relate to patient spectrum issues, verification issues, information bias and incomplete reporting. Issues related to the availability and quality of the reference standard include the overall appropriateness of this standard, partial or differential verification, important time gap between index test and reference standard, and the inclusion of the index test results in the reference standard in case of a composite reference standard.

The magnitude and direction of the resulting bias caused by methodological shortcomings may vary, depending on target condition and clinical setting. In addition, other items can be a cause of bias for specific tests. For example, in studies assessing the accuracy of biochemical serum markers, data-driven selection of the cut-off value may bias diagnostic accuracy. Review authors should therefore think carefully whether items need to be added to the QUADAS list.
Unfortunately, any evaluation of study quality is hampered by incomplete reporting\textsuperscript{25}. Guidelines for complete and transparent reporting have been developed\textsuperscript{26}, but their effects are only gradually becoming visible in the literature\textsuperscript{27}.

The results of quality appraisal can and should be summarized, to offer a general impression of the validity of the available evidence. Using an overall quality score is not recommended, 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\textsuperscript{28}. A way to summarize the quality assessment is shown in Figure 1.1, where stacked bars are used for each QUADAS item.

In the analysis phase, the results of the quality appraisal can be used to guide explorations of the sources of heterogeneity\textsuperscript{30,31}. Possible methods to address quality differences are sensitivity analysis, subgroup analysis or regression analysis, although the number of included studies may often be too low for meaningful meta-regression. The interpretation of results should at least be made bearing in mind the risk of bias.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{QuADAS.png}
\caption{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\textsuperscript{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.}
\end{figure}
1.5 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 or a risk difference or ratio, the results of most diagnostic test accuracy studies are reported with two or more statistics, the sensitivity and the specificity, the positive and negative predictive value, or likelihood ratios for the respective test results, or the ROC curve and quantities based on it\textsuperscript{32,33}.

The first step in the meta-analysis of diagnostic test accuracy is to visually examine the results of the individual studies. The paired results for sensitivity and specificity in the included studies can be plotted in a paired forest plot (see Figure 1.2) or plotted as points in an ROC plot (see Figure 1.3).

Plots of estimated sensitivity and specificity often display a pattern of negative correlation with each other across studies of the same test. A major contributor to this appearance is the trade-off between the true sensitivity and specificity of a test, as the threshold for defining test positivity varies. Decreasing the threshold that defines a test as positive rather than negative will increase sensitivity and decrease specificity (or vice versa), as described by the ROC curve for that test. When studies included in a review vary in positivity thresholds, a ROC-curve like pattern may be discerned across the points on the summary 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.

\begin{table}[h]
\begin{tabular}{|l|cccccc|}
\hline
\textbf{Study} & \textbf{TP} & \textbf{FP} & \textbf{FN} & \textbf{TN} & \textbf{Sensitivity} & \textbf{Specificity} \\
\hline
Abbate 1998 & 59 & 4 & 50 & 69 & 0.54 [0.44, 0.64] & 0.95 [0.87, 0.98] \\
Casella 2000 & 67 & 17 & 63 & 88 & 0.52 [0.43, 0.60] & 0.84 [0.75, 0.90] \\
Chahal 2001 & 7 & 7 & 9 & 73 & 0.44 [0.20, 0.70] & 0.91 [0.83, 0.96] \\
Giannopoulos 2001 & 47 & 16 & 21 & 34 & 0.69 [0.57, 0.80] & 0.68 [0.53, 0.80] \\
Lahme 2001 & 25 & 31 & 15 & 98 & 0.63 [0.46, 0.77] & 0.76 [0.68, 0.83] \\
Landman 1998 & 38 & 7 & 9 & 23 & 0.81 [0.67, 0.91] & 0.77 [0.59, 0.90] \\
Lee 2001 & 53 & 10 & 17 & 26 & 0.76 [0.64, 0.85] & 0.72 [0.55, 0.86] \\
Miyanga 1999 & 20 & 68 & 2 & 219 & 0.91 [0.71, 0.99] & 0.70 [0.71, 0.81] \\
Öge 2001 & 20 & 4 & 7 & 6 & 0.74 [0.54, 0.89] & 0.60 [0.26, 0.88] \\
Paoluzi 1999 & 27 & 22 & 5 & 36 & 0.84 [0.67, 0.96] & 0.62 [0.48, 0.74] \\
Ramakumar 1999 & 30 & 56 & 27 & 83 & 0.53 [0.39, 0.66] & 0.60 [0.51, 0.68] \\
Sharma 1999 & 4 & 33 & 2 & 166 & 0.67 [0.22, 0.96] & 0.83 [0.78, 0.88] \\
Sozen 1999 & 29 & 19 & 11 & 81 & 0.72 [0.56, 0.85] & 0.81 [0.72, 0.88] \\
Zippe 1999 & 18 & 45 & 0 & 267 & 1.00 [0.81, 1.00] & 0.86 [0.81, 0.89] \\
\hline
\end{tabular}
\end{table}

\textbf{Figure 1.2. 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.\textsuperscript{10}.}
Figure 1.3a 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.
Because threshold effects cause sensitivity and specificity estimates to appear as negatively correlated according to a ROC curve shape, and because threshold variation can be assumed to be present in nearly all situations to some degree, robust approaches to meta-analysis estimate the underlying relationship between sensitivity and specificity by constructing a summary ROC (SROC) curve. An average ‘operating point’ on this curve may subsequently be identified that indicates where the centre of the study results lie. Separate pooling of sensitivity and specificity to identify this point has been discredited, because in such an approach may identify a point which does not lie on the SROC curve when there is between study variation.

In 1994, Irwig and colleagues\(^3\) recommended Moses and Littenberg’s linear regression model for the construction of summary ROC curves\(^34\), which is based on regressing the log diagnostic odds ratio against a measure of the proportion reported as test positive. Extending the regression model by adding covariates has been proposed to examine differences between tests and relate them to study or sample characteristics. However, the formulation of the model has limitations in failing to consider the precision of the study estimates, not estimating between study heterogeneity and the explanatory variable in the regression being measured with error. These problems render estimates of confidence intervals and \(P\)-values that are unsuitable for formal inference\(^33,35\).

Two approaches to fitting random effects hierarchical models have been developed to overcome these limitations: the hierarchical summary ROC (HSROC) model\(^33,36,37\) and the bivariate random effects model\(^35,38\). The HSROC model focuses on identifying the underlying ROC curve, estimating the average accuracy 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 sensitivity and specificity and the correlation between them. These two basic models have been shown to be mathematically equivalent. Both can be used to identify the underlying SROC curve and the average operating point\(^35,39\). They can also be used to explore heterogeneity by adding covariates to the models, or by applying separate models to different subgroups. Both models can be fitted with statistical software for fitting mixed models\(^33,35,37,38\).

Some authors have advocated the pooling of likelihood ratios rather than pooling of sensitivity and specificity or pooling of ROC curves\(^40,41\). However, summary likelihood ratios can be easily calculated with the methods described above, while calculating sensitivity and specificity from pooled likelihood ratios may result in sensitivities and specificities above 1 or below 0\(^42\).
Figure 1.4. 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.
Systematic reviews of diagnostic test accuracy

1.5.1 Curves or summary points
The ability to estimate both underlying SROC curves and average operating points allows flexibility in testing hypotheses and estimating diagnostic accuracy. Analyses to estimate underlying SROC curves can be based on all included studies, and facilitate well powered comparisons between different tests, or between subgroups of studies, which are not restricted to investigating accuracy at a particular threshold. An example can be found in Figure 1.3a, where the diagnostic accuracy of a bladder tumour antigen test for diagnosing bladder cancer is summarized with an SROC curve. In contrast, estimation of a summary point specific to a test being used at a common threshold is useful to obtain the best estimate of test accuracy in parameters clinicians understand. The certainty associated with the estimate can be described by confidence regions marked on the SROC plot around the average point. An example of this approach is given in Figure 1.3b.

1.5.2 Comparative analyses
Systematic reviews of diagnostic test accuracy may evaluate more than one 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 of 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 the confounding.

An alternative approach is to only use studies that directly compared the tests in the same patients, or randomized patients to tests. Such direct comparisons do not suffer from confounding. Unfortunately, fully paired studies are not always available. 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 1.4.

1.6 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 is actually suitable for addressing the objectives of the review, and not consist solely of reporting the results. The interpretation of the findings should 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. A decision model could be used to structure the interpretation of the findings. Such a model would incorporate important factors as the disease prevalence and the available diagnostic and therapeutic interventions that may follow the test.
Chapter 1

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 SROC curves. The potential effects of quality differences, or the lack of high quality studies on the results should be considered. Additional information, such as costs or important trade-offs between harms and benefits can be included.

1.7 Conclusion

Important progress has been made in recent years in the methods for developing methodology for systematic reviews of diagnostic test accuracy studies. We know more about searching, about sources of bias in study design, and about quality appraisal. In meta-analysis new hierarchical random effects models have been developed with sound statistical properties. Methods for the estimation of summary ROC curves and of summary estimates of sensitivity and specificity are now available. All these advances are described in detail in the Cochrane Handbook for Diagnostic Test Accuracy Reviews.

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 a suitable spectrum 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\textsuperscript{41,42}. Furthermore, policy makers and guideline developers may be interested in the comparative accuracy only, or in additional information, such as costs and burden. Developing systematic reviews that are really relevant for both policy makers and clinical practice poses a major challenge, and clear thinking about the scope and purpose of the review is a necessary condition.

The Cochrane Diagnostic Test Accuracy Working Group addresses those challenges and will continue developing, evaluating and disseminating the methods for diagnostic test accuracy reviews.
Chapter 1

References


Chapter 1


Appendix 1.1. Search for diagnostic reviews

PubMED search strategy for identification of diagnostic test accuracy reviews in MEDLINE:

("Diagnosis"[Majr] OR diagnos*[ti] OR accuracy[ti]) AND (meta-analysis[tw] OR systematic review[tw]).