REVIEW 10.1111/1469-0691.12474

Systematic reviews and meta-analyses of diagnostic test accuracy

M. M. G. Leeflang

Clinical Epidemiology and Biostatistics and Bioinformatics, Academic Medical Centre, University of Amsterdam, Amsterdam, the Netherlands

Abstract

Systematic reviews of diagnostic test accuracy summarize the accuracy, e.g. the sensitivity and specificity, of diagnostic tests in a systematic and transparent way. The aim of such a review is to investigate whether a test is sufficiently specific or sensitive to fit its role in practice, to compare the accuracy of two or more diagnostic tests, or to investigate where existing variation in results comes from. The search strategy should be broad and preferably fully reported, to enable readers to assess the completeness of it. Included studies usually have a cross-sectional design in which the tests of interest, ideally both the index test and its comparator, are evaluated against the reference standard. They should be a reflection of the situation that the review question refers to. The quality of included studies is assessed with the Quality Assessment of Diagnostic Accuracy Studies-2 checklist, containing items such as a consecutive and all-inclusive patient selection process, blinding of index test and reference standard assessment, a valid reference standard, and complete verification of all included participants. Studies recruiting cases separately from (healthy) controls are regarded as bearing a high risk of bias. For meta-analysis, the bivariate model or the hierarchical summary receiver operating characteristic model is used. These models take into account potential threshold effects and the correlation between sensitivity and specificity. They also allow addition of covariates for investigatation of potential sources of heterogeneity. Finally, the results from the meta-analyses should be explained and interpreted for the reader, to be well understood.

Keywords: diagnosis, diagnostic test accuracy, evidence-based medicine, meta-analyses, sensitivity and specificity, systematic reviews
Article published online: 26 November 2013
Clin Microbiol Infect 2014; **20:** 105–113

Corresponding author: M. M. G. Leeflang, Clinical Epidemiology and Biostatistics and Bioinformatics, Academic Medical Centre, University of Amsterdam, Room J1b-209, PO Box 22700, 1100 DE Amsterdam, the Netherlands
E-mail: m.m.leeflang@amc.uva.nl

Introduction

Practising evidence-based medicine starts with a clinical question [1]. For example, a general physician might want to know whether testing for papilloma virus can replace cytology for the diagnosis of cervical cancer, as it is cheaper and easier to perform, or a haematologist might wonder whether a molecular test is needed on top of clinical judgement before a patient is treated for invasive fungal diseases. For questions such as these, the sensitivity and specificity of a diagnostic test may be helpful.

Systematic reviews are at the heart of evidence-based medicine. These literature overviews are performed in a systematic and transparent way, and they are explicit about where their study base comes from and how included references were selected. The quality of included studies is assessed and, if appropriate, the results are quantitatively summarized in a meta-analysis. These explicit methods limit bias, and improve the reliability of conclusions [2]. Systematic reviews also enable us to establish whether findings are consistent and can be generalized over different situations.

Healthcare professionals looking for evidence about diagnostic tests may turn to systematic reviews of diagnostic test accuracy. These reviews summarize the sensitivity and specificity of a test, and sometimes other measures as well, such as predictive values, likelihood ratios, ORs, or summary receiver operating characteristic (ROC) curves [3]. Sensitivity is defined as the probability of a person with the disease of interest having a positive test result, and specificity is defined as the probability of a person without the disease having a negative test result. These refer to the clinical situation in

which a test is being used, and are different from analytical sensitivity (referring to the ability of the test to measure low concentrations of a substance) and analytical specificity (usually referring to cross-reactivity). They may also be different from more technical definitions of sensitivity and specificity, such as the ability to distinguish between cases and (often healthy) controls. These analytical and technical measures are important in the earlier phases of test development, whereas clinical sensitivity and specificity are used to indicate the performance of a test in clinical practice [4].

The number of diagnostic test accuracy reviews has rapidly increased, especially over the last 5 years. A quick MEDLINE search revealed that the number of systematic reviews or meta-analyses with diagnosis, diagnostic, test, testing, tests or accuracy in the title grew from 748 at the end of 2008 to 2068 in November 2013. However, readers find it difficult to grasp the concept of a diagnostic test accuracy review, and this may limit their use in practice [5].

This overview describes the steps involved in a diagnostic test accuracy systematic review, while focusing on the link with the clinical question. We hope to explain for readers what they can expect from a diagnostic accuracy review, and how the results of these reviews can be used in clinical and laboratory practice. A selection of 20 diagnostic test accuracy reviews in infectious diseases will serve as an illustration [6–25] (see Appendix). These reviews come from a set of reviews published between September 2011 and January 2012 that we used to survey which meta-analytic methods authors use [26].

Review question

The first and most important step in a systematic review is question formulation. The review question guides the rest of the review: it dictates the relevant study design and study characteristics, the potential biases to be expected, the appropriate meta-analysis technique, and the interpretation of results. The review question includes some basic elements: the patients or population who will undergo the test in practice, the test(s) of interest and comparator test(s), and the target condition or disease of interest, as defined by the reference standard. When papilloma virus testing is compared with cytology, the patient population consists of women who will be tested for cervical cancer. The test of interest is called the index test, here being virus testing. Its comparator in this case is cytology. The disease of interest is cervical cancer; the term target condition refers to a more specific definition, e.g. a specific stage of cancer, or treatable cancer [27].

A key element in diagnostic accuracy is the reference standard. This is the test used to define the target condition, and the underlying assumption is that it reflects the truth. For cervical cancer, a valid reference standard is histopathology. By design, the reference standard is assumed to be flawless. The reference standard sets the reference, and sensitivity and specificity are expressed as the proportion of reference standard positives with a positive index test result, and the proportion of reference standard negatives with a negative index test result, respectively. It is therefore impossible to show that an index test is better than the reference standard, even if this would be the case in reality.

To place the review question in a context and to enable better interpretation of the results, the place of the test(s) in the diagnostic pathway should be described [28]. It matters whether the test is used as a first-line test to decide who should be referred for further testing, or whether the test will be used to start treatment on the basis of the test result. If a test is used as a first-line test, then the composition of the sample and the consequences of a false-positive or false-negative test result will be different from those in a more specialized situation. A first-line test, also called a triage test, may be useful even when the sensitivity or specificity is not high, depending on the steps that will be taken after testing. If the test is used to determine who should be treated and who should not be treated, it will be important to not miss any diseased patients (requiring high sensitivity), and it may be also be important to prevent the treatment of non-diseased persons (requiring high specificity), especially when the treatment is invasive or burdensome. A systematic review on molecular assays for neonatal sepsis aimed to investigate whether the sensitivity of these assays would be higher than 98% and the specificity higher than 95%, based on the balance between missing almost no neonate with sepsis and overtreatment of neonates without sepsis [19]. Authors may find it difficult to firmly state a minimally accepted sensitivity and specificity beforehand. Alternatively, one could hypothesize that the sensitivity and specificity in the current study should at least be as high as previously reported, or that the sensitivity and specificity of the index test should at least be as high as those of the comparator test(s).

An important secondary objective of a diagnostic test accuracy review is to investigate potential sources of heterogeneity. How do the sensitivity and specificity of a test differ between adults and children, or between primary care and secondary care, or between different subtypes of the test? For example, the objective of a systematic review on antigen tests for tuberculosis was to estimate the diagnostic accuracy of antigen detection tests using different clinical specimens in adults and children with and without human immunodeficiency virus infection [9].

Sometimes, a review states that it aims to 'characterize the clinical usefulness' [13] or to assess the 'immunodiagnostic efficacy' [24]. These terms refer to outcome measures other than accuracy measures. A test with very high sensitivity and specificity is not necessarily useful or effective: if patient management does not change after a test result, then testing does not influence the patient's outcome. Therefore, phrasing the aim and objective in terms of usefulness or efficacy may confuse the readers.

Searching for Literature

For systematic reviews, the aim of a search strategy is often to find all available evidence that can be used to answer a particular question [29]. Therefore, the search strategy should be as broad as possible. Although missing a random number of studies does not necessarily influence the summary estimates, the credibility of a systematic review may depend on the search strategy used. This search strategy should incorporate an electronic search, checking reference lists of relevant studies and reviews, and some effort should be put in the retrieval of unpublished data (grey literature).

For systematic reviews, at least two electronic bibliographic databases should be searched, such as MEDLINE, EMBASE (which includes MEDLINE), or BIOSIS. These databases can be accessed through a search engine, e.g. OVID (for MEDLINE or EMBASE) or PubMed (for MEDLINE). The development of a broad search strategy requires a wide variety of search terms, combined in a way that is not restrictive. In general, review authors are encouraged to use all existing synonyms for the target condition and all synonyms for the index test(s), and sometimes also for the class of tests that the index test belongs to. A broad search strategy includes these terms both as medical subject heading (if available) and as words in the title or abstract of a study. Medical subject headings are terms that are linked to a certain topic, and studies are indexed in the bibliography by use of these headings. However, indexing may not be perfect, and adding the same or similar terms as words in the title or abstract therefore increases the number of retrieved studies. Similarly, a strategy that focused only on words in titles or abstracts would miss studies not using these exact words, and medical subject headings may retrieve such a study.

Depending on the topic, the search may result in >5000 titles [10,19]. To limit these numbers, some authors use terms such as 'sensitivity and specificity' or 'accuracy', so-called methodological search filters [8,14]. Diagnostic accuracy studies are all described in different ways; there is no standard terminology. This is especially true for older studies, which makes it difficult to filter these studies or to index them. The use of search filters may therefore lead to relevant studies being missed, and so is not recommended [29,30]. Achieving a balance between manageable numbers and being as complete as possible is a complex task that requires support from information specialists.

Readers should be able to assess the likelihood that relevant studies were missed. This is only possible if the search strategy has been reported completely, including all terms used and the way in which these terms were used (as subject headings, or as words in titles and abstracts). Although word count limits may discourage authors to do so, most journals have online supplements or online appendices in which the complete search strategy may be reported.

Selection of Relevant Studies

The first step in the selection process is the selection of potentially relevant publications on the basis of title and abstract. Then, the full texts of these articles are read and included when deemed relevant. The last stage is the exclusion of studies that turn out to be not relevant when data are extracted. In every stage, the selection is performed by two individuals independently. Although the value of independent double selection over selection by one author has not been investigated, the complexity of diagnostic test accuracy studies suggests the need for selection by at least two review authors.

Two major diagnostic test accuracy designs may be distinguished. One is the so-called diagnostic case–control or two-gate design, in which the people with the disease (cases) are selected from a different population than the persons without the disease (controls) [31]. For example, people with malaria may be selected in a field health centre, whereas controls without malaria may be selected from among stored blood samples from donors without any infections. Although case–control designs provide an indication of the maximum accuracy of a test, and are therefore valuable in the technical validation of a test, estimates from these studies are generally not representative of a test's accuracy in clinical practice [32,33].

The alternative design is a more cohort-like approach, a typical cross-sectional design in which all patients suspected of having the disease of interest undergo the index test(s). To verify who has the disease and who does not, all included patients also undergo the reference standard test. The reference standard-positive patients can be seen as cases, and the reference standard-negative patients can be seen as controls. Such a design reflects reality better than the case–control design, and is more likely to provide valid estimates of diagnostic accuracy.

A review may include both case–control designs and cross-sectional designs. In that case, the potential for bias caused by the case–control studies should be assessed [18]. Another disadvantage of case–control design studies is that prevalence or predictive values cannot be estimated. The positive (or negative) predictive value is calculated by dividing the true-positive (or negative) results by all positive results (or negative results). Both prevalence and predictive values depend on the ratio of people with and without the disease. In case–control studies, this ratio is constructed artificially, and thus prevalence and predictive values calculated from such a study are artefacts.

If the aim of the review is to compare two or more tests, then, ideally, cross-sectional designs evaluating both tests against the same reference standard and in the same patients should be included. However, these studies are rare, and limiting the selection to only these comparative studies may result in no included studies at all. Non-comparative studies evaluate one of the tests of interest, and are far more common, but may also lead to a biased comparison [34]. If the studies evaluating test A against a reference standard were all performed in severely ill patients and all studies evaluating test B against a reference standard were performed in patients who were not that ill, then the difference between test A and test B may have been caused by the difference in setting rather than being a real difference between the tests.

Assessment of Methodological Quality

A number of studies have shown that diagnostic accuracy is not a fixed property of a test, although we may have been taught otherwise [35–37]. It varies, depending on where the study was performed, in which patients it was performed, how the test was performed, and whether the study was flawed. Risk of bias refers to a flawed study design and systematic errors in the conduct of the study. Applicability refers to clinical variation: studies performed in specialized clinics may not be applicable when the review is focused on primary-care questions. If the included studies are biased or not applicable to the situation in practice, the results from the review should be taken with caution. Most diagnostic accuracy reviews use the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool, or its revision, QUADAS-2 [38,39]. It assesses the quality of included studies in terms of risk of bias and concerns regarding applicability over four domains, as discussed below.

Patient selection may be biased when a case–control design is used, or when patients who may be difficult to diagnose are excluded from the study. Concerns regarding applicability arise when the patients in the study are not the same as those tested and used in practice. For example, when the study had no risk of bias, but included a combination of adults and children, and the review focuses on children, there may be a concern regarding applicability. The test may behave differently in adults and in children.

The *index test* may be biased when the assessment takes place while the assessors know the diagnosis or reference standard results of the patients, or when the threshold or cut-off value at which the results are reported was selected *ad hoc*. Concerns regarding applicability may also depend on the threshold used in the study. For example, if, in practice, most laboratories would use one threshold, but an included study uses a different threshold, then this may lead to concerns regarding applicability even if this threshold was defined beforehand.

The *reference standard* may be biased when the reference standard is assessed with knowledge of the index test results or when it is likely that the reference standard does not correctly classify the target condition. Concerns regarding applicability arise when the reference standard used in the study defines a different disease or target condition than the target condition of the review.

The fourth domain is *flow and timing*, and it refers to the time interval between the index test and the reference standard, and whether patients have been treated in the meantime. It also refers to the flow of testing: were all patients subjected to all index tests and the (same) reference standard? The last part of this domain concerns the analyses: for example, when studies exclude uninterpretable or intermediate results when calculating sensitivity and specificity, the accuracy may be overestimated.

Quality assessment results are presented in a graph or a table, and may be used to investigate the effect of bias on the results. Sometimes, reviews limit their analyses to high-quality studies only, but this often leads to there being too few studies in the review for any analyses to be performed [40]. Defining high quality is also problematic. Some review authors calculate an overall quality score and set a threshold to define high quality [21]. This is not recommended, as some sources of bias may be more influential than others, and this may differ between topics [41].

Data Analysis

The analysis of the data starts with a description of the included studies and their accuracy results: the number of studies retrieved, the number of diseased and non-diseased participants included, and the main characteristics (setting,

year, etc.). Accuracy results from the individual studies can be described in the same table, in a forest plot, or in a plot of sensitivity vs. 1 – specificity. Other accuracy measures may also be presented: predictive values, likelihood ratios, and ORs. Although any outcome can be meta-analysed, we will focus on sensitivity and specificity. The main reason for this is that any other measure can be calculated on the basis of these estimates, whereas it may not be possible to calculate valid measures of sensitivity and specificity the other way around [42].

A meta-analysis is a quantitative method that uses the results from the included studies to estimate a weighted average. Systematic reviews may contain a meta-analysis, but this is not a requirement. Sometimes, the data are too scarce or too heterogeneous. A key feature of diagnostic test accuracy and an important factor in the meta-analysis of these data is the threshold effect. Continuous tests require a threshold above (or below) which the test is considered to be positive. For example, all test results above a value of 150 units/mL are regarded as positive. If a higher value of the test correlates with more symptoms or a higher likelihood of the disease, then increasing the threshold will lead to more false-negative results (and thus lower sensitivity) and fewer false-positive results (and thus higher specificity). If the included studies use different thresholds, then the sensitivity and specificity will also be different between these studies. The threshold effect is the most obvious source of heterogeneity in an accuracy review.

Because sensitivity and specificity depend on the situation in which the test is being used, and because of the threshold effect, sensitivity and specificity are expected to be very heterogeneous. In combination with the correlation between sensitivity and specificity, this makes testing for heterogeneity or using the *I*-square statistic to indicate the degree of heterogeneity problematic. The review on cervical cancer reported an *I*-square of 85.6% for the specificity of cytology for cervical cancer [11]. Although this indicates that there is much heterogeneity in specificity, this heterogeneity may be largely caused by the variation in sensitivity or by a threshold effect. Rather than testing whether heterogeneity is present, authors are therefore encouraged to investigate where the heterogeneity comes from.

The dependence on threshold and the correlation between sensitivity and specificity also has implications for the meta-analysis. Methods are needed that can deal with heterogeneity, with threshold effects, and with the correlation between sensitivity and specificity. The Cochrane Collaboration currently recommends two random-effects methods: the bivariate model, and the hierarchical summary ROC (HSROC) model [43–45]. The bivariate model meta-analyses a summary estimate for sensitivity and specificity together (Fig. 1a), whereas the HSROC model models the parameters for the summary ROC curve (Fig. 1b). In general, the HSROC model is recommended for continuous tests when the included studies all report a different threshold for test positivity. The bivariate model is recommended for purely binary tests or when different studies report similar thresholds.

These models can also be used to investigate sources of heterogeneity. For example, a review on antigen detection tests for tuberculosis estimated the accuracy of these tests in both adults and children and in people with and without human immunodeficiency virus infection [9]. In this review, the authors separately analysed the data for all subgroups (subgroup analysis). Other ways to investigate heterogeneity

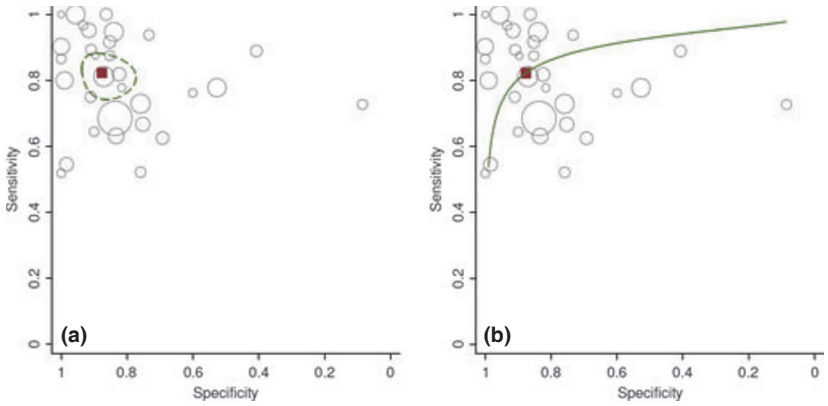

FIG. 1. Summary receiver operating characteristic (ROC) plots. (a) Summary sensitivity and specificity (red square) and their 95% confidence region (dotted ellipse). (b) Summary ROC curve (solid line) and summary point (red square). Every circle represents the sensitivity and specificity estimate from one study, and the size of the circle reflects the sample size. The graphs were made in StataIC 10, based on data from Onishi *et al.* [18].

are the removal of certain subgroups from the overall analyses (sensitivity analysis), or the inclusion of variables as covariates in the meta-analyses. Both the HSROC model and the bivariate model allow the addition of covariates. In the HSROC model, covariates may explain the variation in accuracy, threshold or shape of the ROC curve. In the bivariate model, covariates may explain variation in sensitivity and specificity. Of the 20 reviews in the Appendix, 11 used either the HSROC or the bivariate model.

A comparison between two tests can be performed by treating these tests as a source of heterogeneity, by meta-analysing the two tests separately (subgroup analysis), or by adding test type as a covariate to the bivariate or HSROC model. If both comparative studies and non-comparative studies are included in the analyses, it may be worthwhile performing a sensitivity analysis to assess the effect of the comparative studies on the overall analyses [34].

An important threat to the validity of a systematic review is publication bias. Publication bias occurs if studies with favourable results are more likely to be published than studies with unfavourable results. Although there is no evidence of publication bias in diagnostic accuracy reviews, it seems unlikely that it does not exist at all. Methods to detect publication bias are not very reliable when used in diagnostic accuracy data, although the method of Deeks *et al.* has been shown to be the least biased [46]. An alternative is to compare conference abstracts with published work [47].

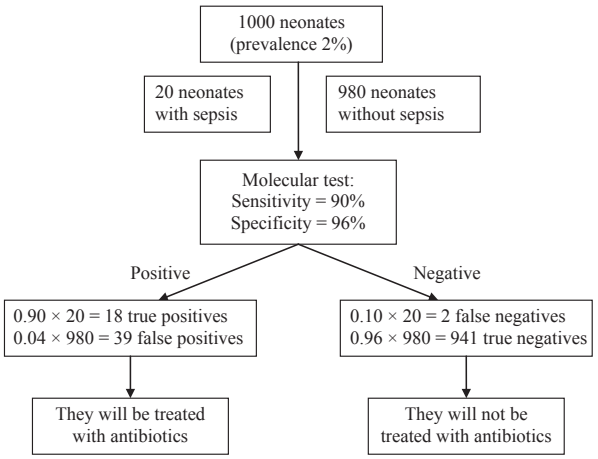

FIG. 2. Consequences of a molecular test for neonatal sepsis in a hypothetical cohort of 1000 neonates. The use of this test will, on average, mean that of 57 neonates treated with antibiotics, 39 do not have sepsis, and of 943 neonates who will not be treated with antibiotics, two should have been treated after all [19].

Interpretation and Conclusions

Interpretation of the results of the review and the concluding remarks should refer back to the review question and the (potential) role of the index test in clinical practice. Readers find it difficult to grasp the implications of the estimated sensitivity and specificity, and should be guided towards better understanding. Key to this explanation is the role of the test in practice and the potential consequences of a positive test result and a negative test result.

As a start, the main results could be presented in a summary of findings table, e.g. like the table presented by Kattenberg *et al.* [10]. In this table, the absolute numbers of true positives, false positives, false negatives and true negatives may be given for a hypothetical cohort of 1000 people. In the review on molecular tests for neonatal sepsis, the hypothetical cohort may exist of 1000 neonates screened for early-onset sepsis [19]. The expected prevalence of early-onset sepsis in this group is 2%. This means that 20 neonates will have sepsis. An assay with a sensitivity of 90% and a specificity of 96% would miss two of 20 cases with sepsis, and would lead to overtreatment in 39 of 980 neonates without sepsis (Fig. 2). These numbers may facilitate the interpretation of accuracy results, especially in the case of a comparative question, where the numbers for one test can be compared with the numbers for its alternative(s). Although high sensitivity and specificity do not necessarily lead to better health for the patient, putting the results in a clinical context and combining the clinical context with hypothetical numbers may provide more insights into the relevant consequences for the false-positive patients and the false-negative patients.

Results from the meta-analysis should also be interpreted in the light of their validity. If the majority of included studies are case–control designs, the actual accuracy will probably be lower than the estimates from the meta-analysis. This should be emphasized. The same is true for the other sources of bias and heterogeneity.

Transparency Declaration

M. Leeflang is co-convenor of the Cochrane Screening and Diagnostic Test Methods group, and is co-author of three of the reviews used for illustration in this article.

Appendix 1: Twenty diagnostic test accuracy reviews in infectious diseases

First Author and year	Topic (*test for target condition*)	Objective/aim	Reference standard	Quality assessment (tool used: results reported)	Meta-analysis (method used)	Publication bias (method used; results)
Chang 2011 [6]	Molecular pyrazinamide susceptibility testing for tuberculosis	To clarify the role of molecular assays by evaluating their test performance in different clinical settings	Phenotypic drug susceptibility of Mycobacterium tuberculosis to pyrazinamide	N/A	Bivariate model	Deeks' method: significant funnel plot asymmetry
Chen 2011 [7]	Interferon-gamma assays for tuberculosis	To support the development of evidence-based guidance on the use of interferon-gamma assays for active tuberculosis in HIV-infected patients	Culture or histopathology	QUADAS: all articles fulfilled at least 13 items.	Separate pooling of sensitivity and specificity	NR
Deshpande 2011 [8]	Real-time polymerase chain reaction for Clostridium difficile	To investigate whether real-time polymerase chain reaction used alone was sufficiently sensitive and specific for the diagnosis of Clostridium difficile in endemic populations	Two different culturing methods	QUADAS: nine out of 11 items assessed fulfilled by all studies.	Separate pooling and MosesLittenberg model	N/A; only mentioned in discussion
Flores 2011 [9]	Antigen detection tests for tuberculosis	To estimate the diagnostic accuracy of antigen detection tests using different clinical specimens for tuberculosis in adults and children with and without HIV infection	Culture, smear, or histopathological examination	QUADAS: all studies satisfied six quality items, studies lacked a representative spectrum and blinding.	HSROC model	N/A; only mentioned in discussion
Kattenberg 2011 [10]	Rapid diagnostic tests for placental malaria	To investigate diagnostic accuracy of rapid diagnostic tests for the diagnosis of malaria infection in pregnant women compared to a reference standard	Histology and microscopy	QUADAS: variable results, many items not reported.	Bivariate model	NR
Kocken 2012 [11]	Papilloma virus testing versus cytology for cervical disease	To summarize and update current knowledge of the value of cytology, papilloma virus testing and co-testing used in post-treatment surveillance	Histology	QUADAS: majority of items was fulfilled by all studies.	Bivariate model	N/A; only mentioned in discussion
Lu 2011a [12]	Polymerase chain reaction for Pneumocystis pneumoniae	To assess the diagnostic accuracy of polymerase chain reaction	Microscopy	N/A	Unclear	NR
Lu 2011b [13]	1,3-beta-D-glucan for fungal diseases	To characterize the clinical usefulness of the assay in at-risk patients and to evaluate which variables affect its performance	EORTC criteria	STARD and QUADAS: all studies had high STARD scores and most studies had high QUADAS scores	Bivariate model	Deeks' method: no funnel plot asymmetry found
Mathews 2011 [14]	Cytology for anal/cervical dysplasia	To meta-analytically compare a summary operating characteristic of the performance of cervical and anal cytology testing in the detection of cervical and anal cancer and their precursors	Colposcope magnified and directed punch biopsy of the uterine cervix or anal canal	QUADAS + study design: all studies fulfilled majority of items; blinding and uninterpretables often unreported.	Separate pooling	Funnel plots and Egger test used: no evidence of funnel plot asymmetry
Medina 2011 [15]	Clinical examination for urinary tract infection	To determine the probability that various symptoms, signs, antecedents and tests predict urinary tract infection in women	Culture	QUADAS: variation in results.	Separate pooling of Likelihood Ratios	N/A; only mentioned in discussion
Minion 2011 [16]	Urine lipoarabinomannan for tuberculosis	NR	Different ways of culturing and microscopy	QUADAS: variation in results.	Bivariate model	NR
Mugasa 2012 [17]	Molecular amplification tests for trypanosomiasis	NR	Microscopy	QUADAS + study design: variable results	Bivariate model	NR
Onishi 2012 [18]	1,3-beta-D-glucan for fungal diseases	To assess the diagnostic accuracy of this assay	EORTC criteria or autopsy	QUADAS: variable results; blinding poorly reported; timing poorly reported; case-control studies overestimated accuracy	Separate pooling and MosesLittenberg	Funnel plots and Egger test used: significant funnel plot asymmetry
Pammi 2011 [19]	Molecular assays for neonatal sepsis	To assess whether molecular assays have sufficient sensitivity (0.98) and specificity (0.95) to replace microbial cultures in the diagnosis of neonatal sepsis	Blood culture	QUADAS: most items were fulfilled	Bivariate model	N/A; only mentioned in discussion
Pant Pai 2012 [20]	Oral versus whole blood rapid point of care tests for HIV	To establish whether a convenient, non-invasive, HIV test that uses oral fluid was accurate by comparison with the same test with blood-based specimens	Immunological testing on whole blood	QUADAS: no further information provided.	Hierarchical Bayesian meta-analytic model	N/A; only mentioned in discussion
Summah 2011 [21]	Receptor expression test for bacterial infection	NR	Culture, clinical criteria and/or repsonse to therapy	STARD and QUADAS: majority of studies had high quality scores	Unclear	Funnel plots and Egger test used: significant funnel plot asymmetry
Sun 2011a [22]			Culture or clinical criteria	N/A		NR

(*Continued*)

First Author and year	Topic (*test for target condition*)	Objective/aim	Reference standard	Quality assessment (tool used: results reported)	Meta-analysis (method used)	Publication bias (method used; results)
	Interferon-gamma assays for pediatric tuberculosis	To compare the sensitivity and specificity of commercial interferon-gamma assays with the skin test in pediatric tuberculosis			Separate pooling of sensitivity and specificity	
Sun 2011b [23]	Polymerase chain reaction on bronchioalveolar lavage fluid for invasive aspergillosis	To assess the accuracy of polymerase chain reaction as a diagnostic test for invasive aspergillosis in immunocompromised patients	EORTC criteria	STARD and QUADAS: quality of all studies was generally high, meeting on average 10 of the 14 QUADAS criteria	Bivariate model	Deeks' method: no funnel plot asymmetry found
Wang 2012 [24]	Immunological tests for Schistosoma japonicum	To assess the immunodiagnostic efficacies of immunological tests for detection of Schistosoma japonicum human infections in the field	Microscopy	N/A	MosesLittenberg	Funnel plot asymmetry: method not specified.
Zhang 2011 [25]	Polymerase chain reaction versus serology for Mycoplasma pneumoniae	To conclude the overall accuracy of polymerase chain reaction	Serology	QUADAS: no further information provided, all studies were case-control design.	Bivariate model/ HSROC model	Deeks' method: no funnel plot asymmetry found

HIV, hunam immunodeficiency virus; N/A, not assessed; NR, not reported at all; EORTC, European Organisation for Research and Treatment of Cancer; QUADAS, Quality assessment of diagnostic accuracy studies; STARD, Standards for the reporting of diagnostic accuracy studies; HSROC, Hierarchical Summary Receiver Operating Characteristic.

Sensitivity: proportion of persons tested positive amongst those having the target condition, TP/(TP+FN).

Specificity: proportion of persons tested negative amongst those without the target condition, TN/(TN+FP).

Prevalence: proportion of persons with the target condition amongst the group suspected of having the condition, (TP+FN)/(TP+FP+FN+TN).

Positive predictive value: proportion having the target condition amongst those tested positive, TP/(TP+FP).

Negative predictive value: proportion not having the target condition amongst those tested negative, TN/(TN+FN).

Positive likelihood ratio: ratio of the proportion of positives amongst those with the target condition compared to the proportion of positives amongst those without the target condition, sensitivity/(1-specificity).

Negative likelihood ratio: ratio of the proportion negatives amongst those with the target condition compared to the proportion negatives amongst those without the target condition, (1-sensitivity)/specificity.

Diagnostic odds ratio: ratio of the odds of testing positive when having the target condition compared to the odds of testing positive without the target condition, (TP/FN):(FP/TN).

Receiver characteristic operating (ROC) curve: the sensitivity and specificity of a test vary depending on the threshold chosen. The ROC curve describes the trade-off between sensitivity and specificity as the threshold changes.

References

1. Straus SE, Richardson WS, Glasziou P, Haynes RB. Evidence-based medicine. In: Straus SE, Richardson WS, Glasziou P, Haynes RB, eds. *How to practice and teach EBM*, 3rd edn. Oxford: Elsevier, 2005; 13–30.
2. Mulrow CD. Rationale for systematic reviews. *BMJ* 1994; 309: 597–599.
3. Honest H, Khan KS. Reporting of measures of accuracy in systematic reviews of diagnostic literature. *BMC Health Serv Res* 2002; 2: 4.
4. Sackett DL, Haynes RB. The architecture of diagnostic research. *BMJ* 2002; 324: 539–541.
5. Zhelev Z, Garside R, Hyde C. A qualitative study into the difficulties experienced by healthcare decision makers when reading a Cochrane diagnostic test accuracy review. *Syst Rev* 2013; 2: 32.
6. Chang KC, Yew WW, Zhang Y. Pyrazinamide susceptibility testing in *Mycobacterium tuberculosis*: a systematic review with meta-analyses. *Antimicrob Agents Chemother* 2011; 55: 4499–5505.
7. Chen J, Zhang R, Wang J *et al.* Interferon-gamma release assays for the diagnosis of active tuberculosis in HIV-infected patients: a systematic review and meta-analysis. *PLoS ONE* 2011; 6: e26827.
8. Deshpande A, Pasupuleti V, Rolston DD *et al.* Diagnostic accuracy of real-time polymerase chain reaction in detection of *Clostridium difficile* in the stool samples of patients with suspected *Clostridium difficile* infection: a meta-analysis. *Clin Infect Dis* 2011; 53: e81–e90.
9. Flores LL, Steingart KR, Dendukuri N *et al.* Systematic review and meta-analysis of antigen detection tests for the diagnosis of tuberculosis. *Clin Vaccine Immunol* 2011; 18: 1616–1627.

10. Kattenberg JH, Ochodo EA, Boer KR, Schallig HD, Mens PF, Leeflang MM. Systematic review and meta-analysis: rapid diagnostic tests versus placental histology, microscopy and PCR for malaria in pregnant women. *Malar J* 2011; 10: 321.
11. Kocken M, Uijterwaal MH, de Vries AL *et al.* High-risk human papillomavirus testing versus cytology in predicting post-treatment disease in women treated for high-grade cervical disease: a systematic review and meta-analysis. *Gynecol Oncol* 2012; 125: 500–507.
12. Lu Y, Ling G, Qiang C *et al.* PCR diagnosis of *Pneumocystis* pneumonia: a bivariate meta-analysis. *J Clin Microbiol* 2011; 49: 4361–4363.
13. Lu Y, Chen YQ, Guo YL, Qin SM, Wu C, Wang K. Diagnosis of invasive fungal disease using serum (1→3)-β-D-glucan: a bivariate meta-analysis. *Intern Med* 2011; 50: 2783–2791.
14. Mathews WC, Agmas W, Cachay E. Comparative accuracy of anal and cervical cytology in screening for moderate to severe dysplasia by magnification guided punch biopsy: a meta-analysis. *PLoS ONE* 2011; 6: e24946.
15. Medina-Bombardó D, Jover-Palmer A. Does clinical examination aid in the diagnosis of urinary tract infections in women? A systematic review and meta-analysis. *BMC Fam Pract* 2011; 12: 111.
16. Minion J, Leung E, Talbot E, Dheda K, Pai M, Menzies D. Diagnosing tuberculosis with urine lipoarabinomannan: systematic review and meta-analysis. *Eur Respir J* 2011; 38: 1398–1405.
17. Mugasa CM, Adams ER, Boer KR *et al.* Diagnostic accuracy of molecular amplification tests for human African trypanosomiasis—systematic review. *PLoS Negl Trop Dis* 2012; 6: e1438.
18. Onishi A, Sugiyama D, Kogata Y *et al.* Diagnostic accuracy of serum 1,3-β-D-glucan for *Pneumocystis jiroveci* pneumonia, invasive candidiasis, and invasive aspergillosis: systematic review and meta-analysis. *J Clin Microbiol* 2012; 50: 7–15.
19. Pammi M, Flores A, Leeflang M, Versalovic J. Molecular assays in the diagnosis of neonatal sepsis: a systematic review and meta-analysis. *Pediatrics* 2011; 128: e973–e985.
20. Pant Pai N, Balram B, Shivkumar S *et al.* Head-to-head comparison of accuracy of a rapid point-of-care HIV test with oral versus whole-blood specimens: a systematic review and meta-analysis. *Lancet Infect Dis* 2012; 12: 373–380.
21. Summah H, Tao LL, Zhu YG, Jiang HN, Qu JM. Pleural fluid soluble triggering receptor expressed on myeloid cells-1 as a marker of bacterial infection: a meta-analysis. *BMC Infect Dis* 2011; 11: 280.
22. Sun L, Xiao J, Miao Q *et al.* Interferon gamma release assay in diagnosis of pediatric tuberculosis: a meta-analysis. *FEMS Immunol Med Microbiol* 2011; 63: 165–173.
23. Sun W, Wang K, Gao W *et al.* Evaluation of PCR on bronchoalveolar lavage fluid for diagnosis of invasive aspergillosis: a bivariate metaanalysis and systematic review. *PLoS ONE* 2011; 6: e28467.
24. Wang W, Li Y, Li H *et al.* Immunodiagnostic efficacy of detection of *Schistosoma japonicum* human infections in China: a meta analysis. *Asian Pac J Trop Med* 2012; 5: 15–23.
25. Zhang L, Zong ZY, Liu YB, Ye H, Lv XJ. PCR versus serology for diagnosing *Mycoplasma pneumoniae* infection: a systematic review & meta-analysis. *Indian J Med Res* 2011; 134: 270–280.
26. Ochodo EA, Reitsma JB, Bossuyt PM, Leeflang MM. Survey revealed a lack of clarity about recommended methods for meta-analysis of diagnostic accuracy data. *J Clin Epidemiol* 2013; 66: 1281–1288.
27. Lord SJ, Staub LP, Bossuyt PM, Irwig LM. Target practice: choosing target conditions for test accuracy studies that are relevant to clinical practice. *BMJ* 2011; 343: d4684.
28. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ* 2006; 332: 1089–1092.
29. de Vet HCW, Eisinga A, Riphagen II, Aertgeerts B, Pewsner D. Searching for studies. In: *Cochrane handbook for systematic reviews of diagnostic test accuracy Version 0.4* (updated September 2008). The Cochrane Collaboration, Oxford, 2008.
30. Beynon R, Leeflang MM, McDonald S *et al.* Search strategies to identify diagnostic accuracy studies in MEDLINE and EMBASE. *Cochrane Database Syst Rev* 2013; 9: MR000022.
31. Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. *Clin Chem* 2005; 51: 1335–1341.
32. Lijmer JG, Mol BW, Heisterkamp S *et al.* Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999; 282: 1061–1066.
33. Whiting PF, Rutjes AW, Westwood ME, Mallett S, QUADAS-2 Steering Group. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *J Clin Epidemiol* 2013; 66: 1093–1104.
34. Takwoingi Y, Leeflang MM, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. *Ann Intern Med* 2013; 158: 544–554.
35. Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Ann Intern Med* 2002; 137: 598–602.
36. Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Stat Med* 1997; 16: 981–991.
37. Moons KGM, Harrell FE. Sensitivity and specificity should be deemphasized in diagnostic accuracy studies. *Acad Radiol* 2003; 10: 670–672.
38. 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.
39. Whiting PF, Rutjes AW, Westwood ME *et al.* QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011; 155: 529–536.
40. Leeflang M, Reitsma J, Scholten R *et al.* Impact of adjustment for quality on results of metaanalyses of diagnostic accuracy. *Clin Chem* 2007; 53: 164–172.
41. 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.
42. Zwinderman AH, Bossuyt PM. We should not pool diagnostic likelihood ratios in systematic reviews. *Stat Med* 2008; 27: 687–697.
43. 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: 982–990.
44. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med* 2001; 20: 2865–2884.
45. Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. *Cochrane handbook for systematic reviews of diagnostic test accuracy Version 1.0*. The Cochrane Collaboration, 2010. Available from: http://srdta.cochrane.org/.
46. 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: 882–893.
47. Brazzelli M, Lewis SC, Deeks JJ, Sandercock PA. No evidence of bias in the process of publication of diagnostic accuracy studies in stroke submitted as abstracts. *J Clin Epidemiol* 2009; 62: 425–430.