METHODOLOGY PAPER

Diagnostic test accuracy: methods for systematic review and meta-analysis

Jared M. Campbell PhD,[1] Miloslav Klugar PhD,[2] Sandrine Ding PhD,[3] Dennis P. Carmody PhD,[4] Sasja J. Hakonsen MScN,[5] Yuri T. Jadotte PhD,[6,7] Sarahlouise White PhD[8] and Zachary Munn PhD[1]

[1]The Joanna Briggs Institute, Faculty of Health Sciences, University of Adelaide, Adelaide, Australia, [2]Department of Social Medicine and Public Health, Faculty of Medicine and Dentistry, The Czech Republic (Middle European) Centre for Evidence-Based Health Care: An Affiliated Centre of The Joanna Briggs Institute, Palacky University Olomouc, Olomouc, Czech Republic, [3]Bureau d'Echanges des Savoirs Pour des praTiques Exemplaires de Soins (BEST): An Affiliate Centre of The Joanna Briggs Institute, Lausanne, Switzerland, [4]Rutgers Robert Wood Johnson Medical School, Institute for the Study of Child Development, New Brunswick, New Jersey, USA, [5]Center for Kliniske Retningslinjer, Institut for Medicin and Sundhedsteknologi, Aalborg Universitet, Aalborg, Denmark, [6]Northeast Institute for Evidence Synthesis and Translation, A Collaborating Center of the Joanna Briggs Institute, Division of Nursing Science, Rutgers School of Nursing, [7]Department of Quantitative Methods, Biostatistics and Epidemiology, Rutgers School of Public Health, Newark, New Jersey, USA, and [8]Department of Speech Pathology and Audiology, Flinders University, Adelaide, Australia

ABSTRACT

Systematic reviews are carried out to provide an answer to a clinical question based on all available evidence (published and unpublished), to critically appraise the quality of studies, and account for and explain variations between the results of studies. The Joanna Briggs Institute specializes in providing methodological guidance for the conduct of systematic reviews and has developed methods and guidance for reviewers conducting systematic reviews of studies of diagnostic test accuracy. Diagnostic tests are used to identify the presence or absence of a condition for the purpose of developing an appropriate treatment plan. Owing to demands for improvements in speed, cost, ease of performance, patient safety, and accuracy, new diagnostic tests are continuously developed, and there are often several tests available for the diagnosis of a particular condition. In order to provide the evidence necessary for clinicians and other healthcare professionals to make informed decisions regarding the optimum test to use, primary studies need to be carried out on the accuracy of diagnostic tests and the results of these studies synthesized through systematic review. The Joanna Briggs Institute and its international collaboration have updated, revised, and developed new guidance for systematic reviews, including systematic reviews of diagnostic test accuracy. This methodological article summarizes that guidance and provides detailed advice on the effective conduct of systematic reviews of diagnostic test accuracy.

Key words: diagnosis, diagnostic test accuracy, meta-analysis, systematic review

Int J Evid Based Healthc 2015; 13:154–162.

Introduction

Diagnostic test accuracy

Diagnostic tests are used to identify the presence or absence of a condition for the purpose of developing an appropriate treatment plan.[1] Examples of diagnostic tests include imaging and biochemical technologies, pathological and psychological investigation, and signs and symptoms observed during history taking and clinical evaluations.[2] Owing to the demands for improvements in speed, cost, ease of performance, patient safety, and accuracy, new diagnostic tests are continuously developed, and there are often several tests available for the diagnosis of a particular condition.[1] In order to provide the evidence necessary for clinicians and other healthcare professionals to make informed decisions regarding the optimum test to use, primary studies need to be carried out on the accuracy of diagnostic tests and the results of these studies synthesized through systematic review.

This methods article aims to provide systematic reviewers with a detailed set of guidance for undertaking

Correspondence: Jared Michael Campbell, PhD, Faculty of Health Science, The Joanna Briggs Institute, University of Adelaide, Adelaide, Australia. E-mail: jared.campbell@adelaide.edu.au

DOI: 10.1097/XEB.0000000000000061

a systematic review and meta-analysis of diagnostic test accuracy. It has been developed as a partner to a chapter in the *Joanna Briggs Institute (JBI) Reviewer's Manual* that can be consulted for further guidance.[3]

Diagnostic test accuracy study designs

In primary studies of diagnostic test accuracy, the test of interest (the 'index test') is compared with an existing diagnostic test (the 'reference test'), which is known to be the best test currently available for accurately identifying the presence or absence of the condition of interest. The outcomes of the two tests are then compared in order to evaluate the accuracy of the index test. The two main types of studies of diagnostic test accuracy are case–control and cross-sectional. In case–control studies, also sometimes called the 'two gate design',[4] people with the condition (cases) come from one population (i.e., a healthcare center for people known to have the condition), whereas people without the condition come from another. Although this design gives an indication of the maximum accuracy of the test, it has been noted that the results will generally exaggerate the test's accuracy in practice.[5] In cross-sectional studies, all patients suspected of having the condition of interest undergo both the index test and the reference test. Those who test positive for the condition by the reference test can be considered the cases, whereas those who test negative are the controls. This study design is held to reflect actual practice better and is more likely to provide a valid estimate of diagnostic accuracy.[5] As such, cross-sectional studies are the preferred evidence resource for forming conclusions regarding diagnostic test accuracy.

Measures of diagnostic test accuracy

Diagnostic accuracy is predominantly represented by two measures, sensitivity and specificity; however, sometimes other measures, including predictive values, odds ratios, likelihood ratios, and receiver-operating characteristic (ROC) curves, are used.[4] Sensitivity refers to the probability of a person with the condition of interest having a positive result (also known as the true positive proportion), whereas specificity is the probability of a person without the condition of interest having a negative result (also known as the true negative proportion).[4] It should be noted that these definitions refer to the clinical situation, and other definitions of sensitivity and specificity exist that are used in different contexts.[6]

Specifically, sensitivity can be calculated as $\frac{\text{True positives}}{(\text{True positives}+\text{False negatives})}$, whereas specificity can be calculated as $\frac{\text{True negatives}}{(\text{True negatives}+\text{False positives})}$ (Table 1). A measure of test accuracy that brings together sensitivity and specificity is the diagnostic odds ratio, which is the ratio of the odds of disease in test positives relative to the odds of disease in test negatives. Sensitivity and specificity have been identified as essential measures of diagnostic accuracy.[4,5,7]

ROC curve analysis is also very common for evaluating the performance of diagnostic tests that classify individuals into categories of those with and those without a condition.[8,9] The data obtained from a diagnostic test will often exist on a scale (i.e., blood pressure, hormone concentration), and a decision will need to be made on whether a certain test value indicates that the condition is present (positive test) or not (negative test). Where this 'cutoff' is made is termed the decision or positivity threshold. Choice of a decision threshold will have a large effect on the sensitivity and specificity of a test. Although setting a low threshold will result in a large proportion of true positives being correctly identified as positive, it will also decrease the rate of true negatives. In other words, a lower threshold increases sensitivity but decreases specificity, whereas the inverse is true for high thresholds. As sensitivity and specificity depend on the selection of a decision threshold, ROC analysis is used to plot the sensitivity (*y*-axis) against 1-specificity (*x*-axis) as the threshold value changes,[10] giving a visual representation of the relationship between the sensitivity and specificity of a diagnostic test (Fig. 1). This can be quantitatively measured by assessing the area under the curve (AUC).[11] The AUC for a perfect test is 1.0, and a test with no differentiation between disorder and no disorder has an AUC of 0.5.[12]

Additional measures of diagnostic test accuracy include predicative values and likelihood ratios. However, as these are not essential concepts for undertaking

Table 1. Classification of patient test results by condition status

Index test outcome	Reference positive	Reference negative	Total
Index test positive	True positives (TPs)	False positives (FPs)	Test positives (TPs + FPs)
Index test negative	False negatives (FNs)	True negatives (TNs)	Test negatives (FNs + TNs)
Total	Reference positives (TPs + FNs)	Reference negatives (FPs + TNs)	*N* (TPs + FPs + FNs + TNs)

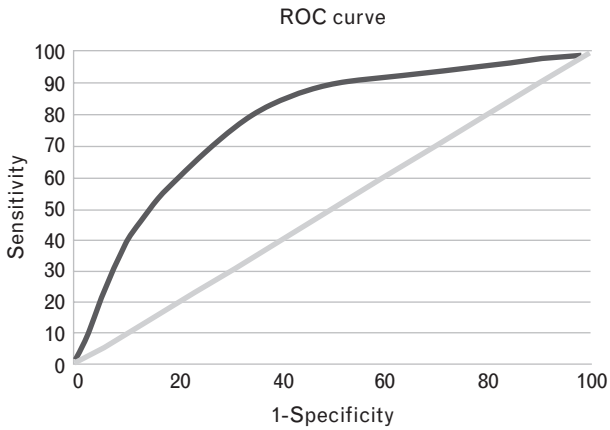

Figure 1. An example receiver-operating characteristic curve generated using mock data; the diagonal line shows the baseline result of a tests with no differential power (area under the curve = 0.5). ROC, receiver-operating characteristic.

a systematic review of diagnostic test, they will not be discussed here.

Systematic reviews of diagnostic test accuracy

Systematic reviews of diagnostic test accuracy are used to provide a summary of test performance based on all available evidence, evaluate the quality of published studies, and account for variation in findings between the studies.[2,5] They represent the highest level of evidence in the diagnostic field. Estimates of test accuracy frequently vary between the studies, often due to differences in how test positivity is defined, study design, patient characteristics, and positioning of the test in the diagnostic pathway.[5] Furthermore, diagnostic test accuracy studies have unique design characteristics that require different criteria for critical appraisal compared with other sources of quantitative evidence, and report a paired set of results ('sensitivity and specificity') rather than a single statistic.[1] Consequently, systematic reviews of diagnostic test accuracy studies require different statistical methods for meta-analytical pooling and different approaches for narrative synthesis.[4]

Conduct of the review

All systematic reviews should be carried out according to an a-priori published protocol in order to support the unbiased inclusion of studies and reporting of findings.

Review question

As with other JBI review methodologies, it is recommended that a systematic review of diagnostic test accuracy should begin with a strong clinical question that includes the key elements of the systematic review's inclusion criteria. In general, it will be likely to take the form of: What is the diagnostic accuracy of (index test) compared with (reference test) in (population) for the diagnosis of (diagnosis of interest)?

Inclusion criteria

The mnemonic PIRD is recommended for structuring the inclusion criteria of a systematic review of diagnostic test accuracy. PIRD stands for population, index test, reference test, and diagnosis of interest.

Population

The individuals who undergo the diagnostic test in the included studies should be reflective of those who will be undergoing the test in actual clinical practice. If test results are extrapolated from one group of patients to another, it may result in an inaccurate estimation of test accuracy. Key characteristics to define vary from review to review; however, useful examples to consider are disease stage, symptoms, age, sex, race, and educational status. The reason for the inclusion or exclusion of participants should be explained and based on clear scientific justifications.

Index test

As discussed in the previous subsection, the index test is the diagnostic test whose accuracy is being investigated in the systematic review. In some cases, more than one iteration of a specific test will exist. If so, it should be considered very carefully whether the tests are similar enough to be combined through meta-analysis. Other considerations include the criteria by which the index test results will be categorized as being positive or negative (the decision threshold), who carries out or interprets the test (level of expertise or qualification may be an issue), the conditions that the test is conducted under (i.e., laboratory, clinical), and specific details regarding how the test will be conducted.

Reference test

The reference test is the 'gold standard' test against which the results of the index test will be compared. Consequently, it should be the best test currently available for the diagnosis of the condition of interest. Otherwise, the same considerations that apply to the index test also apply to the reference test.

Diagnosis of interest

This item relates to what diagnosis is specifically being investigated in the included studies. This may be a disease, injury, disability, or any other condition. In some

cases where the index or reference tests are only used for one purpose or where the 'population' specifies a suspected condition, this factor may seem redundant. However, overall, it is often informative to specify the diagnosis of interest, particularly when it comes to designing the search strategy.

An example review question constructed using the PIRD structure is: 'What is the diagnostic accuracy of currently available laboratory tests for swine flu (H1N1) compared with viral culture as a reference test among individuals presenting with suspected flu?' In this example, the population is 'individuals presenting with suspected flu'; the index test is 'currently available laboratory tests'; the reference test is 'viral culture'; and the diagnosis of interest is 'swine flu (H1N1)'.

Search strategy

As with all other types of systematic reviews, the documentation of search strategies is a key element of the scientific validity of systematic reviews of diagnostic test accuracy. The standard three-step JBI search strategy of an initial limited search to identify relevant keywords and indexing terms, followed by a second thorough search across all included databases (Pubmed, Embase, and CINAHL are likely defaults for inclusion, but it may be appropriate to include specialist databases based on the topic of the review being undertaken), and then a final review of the reference lists of included studies, is recommended for systematic reviews of diagnostic test accuracy.[3] Logic grids that utilize Boolean operators (Table 2) are useful for structuring searches and should be based on the elements of the PIRD. Although the number of articles retrieved by such searches may be very large and methodological filters consisting of text words and database indexing terms have been developed to increase the precision of searches, the use of filters to identify records for diagnostic reviews may miss relevant studies while at the same time not making a big difference to the number of studies that have to be assessed for inclusion. A systematic review assessed the performance of 70 filters (taken from 19 studies) for identifying diagnostic studies in Medline and Embase. It was found that the search filters did not perform consistently, and none of them met the author's minimum criteria: a sensitivity greater than 90% and a precision above 10%.[13] As such, the authors recommended against relying on methodological filter search terms for systematic reviews of diagnostic test accuracy.

Table 2. Structure of a logic grid

Population	Index	Reference	Diagnosis
. . .. . .	. . ...	. . ...	. . ...
OR	OR	OR	OR
. . ....	. . ...	. . ...	. . ...
OR	OR	OR	OR
. . ...	. . ...	. . ...	. . ...
	AND	AND	AND

As studies on diagnostic accuracy are often based on routinely collected data rather than preregistered trials, publication bias may have an increased effect on diagnostic research.[1] Consequently, searching for difficult-to-locate studies (gray or unpublished literature) and studies in languages other than English is of increased importance in systematic reviews of diagnostic test accuracy.

Study selection

Following the completion of the search strategy, the total list of identified references should be screened for duplicates (the use of bibliography management software such as Endnote or Papers is advised to help manage this). Titles and abstracts should then be reviewed for relevance to the inclusion criteria, followed by the retrieval and review of full texts. At each stage, it is important that the total number of included and excluded studies be kept track of, as these should be reported in the final publication. When conducting the full-text review, the reasons for excluding articles should also be recorded as these need to be included in the final report, either in the flow chart or as an appendix.

Critical appraisal

Assessing the methodological quality of the diagnostic studies that have been included is a vital part of the systematic review process. Methodological quality relates to the risk of bias resulting from the design and conduct of the study. The quality of a diagnostic study is determined by its design, methods by which the study sample is recruited, the conduct of tests involved, blinding in the process of interpreting tests, and the completeness of the study report. The process of critical appraisal examines the methodology of a study against predefined criteria, with the aim of considering individual sources of risk of bias. The most widely used tool for examining diagnostic accuracy is the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2) that was released in 2011 following the revision of the original QUADAS tool.[14] We recommended that reviewers use the QUADAS-2 tool when undertaking their critical appraisal and its 'signaling questions' are included in Table 3. The questions in this critical appraisal checklist are split into four domains: Patient selection, which addresses the risk of selection bias created by how patients were selected for the study; Index test, which

Table 3. QUADAS-2 signaling questions

Critical appraisal questions
Domain 1: Patient selection
Was a consecutive or random sample of patients enrolled?
Was a case–control design avoided?
Did the study avoid inappropriate exclusions?
Domain 2: Index test
Were the index test results interpreted without knowledge of the results of the reference standard?
If a threshold was used, was it prespecified?
Domain 3: Reference test
Is the reference standard likely to correctly classify the target condition?
Were the reference standard results interpreted without knowledge of the results of the index test?
Flow and timing
Was there an appropriate interval between index test and reference standard?
Did all patients receive the same reference standard?
Were all patients included in the analysis?

addresses the risk of bias created by how the index test was conducted and interpreted; Reference test, which investigates the same for the reference test; and Flow and timing, which investigates the risk of bias attributable to the order in which the index and reference tests were conducted in the study. If there is a long time delay between the conduct of the two tests, the status of the patient may change and therefore impact the results of the later test. In addition, if the later test is conducted with the knowledge of the results of the previous test, interpretation of the results may be impacted.

It should be noted that the full version of QUADAS-2 includes questions regarding the level of concern that reviewers have for the applicability of the study under consideration for the research question. If the guidance under 'Study selection' is followed, then there should be no concern that the study lacks applicability to the research question. A full description of the criteria that should be met for each question in QUADAS-2 can be found in Whiting *et al.*[15]

Critical appraisal questions should be answered as 'Yes', 'No', 'Unclear', or, on some occasions, 'Not applicable'. All studies included in a systematic review should be independently appraised by two reviewers, with disagreements resolved through discussion or by seeking the opinion of a third reviewer. Disagreements can be minimized by discussing each item of appraisal with regard to what constitutes acceptable levels of information to allocate a positive response compared with a negative, or a response of 'unclear'. This discussion should take place before independently conducting the appraisal. How much weight is placed on specific critical appraisal questions will vary between reviews, and it is up to the reviewers to set what criteria, if any, will result in the exclusion of a study from the review. Many reviewers specify a set of questions that must be answered 'Yes' or the study will be excluded. It is important that these criteria be applied consistently across studies.

The results of critical appraisal should be reported in the final published review report. This should be a narrative summary of the overall methodological quality of the included studies, which may be directly supported by a table showing the results of the critical appraisal.

Data extraction

The use of a standardized data extraction tool is recommended to facilitate the extraction of the same types of data across all of the included studies. The decision threshold that was used to classify results as positive or negative is an item of data extraction unique to studies of diagnostic test accuracy. However, it is one of the most important details to extract. In addition to recording the final results of the primary study, it is also important to extract the details that inform the studies' generalizability and context. The standards for the reporting of diagnostic accuracy studies (STARD) checklist and flow diagram provide detailed guidance on what studies of diagnostic test accuracy should report and has been used as the basis for the suggested data extraction tool included in the supplementary materials of this article (Appendix I).[16]

All studies of diagnostic test accuracy that comply with the STARD statement will include a 2×2 table that classifies patient test results and disease status similar in format to Table 1; this will provide all of the quantitative data that need to be extracted for the systematic review.

Data synthesis

A key element of systematic reviews is data synthesis. Owing to the paired nature (specificity and sensitivity) of diagnostic test accuracy data, this process requires a unique approach.

Graphic representation

The results of systematic reviews of diagnostic test accuracy can be graphically represented through two different major ways. The first way is the use of forest plots; however, in order to present data on diagnostic

test accuracy, 'paired' forest plots must be used wherein two forest plots are presented side by side: one for sensitivity and the other for specificity. In this way, these graphs display the means and confidence intervals for sensitivity/specificity for each of the selected primary studies. These values are also listed in numerical form along with the number of true positives, false positives, true negatives, and false negatives and, wherever appropriate, any covariates (for instance, the type of diagnostic test used). The below example shows a paired forest plot made using mock data (Fig. 2).

It may also be useful to create summary ROC (SROC) curves, which are graphs with 1-specificity on the *x*-axis and sensitivity on the *y*-axis. Each primary study contributes to a unique point to the plot defined by its sensitivity and specificity for a given threshold value. Point size may vary according to sample size and, to indicate more precisely the precision of the estimates, point height may differ from point width, with these being respectively proportional to the number of diseased and control patients. With rigorous meta-analysis, a curve can be added to the graph. Figure 3 shows an SROC curve made using mock data in RevMan5 (Cochrane Collaboration; http://tech.cochrane.org/revman/download).

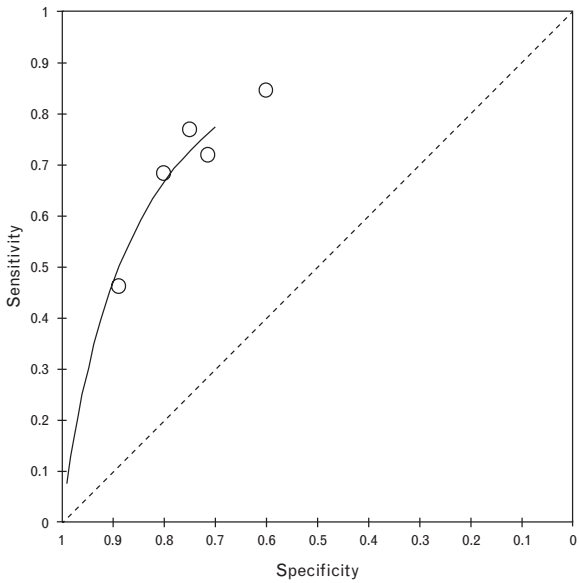

Figure 3. An summary receiver-operating characteristic curve generated using mock data in RevMan5. Sensitivity is on the *y*-axis, and the *x*-axis shows inverted specificity.

Meta-analysis

The paired nature of diagnostic test accuracy data makes meta-analysis complicated. For this reason, the early involvement of a statistician is advisable for reviewers who do not possess considerable statistical expertise of their own. Key issues to consider when planning the meta-analysis of diagnostic test accuracy data are whether a summary sensitivity and specificity should be estimated, and whether an SROC curve should be computed. This depends on the nature of the data available and, more exactly, whether the diagnostic threshold was the same across the selected primary studies. Sometimes all of the included studies will have rigorously used the same diagnostic threshold, but, on other occasions, variations in the threshold will exist. This is often the case when there is no explicit numerical cutoff point or when the index test is based on an observer's judgment.

The basic strategy is that when the same threshold is used throughout the primary studies, the summary sensitivity/specificity should be estimated, and when different thresholds are used, an SROC curve should be computed to display the data graphically, whereas summary sensitivity/specificity should be estimated for each different threshold value used in the articles. It should be noted that if a study has referred to sensitivity/specificity values for several thresholds, it can contribute to several estimations of summary sensitivity/specificity.

The optimum methods for performing meta-analyses regarding diagnostic test accuracy and when to use them are still being debated in the literature and new statistical developments are underway.[17,18] However, three main models exist with one corresponding to fixed-effect models and the other two being random-

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Example 1	6	8	7	64	0.46 [0.19, 0.75]	0.89 [0.79, 0.95]
Example 2	13	4	6	16	0.68 [0.43, 0.87]	0.80 [0.56, 0.94]
Example 3	23	18	4	27	0.85 [0.65, 0.96]	0.60 [0.44, 0.74]
Example 4	18	12	7	30	0.72 [0.51, 0.88]	0.71 [0.55, 0.84]
Example 5	20	10	6	30	0.77 [0.56, 0.91]	0.75 [0.59, 0.87]

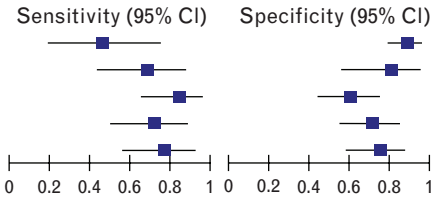

Figure 2. An example paired forest plot generated using mock data in RevMan5. Numerical values for sensitivity and specificity are presented alongside graphical representations where the boxes mark the values and the horizontal lines show the confidence intervals (CIs). FN, false negative; FP, false positive; TN, true negative; TP, true positive.

effects models that take into account the variability present within studies and between studies. These are described further in brief.

The Moses–Littenberg model[19,20] has been extensively used for meta-analyses of diagnostic test accuracy.[21] However, it is principally a fixed-effect model, whereas for many such analyses, a random-effects model is required. It allows the performance of SROC curves in an exploratory approach. But, as a fixed-effect model, it does not take into account and does not consider the variability between studies. As such, the Moses–Littenberg model can, in some circumstances, produce very different SROC curves compared with the other two main models.[22] The Cochrane Collaboration recommends careful use of the Moses–Littenberg model that should be limited to preliminary analyses. The estimation of confidence intervals and investigations of heterogeneity should not be carried out with this model.[10]

The bivariate model[23] estimates the summary parameters sensitivity and specificity across included studies. It is held to be a method of choice in the Cochrane Handbook[10] and in the article by Leeflang.[4] In this method, following Chu and Cole,[24] the within-study variability is modeled by binomial distributions, one for the sensitivity and the other for specificity. These distributions are jointly treated as estimates of sensitivity and specificity and are nonindependent.

The hierarchical SROC model was specifically created to deal with variability in positivity threshold values. It produces an SROC in which each study provides only one pair of values for sensitivity and specificity. It is presented as a method of choice to obtain SROC curves in the Cochrane Handbook and in the article by Leeflang.[4,10]

Heterogeneity

A finding of heterogeneity between studies is especially common in systematic reviews of diagnostic test accuracy.[10] This is often due to differences in study populations, procedures followed for carrying out tests, and the conditions or context of testing; however, heterogeneity can also be the result of telling differences in how studies have been conducted or their data analyzed that have biased the results (i.e., one study may include all results in their analysis, whereas another may exclude inconclusive outcomes, consequently making the test appear more accurate than it is). As such, when heterogeneity is found between diagnostic accuracy studies, it should be carefully investigated. The graphic display of diagnostic accuracy data through paired forest plots (Fig. 2) or SROC curves (Fig. 3) can help provide a subjective assessment of the presence or absence of heterogeneity as large differences between studies, if present, will be recognizable. However, if there are differences in the diagnostic threshold between studies, paired forest plots should not be used as variability will exist because of the interdependence of sensitivity and specificity. In these cases, heterogeneity should be estimated by judging how well studies fit to the SROC curve (and not by how scattered they are). The χ^2 or Fisher exact tests can be used to more objectively assess heterogeneity, but their power has been found to be low.[25] The I^2 test is not routinely used in systematic reviews of diagnostic test accuracy as it does not account for the influence of differing decision thresholds. Subgroup analysis can be used to investigate potential sources of heterogeneity (keeping in mind that it is best practice that all subgroup analyses undertaken be prespecified in the a-priori published protocol). However, when the extent and cause of heterogeneity cannot be explained, reviewers should refrain from meta-analysis and instead conduct a narrative synthesis.

Reporting and discussing results

The process followed, from the search to the final selection of studies for extraction and synthesis, should be reported at the beginning of the results section. The use of a flow chart conforming to the PRISMA (Preferred Reporting Items for Systematic Review and Meta-analysis) statement is suggested (and is a requirement of a growing proportion of journals that accept systematic reviews for publication).[26] Systematic reviews of diagnostic studies should also be accompanied by a summary of findings table, which can be created using the software program Guideline Development Tool (www.guidelinedevelopment.org) available free online. Further guidance on creating a summary of findings table for systematic reviews of diagnostic test accuracy, specifically using the Grading of Recommendations Assessment, Development, and Evaluation approach, can be found in Gopalakrishna *et al.*[27]

In order to provide context for the review findings, the results section should also include an overall description of all the included studies. This should provide sufficient detail for readers to assess the similarities and differences between studies, with a view to informing the appropriateness of meta-analysis. Items of relevance that reviewers may wish to highlight here include characteristics of the participants, the settings in which the tests were conducted, and specific study designs used. Generally, tables will be the most appropriate form for presenting this data. It is important to note that the presence of extensive detail on study characteristics may obscure the actual findings and make the review less accessible to the reader. The

methodological quality of the included studies, as determined by critical appraisal checklist used, should also be reported. This should take the form of a narrative summary of the overall methodological quality of the included studies, directly supported by a table showing the results of the critical appraisal. The main findings of the systematic review, those relating to the primary and secondary outcomes, should be presented in the same order as the relevant review questions in order to create a logical flow. The use of tables and appendices should again be considered in order to avoid obscuring important details with an excess of less-important items. The discussion section should then focus on the effects of the review findings on the field of diagnostics related to the test(s) under review, as well as their influence on patients and other relevant issues. The discussion should also include a final overview of the results that address any limitations or issues arising from the results or conduct of the review. Recommendations for practice and future research should also be made.

Conclusion

The systematic review of studies of diagnostic test accuracy is vital for effective healthcare and it is therefore imperative that they be carried out in a rigorous, transparent, and replicable manner. It is hoped that the guidance provided in this article helps toward that goal. Other references that reviewers may find useful to this end include the diagnostic test accuracy section of the Cochrane Handbook,[5] '*Systematic reviews and meta-analyses of diagnostic test accuracy*' by Leeflang,[4] and '*Systematic reviews of evaluations of diagnostic and screening tests*' by Deeks.[2] For the conduct of meta-analysis, the JBI is undertaking to develop a comprehensive diagnostic test accuracy module for its SUMARI (System for the Unified Management Assessment and Review of Information) suite. At the time of this publication that will not have been completed, however, later readers are urged to check. The meta-analysis suite provided by the Cochrane collaboration RevMan includes some statistical tools for diagnostic test accuracy; however, it is not comprehensive. For reviewers with a high level of statistical competency, nonspecialized statistical software such as SAS or R may provide the optimum tools.

Acknowledgements

The authors report no conflicts of interest.

References

1. White S, Schultz T, Enuameh YAK (Eds): *Synthesizing evidence of diagnostic accuracy*. Philadelphia: Lippincott Williams and Williams, 2011.
2. Deeks JJ. Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. *BMJ* 2001; 323: 157–62.
3. Institute TJB. Joanna Briggs Institute Reviewers' Manual: 2014 ed. Adelaide: The Joanna Briggs Institute; 2014.
4. Leeflang MM. Systematic reviews and meta-analyses of diagnostic test accuracy. *Clin Microbiol Infect* 2014; 20: 105–13.
5. Leeflang MM, Deeks JJ, Takwoingi Y, Macaskill P. Cochrane diagnostic test accuracy reviews. *Syst Rev* 2013; 2: 82.
6. Sackett DL, Haynes RB. The architecture of diagnostic research. *BMJ* 2002; 324: 539–41.
7. Habbema J, Eijkemans R, Krijnen P, Knottnerus J. Analysis of data on the accuracy of diagnostic tests. In: Knottnerus J, Buntinx F, editors. *The evidence base of clinical diagnosis: theory and methods of diagnostic research.* 2nd ed. London: BMJ Publishing Group, 2009. 118–145.
8. Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation* 2007; 115: 654–7.
9. Metz CE. Basic principles of ROC analysis. *Sem Nuclear Med* 1978; 8: 283–98.
10. Macaskill P, Gatsonis C, Deeks J, *et al*. Chapter 10: Analysing and presenting results. In: Deeks J, Bossuyt P, Gatsonis C, editors. *Cochrane handbook for systematic reviews of diagnostic test accuracy*. London: the Cochrane Collaboration, 2010
11. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143: 29–36.
12. Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity. *Contin Educ Anaesth Crit Care Pain* 2008; 8: 221–3.
13. 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.
14. Whiting P, Rutjes AWS, Westwood ME, *et al*. QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Int Med* 2011; 155: 529e36.
15. 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: 529–36.
16. Gatsonis C. Do we need a checklist for reporting the results of diagnostic test evaluations? The STARD proposal. *Acad Radiol* 2003; 10: 599–600.
17. Eusebi P, Reitsma JB, Vermunt JK. Latent class bivariate model for the meta-analysis of diagnostic test accuracy studies. *BMC Med Res Methodol* 2014; 14: 88.
18. Borenstein M, Hedges LV, Higgins J, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res Synth Methods* 2010; 1: 97–111.
19. Littenberg B, Moses LE, Rabinowitz D. Estimating diagnostic-accuracy from multiple conflicting reports – a new meta-analytic method. *Clin Res* 1990; 38: A415–20.
20. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic-test into a summary ROC curve –

data-analytic approaches and some additional considerations. *Stat Med* 1993; 12: 1293–316.
21. Holling H, Bohning W, Bohning D. Meta-analysis of diagnostic studies based upon SROC-curves: a mixed model approach using the Lehmann family. *Stat Model* 2012; 12: 347–75.
22. Harbord RM, Whiting P, Sterne JAC, *et al*. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. *J Clin Epidemiol* 2008; 61: 1095–103.
23. Retisma JB, Rutjes AWS, Whiting P, Vlassov VV, Leeflang MMG, Deeks JJ. Chapter 9: Assessing methodological quality. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.0*. York: The Cochrane Collaboration, 2009.
24. Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *J Clin Epidemiol* 2006; 59: 1331–2.
25. Dinnes J, Deeks J, Kirby J, Roderick P. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. *Health Tech Assess* 2005; 9: 1–113.
26. Liberati A, Altman DG, Tetzlaff J, *et al*. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* 2009; 339: b2700.
27. Gopalakrishna G, Mustafa RA, Davenport C, *et al*. Applying Grading of Recommendations Assessment, Development and Evaluation (GRADE) to diagnostic tests was challenging but doable. *J Clin Epidemiol* 2014; 67: 760–8.

Appendix I. Data extraction instrument

Author/Date	
Inclusion/exclusion criteria (i.e., presenting symptoms, results from previous tests)	Inclusion: Exclusion:
Sample size	
Participant demographics (i.e., age, sex, spectrum of presenting symptoms, comorbidity, current treatments, recruitment centers)	
Study methodology (consecutive or random; retrospective or prospective)	
Period that study was carried out (beginning and end date)	
Index test description (including criteria for positive test)	
Reference test description (including criteria for positive test)	
Geographical location of data collection	
Setting of data collection	
Persons executing and interpreting index tests (numbers, training, and expertise)	
Persons executing and interpreting reference test	
Index/reference time interval (and treatments carried out in between)	
Distribution of severity of disease in those with target condition	
Other diagnoses in those without target condition	
Adverse events from index test	
Adverse events from reference test	

Index test results threshold =	Condition positive	Condition negative	Total
Index test positive (T+)			
Index test negative (T–)			
Total			