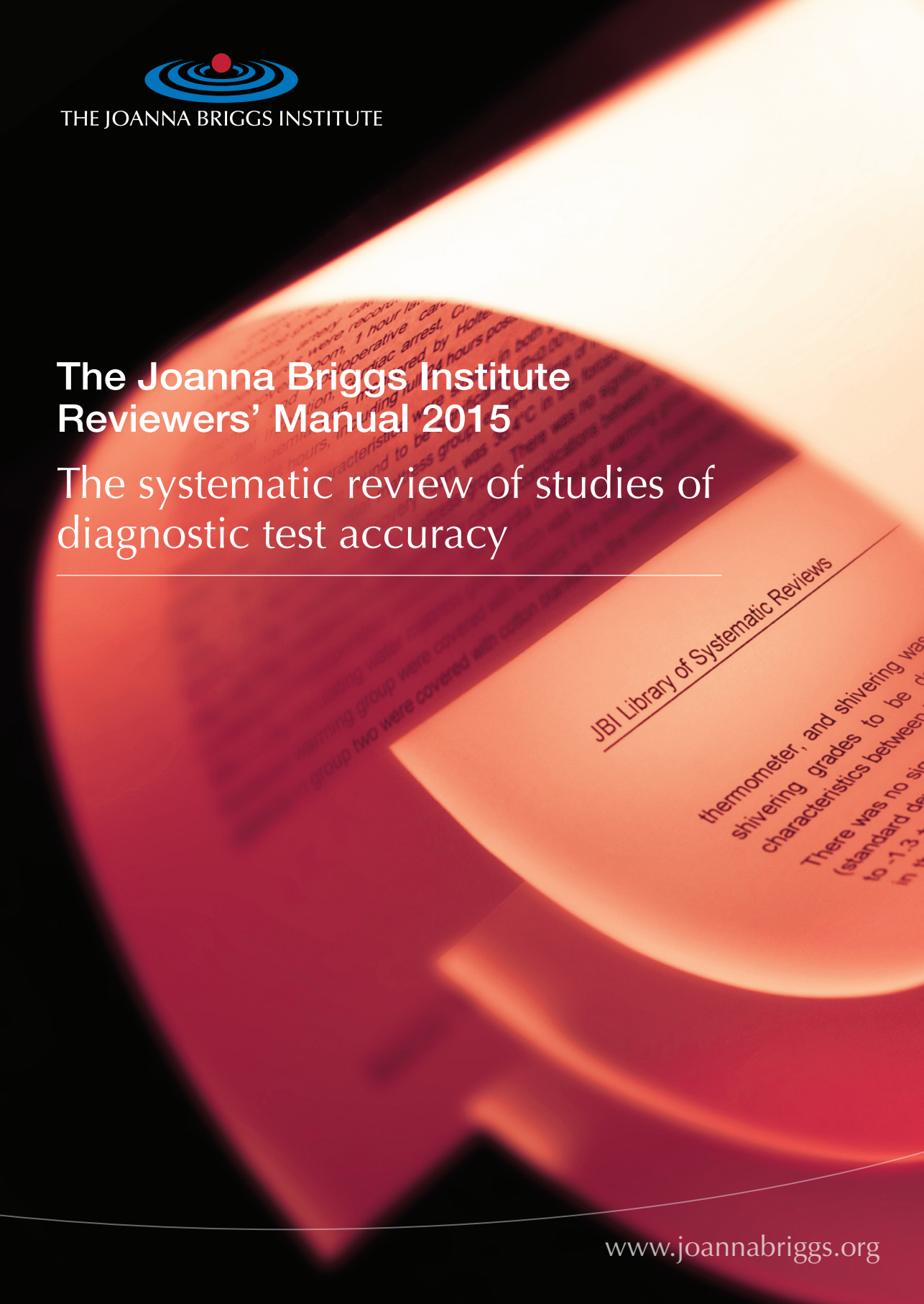
THE JOANNA BRIGGS INSTITUTE

The Joanna Briggs Institute Reviewers' Manual 2015

The systematic review of studies of diagnostic test accuracy



www.joannabriggs.org

Joanna Briggs Institute Reviewers' Manual: 2015 edition/supplement

The Joanna Briggs Institute
The University of Adelaide
South Australia 5005

AUSTRALIA

ABN: 61 249 878 937

Phone: +61 8 8313 4880
Fax: +61 8 8313 4881
Email: jbi@adelaide.edu.au
Web: www.joannabriggs.org

Published by the Joanna Briggs Institute, 2015

Author:	The Joanna Briggs Institute
Tile:	Joanna Briggs Institute Reviewers' Manual: 2015 edition / Supplement
Publisher:	The Joanna Briggs Institute
Subjects:	The systematic review of studies of diagnostic test accuracy

Foreword

Every year the Joanna Briggs Institute publishes a Reviewers' Manual, which is designed to support individuals who are undertaking systematic reviews following JBI methodologies and methods. This chapter represents the latest work and methodological development of the Institute that was not ready for inclusion in the 2014 edition of the Reviewers' Manual that was published in January.

As with the Reviewers' Manual we recommend that this chapter be utilized in conjunction with the JBI SUMARI User Guide. Please note that this chapter makes reference to forthcoming analytical modules that do not currently exist in the JBI SUMARI software suite, but should be available in 2015. For advice on how to best apply the current software to accommodate this new methodology please contact the Synthesis Science Unit of the Institute at jbisynthesis@adelaide.edu.au

We hope that the information contained herewith provides further insight into how to analyze and synthesise different types of evidence to inform clinical and policy decisions to improve global health outcomes.

Associate Professor Edoardo Aromataris

Director, Synthesis Science

Contents

The systematic review of studies of diagnostic test accuracy

Jared M. Campbell
The Joanna Briggs Institute, Faculty of Health Sciences, University of Adelaide, 115 Grenfell Street, Adelaide, Australia

Miloslav Klugar
The Czech Republic (Middle European) Centre for Evidence-Based Health Care: An Affiliated Centre of The Joanna Briggs Institute, Department of Social Medicine and Public Health, Faculty of Medicine and Dentistry

Sandrine Ding
Bureau d'Echanges des Savoirs pour des praTiques exemplaires de soins (BEST): an Affiliate centre of The Joanna Briggs Institute, Av. Beaumont 21, 1011 Lausanne, Switzerland

Dennis P. Carmody
Rutgers Robert Wood Johnson Medical School, Institute for the Study of Child Development, 89 French Street, New Brunswick, NJ 08901, United States.

Sasja J. Hakonsen
Center for Kliniske Retningslinjer, Institut for Medicin & Sundhedsteknologi, Aalborg Universitet, Denmark

Yuri T. Jadotte,
Northeast Institute for Evidence Synthesis and Translation, a Collaborating center of the Joanna Briggs Institute, Division of Nursing Science, Rutgers School of Nursing

Department of Quantitative Methods, Biostatistics and Epidemiology, Rutgers School of Public Health, Newark, NJ 07103, USA

Sarahlouise White
The Joanna Briggs Institute, Faculty of Health Sciences, University of Adelaide, 115 Grenfell Street, Adelaide, Australia

Zachary Munn
The Joanna Briggs Institute, Faculty of Health Sciences, University of Adelaide, 115 Grenfell Street, Adelaide, Australia

Background

Diagnostic tests are used by clinicians to identify the presence or absence of a condition in a patient for the purpose of developing an appropriate treatment plan.[1] They can include imaging and biochemical technologies, pathological and psychological investigation, and signs and symptoms observed during history taking and clinical evaluations.[2] New diagnostic tests are continuously developed, driven by demands for improvements in speed, cost, ease of performance, patient safety and accuracy.[1] Consequently there are often several tests available for the diagnosis of a particular condition. This highlights the importance of clinicians and other healthcare practitioners having access to high level evidence on the accuracy of the diagnostic tests they use or are considering using. The end goal of diagnostic tests is that they result in improved outcomes in areas that are important to the patient. Systematic reviews that investigate whether diagnostic tests improve outcomes are reviews of effectiveness, however, and should be carried out using the methodology from the chapter on effectiveness. Primary studies that investigate the accuracy of diagnostic tests are termed diagnostic test accuracy (DTA) studies, and it is the systematic review of these which will be the focus of this chapter.

Diagnostic test accuracy studies compare a diagnostic test of interest (the 'index test') to 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 with one another in order to evaluate the accuracy of the index test. There are two main types of studies of DTA. The first is the diagnostic case-control design, also sometimes called the 'two gate design'. In this study design people with the condition (cases) come from one population (i.e. a health care centre for people known to have the condition), while people without the condition come from another. Although this design gives an indication of the maximum accuracy of the test, the results will generally give an exaggerated indication of the test's accuracy in practice.[3]

The second study design is cross-sectional, and involves all patients suspected of having the condition of interest undergoing the index test and the reference test. Those who test positive for the condition by the reference test can be considered to be 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.[3]

Systematic reviews of diagnostic test accuracy provide a summary of test performance based on all available evidence, evaluate the quality of published studies, and account for variation in findings between studies.[2, 3] Estimates of test accuracy frequently vary between studies, often due to differences in how test positivity is defined, study design, patient characteristics and positioning of the test in the diagnostic pathway.[3] Furthermore, DTA studies have unique design characteristics which require different criteria for critical appraisal compared to other sources of quantitative evidence, and report a pair of related summary statistics ('sensitivity and specificity', as discussed below) rather than a single statistic such as an odds ratio. Consequently systematic reviews of DTA studies require different statistical methods for meta-analytical pooling, and different approaches for narrative synthesis.[4]

Diagnostic accuracy is predominantly represented by two measures, sensitivity and specificity; however sometimes other measures, including predictive values, odds-ratios, likelihood ratios, and summary 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 [TPP]), while specificity is the probability of a person without the condition of interest having a negative result (also known as the true negative proportion [TNP]).[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.[5] Sensitivity and specificity have been identified as essential measures of diagnostic accuracy.[3, 4, 6, 7]

Measures of diagnostic accuracy

Several pairs of measures are used to determine how well a diagnostic test performs relative to the known proportions of individuals with and without the disorder. Diagnostic accuracy is critical in the evaluation of medical diagnostic tests.[7] Methods to summarize the results of diagnostic studies are available for both binary and continuous data.[8, 9] Measures of overall accuracy are affected by the prevalence of the disorder.[10] In addition, estimates may vary greatly between studies due to differences in the criteria used to declare a test positive, patient characteristics, and study design.[3]

Sensitivity and specificity

The most commonly used measures are sensitivity and specificity. Sensitivity is the probability that a person with the condition of interest will have a positive result, while specificity is the probability of a person without the condition having a negative result.[11]

Specifically, sensitivity can be calculated as $\frac{True\ positives}{(True\ positives+False\ negatives)}$ while specificity can be calculated as $\frac{True\ negatives}{(True\ negatives+False\ positives)}$

Table 1: Classification of patient test results by condition status

Index Test Outcome	Reference positive	reference negative	Total
Index test positive (T+)	True positives (TP)	False positives (FP)	Test positives (TP+FP)
Index test negative (T-)	False negatives (FN)	True Negatives (TN)	Test negatives (FN+TN)
Total	Reference positives (TP+FN)	Reference negatives (FP+TN)	N (TP+FP+FN+TN)

Sensitivity and specificity co-vary with the decision threshold used to identify the disorder.[12]

In Table 2 example data is presented from Mulligan EP et al. 2011, who investigated the diagnostic test accuracy of the Lachman test performed in prone position for the diagnosis of torn anterior cruciate ligament (ACL).[13]

Table 2: Results from Mulligan EP et al. 2011

Prone Lachman	Reference Positive	Reference Negative	Total
Positive	16	1	17
Negative	7	28	35
Total	23	29	52

For this study the sensitivity can be calculated as $\frac{16}{16+7} = 0.70$ while the specificity is $\frac{28}{28+1} = 0.97$

Predictive values

While sensitivity and specificity measure the accuracy of a diagnostic test, they do not provide the probability of the diagnostic value of the result of the test. Predictive values provide the proportion of patients who are correctly diagnosed.[14] The positive predictive value $PPV = \frac{TP}{(TP+FP)}$ is the proportion of individuals with positive test results who were correctly diagnosed, while the negative predictive value $NPV = \frac{TN}{TN+FN}$ is the proportion of individuals with negative test results who were correctly diagnosed.

From the example presented in Table 2, $\mathbf{PPV\ is}\frac{16}{16+1} = 0.94$ while the NPV is $\frac{28}{28+7} = 0.8$

As prevalence does influence predictive values, it is important to account for the prevalence of the disorder in the population under study, given that the higher the prevalence the higher the PPV.[15]

Likelihood ratios

Likelihood ratios assess the probability or likelihood that the test result obtained would be expected in a person with the condition, compared to the probability or likelihood that the same result would be seen in a person without the condition.[2] The positive likelihood ratio $LR+= \frac{sensitivity}{(1-specificity)} = \frac{TP}{(TP+FN)} \div \frac{FP}{(FP+TN)}$ expresses how many times more likely people with the condition are to receive a positive test result compared to those who do not have the condition, while the negative likelihood ratio $LR-= \frac{(1-sensitivity)}{(specificity)} = \frac{FN}{(TP+FN)} \div \frac{TN}{(FP+TN)}$ expresses how likely it is that people with the condition will receive a negative test result compared to those who do not have the condition.

From the example presented in Table 2, $\mathrm{LR+\ is}\frac{0.70}{1-0.97} = 23.33$ while $\mathrm{LR\text{-}\ is}\frac{1-0.70}{0.97} = 0.31.$

The initial assessment of the likelihood of a disorder, that is the *a priori* probability, is modified by the results of the diagnostic test for a *posteriori* probability (the probability actually observed). A suggestion on the limited use of likelihood ratios is that their interpretation requires a calculator to convert between probabilities and odds of the disorder.[16]

ROC analyses

Receiver Operating Characteristic (ROC) curve analysis is useful for evaluating the performance of diagnostic tests that classify individuals into categories of those with and those without a condition.[17, 18] 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 'line' is drawn is termed the decision or positivity threshold. For example, a blood pressure cut-off value for hypertension is 135/80.

The choice of a decision threshold will have a large effect on the sensitivity and specificity of a test. While 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. The inverse is also 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.[19] This gives a visual representation of the relationship between sensitivity and specificity of a diagnostic test as the threshold value changes. This can be measured quantitatively by assessing the area under the curve (AUC).[20] 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]

Figure 1 shows an ROC curve from Erol B. et al 2014 with an AUC of 0.81 (95%CI 0.80 to 0.82). [21] The diagonal line shows the baseline result of a test with no differential power (AUC=0.5).

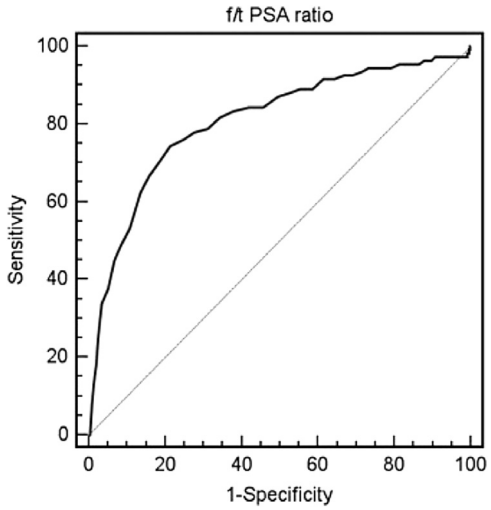


Figure1: ROC graph for the use of prostate specific antigen free/total ratios for the diagnosis of prostate cancer

Protocol and title development

Title

The title should be clear, explicit and reflect the core elements of the review. This creates the best chance of correct indexing by databases and easy identification by interested readers. The title should explicitly state that it is on 'diagnostic test accuracy' and include the phrase '…: a systematic review protocol'. Titles should not be phrased as questions or conclusions and there should be congruency between the title, review objectives/questions and inclusion criteria.

The title should include each of the elements of the PIRD acronym (discussed below), and approximately follow the form of: "The accuracy of **INDEX** relative to **REFERENCE** for the diagnosis of **DIAGNOSIS** in **POPULATION** a systematic review of diagnostic test accuracy"

Example titles are:

- "The accuracy of laboratory testing relative to viral culture for the diagnosis of influenza A (H1N1) 'swine flu' in people presenting with suspected flu: a systematic review protocol of diagnostic test accuracy"
- "The accuracy of endoscopic retrograde cholangiopancreatography relative to confirmed bile stone extraction for the diagnosis of common bile duct stones in in patients with clinical symptoms of common bile duct stones: a systematic review protocol of diagnostic test accuracy"
- The accuracy of automated semen analysis for the diagnosis of infertility in the male partner of an infertile couple relative to laboratory technician analysis: a systematic review protocol of diagnostic test accuracy"

Review question or objective

Developing a good review question/objective is an important step in undertaking a quality systematic review as it sets out the key components of the review (i.e. population, index test, reference test, objectives).

An example of a well written review objective is "To determine the diagnostic accuracy of currently available laboratory tests for swine flu (H1N1) using viral culture as a reference test amongst people presenting with suspected flu" which could alternatively be phrased as the question "What is the diagnostic accuracy of currently available laboratory tests for swine flu (H1N1) compared to viral culture as a reference test amongst people presenting with suspected flu?"

Inclusion/exclusion criteria

The mnemonic PIRD is recommended for setting the inclusion criteria for systematic reviews of diagnostic test accuracy:

- Population
- Index test
- Reference test
- Diagnosis of interest

Population

The types of participants should be appropriate for the review objectives and reflect who will undergo the diagnostic test in clinical practice. If test results are extrapolated to other populations, this may result in an inaccurate estimation of test accuracy and should therefore be avoided. The reasons for the inclusion or exclusion of participants should be explained in the Background section and be based on clear scientific justifications. Population characteristics that may be relevant to outline in detail include disease stage, symptoms, age, sex, race, educational status, etc. An example 'population' is "individuals presenting with flu symptoms".

Index test

The index test(s) is the diagnostic test whose accuracy is being investigated in the review. Sometimes multiple iterations of a specific test will exist, and it should be specified at the protocol stage what criteria will be used to determine if the tests are similar enough to combine in meta-analysis. The criteria by which the index test results will be categorized as being positive or negative (the threshold value) can also be specified at the protocol stage. It may be appropriate for reviewers to specify who carries out or interprets the test, the conditions under which the test is conducted (i.e. laboratory, clinical), and specific details regarding how the test will be conducted.

An example of 'index test' is "currently available laboratory tests (PCR test)".

Reference test

The reference test is the 'gold standard' test to which the results of the index test will be compared. It should be the best test currently available for the diagnosis of the condition of interest. The same standards for describing the index test should be followed for describing the reference test in the protocol; the details of what criteria will be used to determine which tests 'count' as being the reference test, and how results will be categorized as positive or negative should be specified. Systematic reviews of diagnostic test accuracy must specify a reference test. An example 'reference test' is "viral culture".

Diagnosis of interest

This item relates to what diagnosis is being investigated in the systematic review. This may be a disease, injury, disability or any other pathological condition. In some cases (i.e. where the index or reference tests are only used for one purpose or the 'population' specifies "patients suspected of…") this factor may seem redundant. However, as a general rule it is useful to explicitly specify the diagnosis of interest. An example 'diagnosis of interest' is "swine flu (H1N1)"

Types of studies

In this section the types of studies which will be considered for inclusion in the review are described. As detailed above, diagnostic studies generally use cross-sectional or case-control study designs. It should be noted however that restricting database searches by study design may result in studies which contain accuracy data being missed.

Search strategy

This section should detail how the reviewers will search for relevant papers. The documentation of search strategies is a key element of the scientific validity of a systematic review. It enables readers to look at and evaluate the steps taken and decisions made to consider the comprehensiveness and exhaustiveness of the search strategy for each included database. Initial keywords should be specified in the protocol along with the databases to be searched. A three-stage search strategy is recommended including an initial search of the select databases using the pre-specified keywords to identify additional relevant keywords and index terms, a second thorough search across all included databases, and then a final review of the reference lists of included studies in order to identify any studies that may have been missed. If searching is restricted to a specific data range, then that should be specified in the protocol, as well as any language restrictions which may be applied. For further information on searching refer to the 'Searching for studies of diagnostic test accuracy' section.

Assessment of methodological quality

Assessing the quality of diagnostic studies being considered for inclusion 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, the 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 pre-defined criteria, with the aim of considering individual sources of risk of bias and is used to evaluate the extent to which the results of a study should be believed or to be deemed valid after rigorous assessment (Reitsma et al., 2009).[22]

Table 3 is modified and expanded from "Synthesizing evidence of diagnostic accuracy" [1, 22] and highlights the major types of bias that can occur in diagnostic accuracy studies as a result of flawed or incomplete reporting. Attempts such as those by the Standards for Reporting of Diagnostic Accuracy (STARD) initiative,[23, 24] have been made to improve reporting, methodological quality and to aid primary researchers to address and avoid sources of bias.

Table 3: Types of bias in studies of diagnostic test accuracy

	Type of bias	When does it occur?	Impact on accuracy	Preventative measures
Patients/Subjects	Spectrum bias	When included patients do not represent the intended spectrum of severity for the target condition or alternative conditions	Depends on which end of the disease spectrum the included patients represent	Ensure that the included patients represent a broad sample of those that the test is intended for use with in clinical practice
	Selection bias	When eligible patients are not enrolled consecutively or randomly	Usually leads to overestimation of accuracy	Consider all eligible patients and enrol either consecutively or randomly

Index test	Information bias	When the index results are interpreted with knowledge of the reference test results, or with more (or less) information than in practice	Usually leads to overestimation of accuracy, unless less clinical information is provided than in practice, which may result in an under estimation of accuracy	Index test results should be interpreted without knowledge of the reference test results, or with more (or less) information than in practice
Reference test	Misclassification bias	When the reference test does not correctly classify patients with the target condition	Depends on whether both the reference and index test make the same mistakes	Ensure that the reference correctly classifies patients within the target condition
	Partial verification bias	When a non-random set of patients does not undergo the reference test	Usually leads to overestimation of sensitivity, effect on specificity varies	Ensure that all patients undergo both the reference and index tests
	Differential verification bias	When a non-random set of patients is verified with a second or third reference test, especially when this selection depends on the index test result	Usually leads to overestimation of accuracy	Ensure that all patients undergo both the reference and index tests
	Incorporation bias	When the index test is incorporated in a (composite) reference test	Usually leads to overestimation of accuracy	Ensure that the reference and test are performed separately
	Disease/ Condition progression bias Perform the reference and index with minimal delay.	When the patients' condition changes between administering the index and reference test	Under- or Over-estimation of accuracy, depending on the change in the patients' condition	Ideally at the same time where practical
	Information bias	When the reference test data is interpreted with the knowledge of the index test results	Usually leads to overestimation of accuracy	Interpret the reference and index data independently
Data analysis	Excluded data	When uninterpretable or intermediate test results and withdrawals are not included in the analysis	Usually leads to overestimation of accuracy	Ensure that all patients who entered the study are accounted for and that all uninterpretable or intermediate test results are explained

The most widely used tool for examining diagnostic accuracy is the QUADAS 2 which was released in 2011 following the revision of the original QUADAS (Quality Assessment of Diagnostic Accuracy Studies) tool.[25] The Joanna Briggs Institute (JBI) encourages the use of QUADAS 2, and this chapter includes a checklist which incorporates the "signaling questions" from QUADAS 2 (Appendix I). It should be noted that QUADAS 2 includes questions regarding the level of concern that reviewers have for the applicability of the study under consideration to the research question. For JBI DTA systematic reviews, a primary research study should not proceed to critical appraisal if there is concern that the study does not match the inclusion criteria and research question. As such, this element of QUADAS2 is not addressed in the below checklist (Domains 1, 2, 3, 4).

Domain 1: Patient selection

In this section the risk of selection bias is assessed by how patients were selected for the study.

- Was a consecutive or random sample of patients enrolled?
- Was a case-control design avoided?
- Did the study avoid inappropriate exclusions?

Domain 2: Index tests

In this section consideration is on whether the conduct and interpretation of the index test being investigated could have introduced bias.

- Were the index test results interpreted without knowledge of the results of the reference standard?
- If a threshold was used, was it pre-specified?

Domain 3: Reference standard/test

The focus of this section is to determine if and the extent that the way in which the reference test was conducted and interpreted could introduce bias into the study.

- 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?

Domain 4: Flow and timing

The aim of this section is to determine 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 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 knowledge of the results of the previous test, interpretation of the results may be impacted.

- Was there an appropriate interval between the index test and reference standard?
- Did all patients receive the same reference standard?
- Were all patients included in the analysis?

The primary and secondary reviewer should discuss each item of appraisal for each study design included in their review. In particular, discussions should focus on what is considered acceptable for the review in terms of the specific study characteristics. The reviewers should be clear on what constitutes acceptable levels of information to allocate a positive appraisal compared with a negative, or a response of "unclear".

This discussion should take place before independently conducting the appraisal. The weight placed on specific critical appraisal questions will vary between reviews and it is up to the reviewers to set what criteria will result in the inclusion/exclusion of a study. Many reviewers select a set of questions which must be answered "Yes" or the review will be excluded. It is important that these criteria be applied consistently across studies.

Data extraction

Data extraction is the process of sourcing and recording relevant results and details from the primary research studies included in the systematic review. Standardized data extraction tools facilitate the extraction of the same types of data across all of the included studies and are required for JBI systematic reviews. Reviewers should practice using the data extraction tool so they are consistently applied. The protocol should detail what data the reviewers will extract from the included studies and the data extraction tool should be attached in the appendices. Among the most important detail to extract is the decision threshold used.

As well as recording the final results of the study it is important to extract the details that inform generalizability and context of the primary studies. The STARD (Standards for Reporting of Diagnostic Accuracy) checklist and flow diagram provides detailed guidance on what studies of DTA to report and the majority of items are incorporated into the standard data extraction template that is appended to this chapter (Appendix II).26 You can download the STARD checklist and STARD flow diagram: http://www.stard-statement.org/

To reduce errors in data extraction it is recommended that two independent reviewers extract data and use the standardized instrument.

Studies of diagnostic test accuracy that comply with the STARD statement should include a 2×2 table that classifies patient test results and disease status as shown below (Table 4):

Table 4: Condition status (reference test results)

Index test Outcome	Condition positive	Condition negative	Total
Index test positive	True positives (a)	False positives (b)	Test positives (a + b)
Index test negative	False negatives (c)	True negatives (d)	Test negatives (c + d)
Total	Disease/condition positives (a + c)	Disease/condition negatives (b + d)	N (a + b + c + d)

This should essentially include all quantitative data that is needed for the extraction.

Data synthesis

Finally, the protocol should describe how the outcome data of the primary studies will be combined and reported, i.e. meta-analysis, narrative synthesis, graphical representation, etc. Options for summarizing and presenting data are discussed further below.

Graphic representation

Results of diagnostic test accuracy systematic reviews can be graphically represented through two different major ways.

As for systematic reviews of effectiveness, forest plots can be performed. In the case of diagnostic test accuracy, two forest plots are presented side by side: one for sensitivity and the other for specificity. These graphs thus show the means and confidence intervals for sensitivity/specificity for each of the selected primary studies. These values are also listed in numerical form. Moreover, the number of true positives, false positives, true negatives and false negatives are also reported, as well as, where appropriate, any covariates (for instance the type of diagnostic test used). The below example shows a paired forest plot made using mock data.[27]

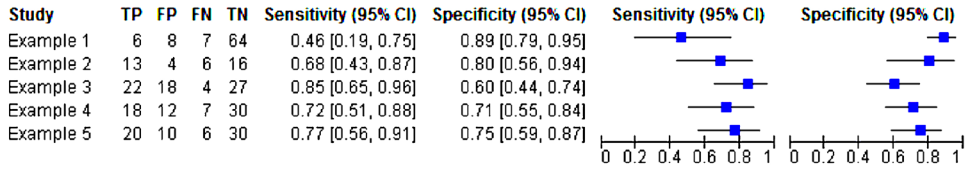


Figure 2: An example paired forest plot generated using mock data in RevMan5

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.[27]

It is also possible to create Summary ROC (SROC) curves. They are graphs with 1-specificity on the x-axis and sensitivity on the y-axis, in which each primary study contributes to a unique point defined by its sensitivity and specificity for a given threshold. If several thresholds are reported in a single primary study, only one sensitivity/specificity pair for that study can be plotted on the SROC graph. Point size may vary according to sample size. To indicate more precisely the precision of the estimates, point height is proportional to the number of diseased patients, while point width is proportional to the number of control patients.

Following a rigorous meta-analysis, a curve can be added in the graph. A Summary ROC curve represents the expected ROC curve at many different positivity threshold levels. If the same positivity threshold has been used in each of the primary studies, it is appropriate to calculate and plot the summary sensitivity and specificity, and their confidence region. A prediction region can also be provided, corresponding to the area where the true sensitivity/specificity of a future study should be found in 95% of the cases. Figure 3 shows a SROC curve from made using mock data in RevMan5.

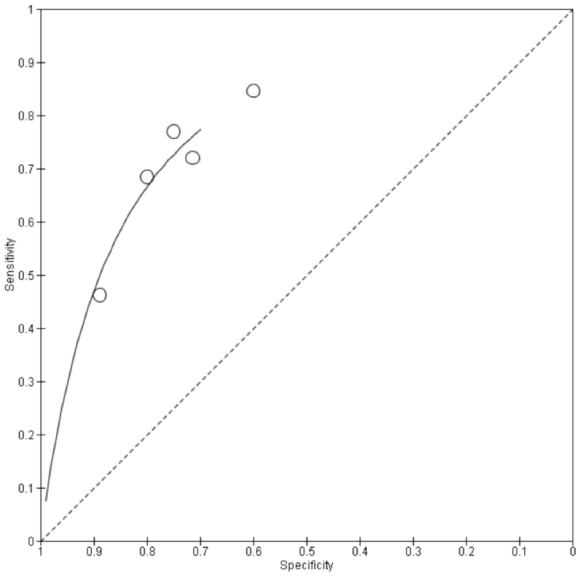


Figure 3: An SROC curve generated using mock data in RevMan5

Sensitivity is shown on the y-axis, the x-axis shows inverted specificity.[27]

Meta-analysis

Meta-analysis of data from studies of diagnostic test accuracy is more complicated than most other forms of meta-analysis (principally due to the paired nature of the main outcome measures sensitivity/specificity). As such the early involvement of a statistician is advizable.

Context

Authors of diagnostic test accuracy systematic reviews need to define the kind of meta-analysis to perform. Questions to consider are:

- Should we estimate summary sensitivity and specificity?
- Should we compute a summary ROC curve?

The answer to these questions lies in the kind of data available and more exactly whether the diagnostic threshold is the same across the selected primary studies. Sometimes retrieved studies may rigorously use the same diagnostic threshold, but, on other occasions variations in the threshold may exist. This is often the case when there is no explicit numerical cut-off-point or when the index test is based on an observer's judgment.

The basic strategy is as follows:

- When the same threshold is used through the primary studies, then:
  - estimate the summary sensitivity/specificity.
- When different thresholds are used, then:
  - produce a SROC curve; and
  - estimate the summary sensitivity/specificity for each threshold provided in the articles.

If a study has referred to sensitivity/specificity values for several threshold, it can contribute to several estimations of summary sensitivity/specificity.

Models

Methods for performing meta-analyses regarding diagnostic tests are still being debated in the literature and new statistical developments are underway.[28] Three main models exist. The first one corresponds to a fixed effect model whereas the other two are random effect models. These last two are based on a hierarchical model, taking into account the variability present within studies and between studies. Exact mathematical details for each model discussed are provided in Appendix III.

- **The Moses-Littenberg model** [29, 30] has been extensively used for meta-analyses of DTA.[31] However, it is principally a fixed effect model, whereas for many such analyses a random effect model is required. It allows the performance of SROC curves in an exploratory approach. As a fixed effect model, it does not take into account and does not consider the variability between studies.

Due to its evident simplicity (it notably does not integrate the inter-study variability), this model can, in some circumstances, produce very different SROC curves compared to the hierarchical model described below.[32] The Cochrane Collaboration recommends careful use of this model which should be limited to preliminary analyses. Confidence intervals in statistics estimates or investigations of heterogeneity should not be studied with this model. [19]

- **The Bivariate model**[22] estimates the summary parameters: sensitivity and specificity across primary studies. It is presented in the Cochrane Handbook[19] and in the article of Leeflang (2014)[4] as a method of choice.

In this method, following Chu & Cole (2006),[33] the within study variability is modelled through binomial distributions, one for sensitivity and the other for specificity. These distributions are treated jointly since estimates of sensitivity and specificity, within each study, are non-independent.

- To deal with variability in positivity cutpoint values, Rutter and Gatsonis (2001) developed the hierarchal **SROC (HSROC) model**. It produces a SROC in which each study provides only one pair of values for sensitivity and specificity. It is presented in the Cochrane Handbook and in the article by Leeflang (2014) as a method of choice to obtain SROC curves.[4, 19]

Heterogeneity

Systematic reviews of DTA frequently find heterogeneity between studies.[19] This can be due to differences in study populations, procedures followed for carrying out the test (index or reference), and the conditions or context of testing.

Additionally, heterogeneity can be the result of differences in how studies have been conducted or their data analyzed which have biased the results (for example, one study may have included all test results in the final analysis, whereas another may exclude inconclusive outcomes, thereby making the test appear more accurate than it actually is). As such the presence of heterogeneity between studies should be carefully investigated. Displaying data graphically through paired forest plots or SROC curves can help to identify the presence or absence of heterogeneity (albeit subjectively), as large differences between studies, if present, will be recognizable. If there are differences in the diagnostic threshold between studies, then paired forest plots should not be used to estimate heterogeneity as variability should exist due to the interdependence of sensitivity and specificity. In these cases heterogeneity can be estimated by judging how well studies fit with the SROC curve (and not by how scattered they are). The Chi-square test or Fisher's exact test can be used for more objective assessments of heterogeneity, however their power has been noted to be low.[34] The I2 test is not routinely used in DTA systematic reviews as it does not account for the influence that differing positivity thresholds can have.

Where heterogeneity is found, its cause should be carefully investigated by comparing the characteristics of the differing studies. If this comparison suggests that the heterogeneity is due to the existence of specific risks of bias in some studies, then meta-analysis should be restricted to studies which do not possess the identified risks (as with all systematic reviews, efforts should be made to identify potential subgroup analyzes and the intention to carry them out declared a priori in the protocol). [1] Unfortunately, subgroup analysis carries its own difficulties, as when subgroups contain a low number of studies, they are prone to heterogeneity.[1] The use of random effects models of meta-analysis (discussed above) can help to identify heterogeneity by adding a covariate into the model. The covariate, either categorical or continuous, is accordingly assumed to be the heterogeneity source. These values are not easily interpreted, however, as they show variation in parameters expressed on log odds scales.[3] When the extent of heterogeneity cannot be explained, reviewers should refrain from meta-analysis and instead conduct a narrative synthesis.

Conflicts of interest

The protocol must include a statement which either declares the absence of any conflicts of interest or which describes a specific or potential conflict of interest.

Acknowledgements

The protocol should include the acknowledgement of sources of external funding or the contribution of colleagues or institutions. If the systematic review is being performed as part of a degree award it should be noted here.

References

The protocol should include all references in full, using the Vancouver referencing style, in the order in which they appear.

Appendices

The protocol should include the critical appraisal and data extraction tools appended as appendices. These tools must match the criteria specified in the Inclusion Criteria and critical Appraisal sections. Appendices should be numbered using Roman numerals.

Searching for studies of diagnostic test accuracy

The aim of the search strategy is to generate a list of studies from the literature which is as comprehensive as possible and which may be suitable for answering the research question posed by the systematic review. The literature encompasses several types of published and unpublished material (grey literature), including journal articles, dissertations, editorials, conference proceedings and reports. Methods by which these sources can be found vary from searching electronic databases to hand searching journals and conference proceedings, checking reference lists of relevant publications, tracking citations of relevant studies and contacting experts.[1, 35]

The timeframe chosen for the search should be justified and any language restrictions stated (e.g. if only studies published in English are considered for inclusion).

The search strategy for a JBI systematic review should be conducted in three phases:

Stage 1: Identification of keywords and search terms

A limited search should be undertaken in major databases (such as MEDLINE) using the initial search terms. The aim of this stage is to locate some papers relevant to the review and determine whether those papers can provide any additional keywords, indexing terms, or subject headings that may help in the search for similar papers. This is done by analyzing words contained in the title, keywords, abstract and indexing list.

Stage 2: Conducting the search across the specified databases

The second phase is to construct database-specific searches (see Appendix IV for general and subject specific databases) for each database included in the protocol. This may involve making slight modifications in how the search terms are entered as each database may have differences in how articles are indexed and terms used to describe articles.

Stage 3: Reference list searching

The final phase of searching involves the review of the reference lists of all studies included in the systematic review for additional studies. Additionally, researchers who are experts in the field of interest may also be considered as a potential source of articles and/or unpublished data.

Unpublished data

The comprehensiveness of searching and the documentation of the databases is a core component of the credibility of a systematic review. In addition to databases of commercially published research, there are several online sources of grey or unpublished literature that should be considered. Grey or gray literature is also known as Deep or Hidden Web material and refers to papers that have not been commercially published and includes: theses and dissertations, reports, blogs, technical notes, non-independent research or other documents produced and published by government agencies, academic institutions and other groups that are not distributed or indexed by commercial publishers. Rather than compete with the published literature, grey literature has the potential to complement and communicate findings to a wider audience, as well as to reduce publication bias. However, an important thing to remember is that the group of databases should be tailored to the particular review topic.[1] Examples of sources of grey literature are included in Appendix IV.

Search strategies

Search strategies for identifying diagnostic studies should not be restricted to a particular study design and are predominantly focused on terms for the diagnostic test(s) of interest (index test) and the clinical disorder or disease stage the test is seeking to detect (target condition). If further restriction of search results is required, we recommend exploring the use of additional topic-specific terms first before a methodology search filter for diagnostic test accuracy studies is considered.[35] If methodology-specific terms are used to filter the search, examples which have been shown to have good sensitivity include false positive, false negative, sensitivity, specificity, diagnos*, detect*, accura*.[36] The terms used will need to be tailored to the database searched, and multiple terms linked with "OR" will be necessary.

Minimizing publication bias

Identifying as many relevant studies as possible and documenting the search for studies with sufficient detail so that it can be reproduced is a key feature that distinguishes a systematic review from a traditional narrative review, and should help to minimize bias and assist in achieving more reliable estimates of diagnostic accuracy. It is important to ensure that the process of identifying studies is as thorough and unbiased as possible, and to be aware of the range of potential biases which might need to be addressed through a variety of search methods. Although the importance of publication bias in diagnostic studies is not yet fully explored, recent research indicates that to achieve as comprehensive a search as possible and thereby minimizing the risk of bias, it is advisable to search several electronic databases and use other methods to retrieve studies (such as checking reference lists, citation searches, hand searching, contacting experts, etc.)[35]

A basic Boolean strategy for searching bibliographic databases is to list synonyms for each element of the PIRD and combine them using "OR" within column and "AND" between columns (Table 5).

Table 5: Structure of a logic grid

Population		Index		Reference		Diagnosis
Flu symptoms **OR** Influenza symptoms **OR** Influenza-like	**AND**	Laboratory testing **OR** PCR assay **OR** PCR test	**AND**	Viral culture **OR** Viral test **OR** Viral assay	AND	Swine flu **OR** Swine influenza **OR** H1N1

Depending on the topic area, the number of articles retrieved by such searches may be very large. Methodological filters consisting of text words and database indexing terms have been developed in the hope of improving the searches by increasing their precision when these filters are added to the search terms for the disease and diagnostic test. On the other hand, using 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 published in 2013 by Beynon et al. assessed the performance of 70 filters (reported in 19 studies) for identifying diagnostic studies in the two main bibliographic databases in health, MEDLINE and EMBASE. The results showed that search filters do not perform consistently, and should not be used as the only approach in formal searches to inform systematic reviews of diagnostic studies. None of the filters reached their minimum criteria of a sensitivity greater than 90% and a precision above 10%.[36] The findings support the current recommendation in the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy, that the combination of methodological filter search terms with terms for the index test and target condition should not be used as the only approach when conducting formal searches to inform systematic reviews of DTA.[36]

Studies on diagnostic accuracy are often based on routinely collected data rather than pre-registered trials, so publication bias may be more prevalent in diagnostic research than in therapeutic research. Searching for studies in languages other than English and for studies that are difficult to locate (grey literature), such as conference proceedings, various types of reports, ongoing studies etc., may be necessary to gain a more complete overview and to get an idea about the size and direction of any publication bias in diagnostic research.[1]

Subject-specific databases[1, 35]

In addition to MEDLINE and EMBASE, which are generally considered to be major international general healthcare databases, many countries and regions produce electronic bibliographic databases that concentrate on the literature produced in these regions, and which often include journals and other literature not indexed elsewhere. Access to many of these databases is available free of charge on the internet. Others are only available by subscription or on a 'pay-as-you-go' basis. Indexing complexity and consistency varies, as does the sophistication of the search interface, but they can be an important source of additional studies from journal articles not indexed in other international databases such as MEDLINE or EMBASE. When supplementing a search of MEDLINE and EMBASE with databases from other regions, where the prevalence of the target condition of interest in the population may be markedly different, it may be particularly important for authors of DTA reviews to consider possible sources of publication bias.

The subject-specific databases to search in addition to MEDLINE, EMBASE and Cochrane Register of Diagnostic Test Accuracy Studies, will be influenced by the topic of the review, and access to specific databases. Examples of subject-specific databases are included in Appendix IV

Dissertations and theses databases

Some studies have found that dissertations and theses are more likely to be published in full if results are positive and that, on average, dissertations that remain unpublished have lower effect sizes than published literature.[1] It is not yet known whether dissertations in diagnostic test accuracy follow a similar publication pattern, but to minimize possible effects of publication bias authors may wish to consider searching for dissertations and theses. These are not normally indexed in general bibliographic databases such as MEDLINE or EMBASE but there are exceptions, such as CINAHL which indexes nursing dissertations and PsycINFO which indexes dissertations relevant to psychology and psychiatry. Some example databases of theses are included in Appendix IV.

Grey literature databases

As discussed above, the inclusion of grey or unpublished literature is important for minimizing bias in a systematic review as grey literature has been found to be more likely to contain intervention studies reporting non-significant results than those published in healthcare journals. Examples of databases covering grey literature sources are included in Appendix IV.

The systematic review of studies of diagnostic test accuracy

This section provides information on how to synthesize and write the results of a properly carried out systematic review of diagnostic test accuracy. Additionally, it includes a brief outline of how the systematic review report should be formatted and the stylistic conventions that should be used to ensure the review meets the publication criteria of the JBI Database of Systematic Reviews and Implementation Reports (JBISRIR).

Title

The title should be clear and explicit, and reflect the core elements of the review. As per the advice for the protocol title it should state that it is on 'diagnostic test accuracy' and include the phrase '…: a systematic review' as well as make reference to each of the elements of the PIRD. Titles should not be phrased as questions or conclusions and there should be congruency between the title, review objectives/questions and inclusion criteria.

Reviewers

Each reviewer should have their affiliations listed, including affiliations with a JBI collaborating centre if applicable. An email address should be provided for the corresponding author.

Executive summary

The executive summary is a structured abstract that reflects and summarizes the main features of the systematic review. Maximum word length is 500 words, abbreviations and references should not be used.

The executive summary should include the following headings:

Background

This section should briefly describe and justify the choice of condition and tests under review, as well as provide sufficient detail to justify why the review was conducted.

Objectives/questions

The review objectives or questions should be stated in full, as detailed in the protocol section.

Inclusion criteria

Population: This section should provide the details of the population as described in the protocol

Index test: This section should provide the details of the index test as described in the protocol, including which iterations of the index test are included and how positive or negative outcomes are specified, i.e. the threshold values.

Reference test: This section should provide the details of the reference test as described in the protocol, including which iterations of the index test are included and how positive or negative outcomes were specified.

Diagnosis of interest: This section should state the disease/illness/injury/disability that is being investigated by the diagnostic test and the formal definition, if any, by which it is described.

Types of studies

Detail the study types which are eligible for inclusion in the systematic review as per the protocol – not the study types which are ultimately found and included. These will be diagnostic case-control and/or diagnostic cross-sectional.

Search strategy

Write a brief description of the systematic review's search strategy (e.g. relevant databases searched, initial search terms or keywords, and any limitations) as specified in the protocol.

Methodological quality

Describe the method or criteria that are used to appraise the included studies.

Data extraction

This section should include a brief description of the types of data extracted and the tool (as specified in the protocol) that is used.

Data synthesis

A brief description of how the data is synthesized.

Results

A brief description of the findings of the review.

Conclusions

A brief description of the conclusions of the review.

Implications for practice

A brief description of any implications that the findings may have for current practice.

Implications for research

A brief description of the implications that the review has for the direction of future research.

Summary of findings table

Systematic reviews of diagnostic studies should be accompanied by a summary of findings table, which should include the question being investigated, the index test, the reference test, the population, the estimates rate of true positives, false negatives, true negatives and false positives and the absolute difference between the index and reference tests for these values per 1000 patients, the sample size as well as the number of studies which contributed to the sample, the GRADE (Grading of Recommendations Assessment, Development and Evaluation) quality of evidence for each finding, and any comments (including decisions as to why the reviewers assigned the final GRADE ranking).[37] These Summary of Findings tables can be created using the software program Guideline Development Tool (GDT, http://www.guidelinedevelopment.org/) and should appear in the Executive Summary section in JBI systematic reviews, following Implications for Research.

To determine a GRADE ranking of the evidence, the GRADE approach begins by assigning a starting level of quality to findings. For studies of diagnostic test accuracy, cross-sectional or cohort studies are considered to provide 'high quality' evidence, whereas for other quantitative studies they are 'low'. There are two other levels in the GRADE systems, with four levels in total. These are high, moderate, low, and very low.[37-39]

Different factors are then considered that lead to downgrading the GRADE ranking. These are: Risk of bias (as determined by critical appraisal; -1 if serious risk of bias, -2 if very serious risk of bias), Inconsistency or heterogeneity of evidence (-1 if serious inconsistency, -2 if very serious inconsistency), Indirectness of evidence (-1 if serious, -2 if very serious), Imprecision of results (-1 if wide confidence interval, -2 if very wide confidence interval) and Publication bias (-1 if likely, -2 if very likely).[37-39]

For other review types there are factors which can increase the GRADE quality of evidence (i.e. large magnitude of effect, dose response, all plausible confounding factors would reduce the demonstrated effect, or create a spurious effect where results suggest no effect). However, no such factors have been endorsed for studies of diagnostic test accuracy. For further guidance on the GRADE approach visit the GRADE working group website.

Table 6: Summary of Findings template

Test result	Number of results per 1000 patients tested (95% CI)		Number of participants (Studies)	Quality of the evidence (GRADE)	Comments
	Prevalence 0%				
	[index test]	[comparator test]			
True positives (patients with [target condition])					
	TP absolute difference: 0 more				
False negatives (patients incorrectly classified as not having [target condition])					
	FN absolute difference: 0 more				
True negatives (patients without [target condition])					
	TP absolute difference: 0 more				
False positives (patients incorrectly classified as having [target condition])					
	FP absolute difference: 0 more				

JBI endorses GDT for the development of Summary of Findings tables. All Summary of Findings tables created for JBI DTA reviews must use the GDT software.

When developing a Summary of Findings table within GDT, there are different format options for exporting the table. JBI reviews must use the Summary of Findings table (layer one) option (Table 6).

Background

The background section of the systematic review report should cover all the main elements of the topic under review. The Background section prepared for the protocol generally makes a good starting point; however it will often need an extension or modification following the review. The Background should detail any definitions important to the review. The information in the Background section must be sufficient to put the inclusion criteria in context. Reasons for investigating the index test, as well as the choice of reference test should be a particular area of focus.

At the conclusion of the Background section there should be a statement that a preliminary search for previous systematic reviews on the topic has been conducted (state the databases searched, e.g. JBISRIR, Cochrane Library, CINAHL, PubMed). If a previous systematic review has been found, it should be specified how the conducted review is different from the previous one. JBI places significant emphasis on a comprehensive, clear and meaningful background section to every systematic review, particularly given the international circulation of systematic reviews, variation in local understandings of clinical practice, health service management and client or patient experiences. It is recommended that all JBI systematic reviews contain a sentence clearly indicating:

"The objectives, inclusion criteria and methods of analysis for this review were specified in advance and documented in an a priori published protocol. Ref" (the reference should be to the appropriate citation in JBISRIR).

This sentence should appear as the final line of the background/introduction section of the systematic review report and complies with the recommendations for reporting of systematic reviews detailed in the PRISMA guidelines.

Methodology

Review objectives and review questions

The review objectives should be the same as stated in the protocol (aside from tense adjustments). As discussed previously they should be followed by the specific questions.

Inclusion criteria

The inclusion criteria should be the same as described in the protocol (PIRD: population, index test, reference test, diagnosis of interest). They should be as clear and as unambiguous as possible.

Search strategy

This section should report on how the reviewers searched for relevant papers. The databases that were searched must be listed along with the search dates. This should be the same as described in the protocol. A detailed search strategy for at least one of the major databases searched should be appended to the review as an appendix. The documentation of search strategies is a key element of the scientific validity of a systematic review as it enables readers to look at and evaluate the search strategy.

Assessment of methodological quality

This section should detail the methodology followed for critical appraisal in the systematic review, including the criteria used to determine the inclusion or exclusion of studies. The process described should be the same as that specified in the protocol, with reasons for deviation given. The critical appraisal tool should be appended to the review.

Data extraction

This section should detail the types of data extracted from the included studies, which should be the same as those specified in the protocol. The data extraction tool used to facilitate this process should be appended to the review.

Data synthesis

This section details the data synthesis approach, as opposed to the results of the synthesis itself. The protocol should have specified which methods of synthesis (narrative, graphical, tabular, meta-analysis) would be considered and under which circumstances. This section should detail the actual method used along with why it has been chosen (i.e. if narrative synthesis is chosen over meta-analysis due to the presence of heterogeneity, this should be explained along with the factors that are causing the studies to be heterogeneous). If a meta-analysis is performed, the statistical software should be specified.

Results

The results section should begin with a summary of the process followed from the search to the final selection of studies for extraction and synthesis, including how many articles have been included or excluded at each stage. This should be accompanied by a flow chart conforming to the PRISMA statement (Figure 4).[40] Lists of included and excluded studies should be included as separate appendices in the systematic review report. It is important that all studies excluded at and from the 'full text review' stage should have their reason for exclusion given as a part of this list.

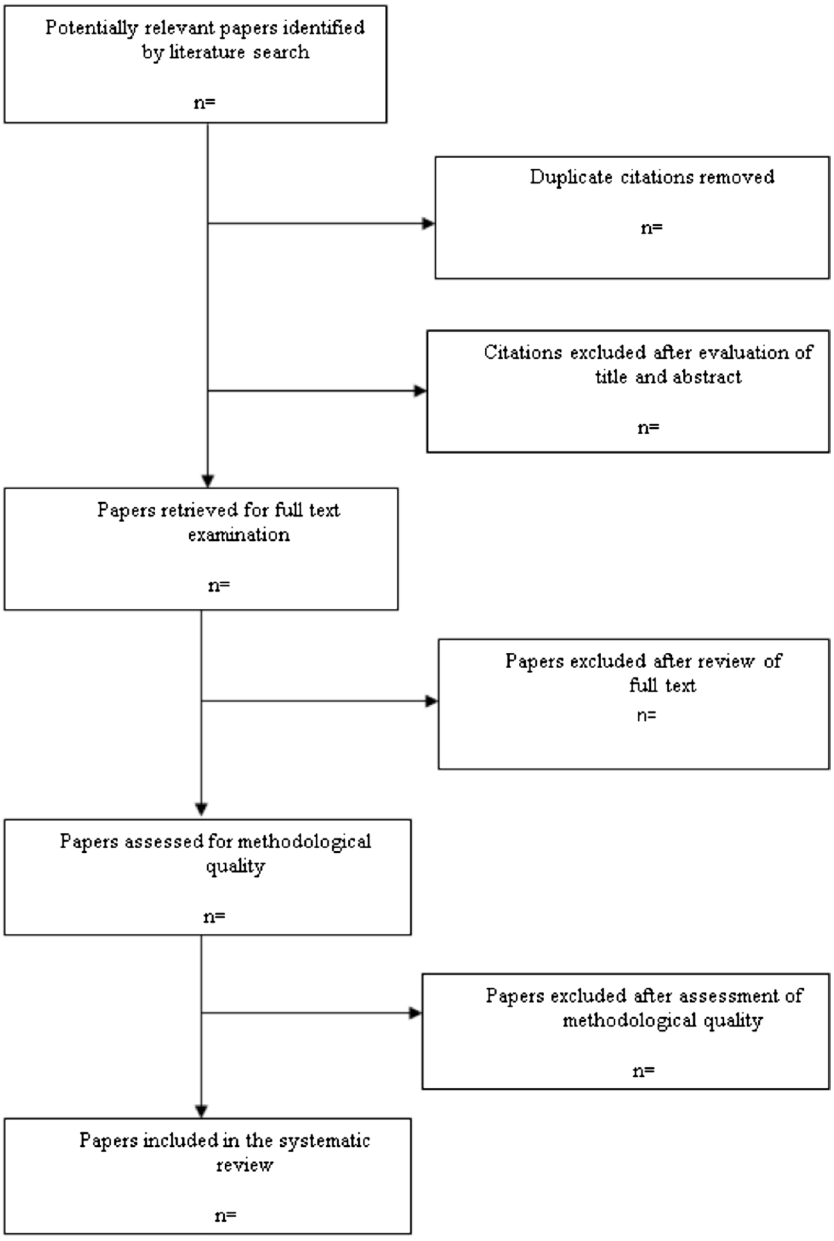


Figure 4: Flowchart detailing identification and selection of studies for inclusion in the review

Description of studies

To provide a context for the findings of the review, the results section should also include an overall description of the included studies. This should provide sufficient detail for the readers to assess how similar the studies are to one another, with a view to informing the appropriateness of meta-analysis. Specific items of interest from the studies may also be highlighted here. These may include: characteristics of the participants, the settings in which the tests have been conducted and specific study designs used. Tables are the most appropriate form for presenting this data, and the use of appendices should also be considered. The presence of extensive detail on study characteristics may obscure the actual findings, and make them less accessible to the reader.

Methodological quality

This section should detail the methodological quality of the included studies, as determined by the critical appraisal checklist used. It should include 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 (see Table 7 for example; if this table is not included in the results it should be included in the appendix). If any studies have been excluded due to critical appraisal, this is an appropriate area to provide justification.

Table 7: Critical appraisal results for included studies using the JBI critical appraisal checklist

Study	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Author(s) [ref]	Y	Y	Y	N	Y	U	Y	N	Y	U

Y - Yes, N - No, U - Unclear

Findings of the review

There is no accepted standard for the structure for reporting the findings of systematic reviews; however it is recommended that findings be presented in the same order as the relevant review questions in order to create a logical flow. Again, the use of tables and appendices should be considered in order to avoid obscuring important details with an excess of less important items. As a general rule, findings are discussed textually and then supported with meta-graphs, tables and figures as appropriate.

Discussion

The discussion section should focus on considering the results in light of the review objectives, as well as how the review findings will influence the course of diagnosis in the area of the review. Specifically, 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, should be considered.

Conclusion

The discussion should also include a final overview of the results that address any issues arising from the review's conduct, including any limitations as well as issues arising from the results of the review. Recommendations for practice and research should also be made.

Recommendations for practice

Recommendations for practice should be detailed, specific and based on documented results, not reviewer opinion. Where the results of the review do not support any specific recommendation for practice this should be noted.

Recommendations for research

Recommendations for research should be derived from the results of the review and based on identified gaps in the literature or methodological weakness. Generalized statements calling for further research should be avoided in favor of the identification of specific issues. Where the findings of a review suggest that no further research be performed (saturation may be apparent, or a test may have been identified as containing an unacceptable risk), this should be noted as a recommendation.

Conflicts of interest

A statement which either declares the absence of any conflicts of interest or which describes a specified or potential conflict of interest should be made by the reviewers in this section.

Acknowledgements

Any acknowledgements should be made in this section, e.g. sources of external funding or the contribution of colleagues or institutions. It should also be noted if the systematic review is to count towards a degree award.

References

All references should be listed in full using the Vancouver referencing style, in the order in which they appear in the review. The references should be appropriate in content and volume and include background references and studies from the initial search.

Appendices

Appendices should be numbered using Roman numerals in the order in which they are referred to in the body of the text. There are several required appendices for a JBI review:

Appendix I: Search strategy

A detailed search strategy for at least one of the major databases searched must be appended.

Appendix II: Critical appraisal instrument

The critical appraisal instrument used must be appended.

Appendix III: Data extraction template

The data extraction template used must be appended.

Appendix IV: Table of included studies

A table of included studies is crucial to allow a snapshot of the studies included in the review.

Appendix V: List of excluded studies

At a minimum, a list of studies excluded at the critical appraisal stage must be appended and reasons for exclusion be provided for each study (these reasons should relate to the methodological quality of the study, not study selection). Studies excluded following examination of the full-text may also be listed along with their reason for exclusion at that stage (i.e. a mismatch with the inclusion criteria). This may be as a separate appendix or itemized in some fashion within the one appendix.

Other appendices should be included in the order that they were referred to in the review.

Appendices

Appendix I: Critical appraisal checklist

JBI Critical Appraisal Checklist for Diagnostic Test Accuracy Studies

Reviewer_____ Date_____

Author_____ Year_____ Record Number_____

	Yes	No	Unclear	Not applicable
1. Was a consecutive or random sample of patients enrolled?	☐	☐	☐	☐
2. Was a case-control design avoided?	☐	☐	☐	☐
3. Did the study avoid inappropriate exclusions?	☐	☐	☐	☐
4. Were the index test results interpreted without knowledge of the results of the reference standard?	☐	☐	☐	☐
5. If a threshold was used, was it pre-specified?	☐	☐	☐	☐
6. Is the reference standard likely to correctly classify the target condition?	☐	☐	☐	☐
7. Were the reference standard results interpreted without knowledge of the results of the index test?	☐	☐	☐	☐
8. Was there an appropriate interval between the index test and the reference standard?	☐	☐	☐	☐
9. Did all patients receive the same reference standard?	☐	☐	☐	☐
10. Were all patients included in the analysis?	☐	☐	☐	☐

Overall appraisal: Include ☐ Exclude ☐ Seek further info ☐

Explanation of diagnostic test accuracy studies critical appraisal

Diagnostic Test Accuracy Studies Critical Appraisal Tool

Answers: Yes, No, Unclear or Not/Applicable

Patient selection

1. Was a consecutive or random sample of patients enrolled?

   Studies should state or describe their method of enrolment. If it is claimed that a random sample was chosen the method of randomization should be stated (and appropriate). It is acceptable if studies do not say 'consecutive' but instead describe consecutive enrolment; i.e. 'all patients from …. till …. were included'.

2. Was a case-control design avoided?

   Case control studies are described in detail in the reviewers manual. In essence, if a study design involves recruiting participants who are already known by other means to have the diagnosis of interest and investigating whether the test of interest correctly identifies them as such, the answer is 'No'.

3. Did the study avoid inappropriate exclusions?

   If patients are excluded for reasons that would likely influence the conduct, interpretation or results of the test, this may bias the results. Examples include: excluding patients on which the test is difficult to conduct, excluding patients with borderline results, excluding patients with clear clinical indicators of the diagnosis of interest.

Index test

4. Were the index test results interpreted without knowledge of the results of the reference standard?

   The results of the index test should be interpreted by someone who is blind to the results of the reference test. The reference test may not have been conducted at the point that the index test is carried out, if so the answer to this question will be 'Yes'. If the person who interprets the index test also interpreted the reference test then it is assumed that this question will be answered 'No' unless there are other factors in play (for instance, the interpretation of the results may be separate from their collection, in which case the interpreter may be blinded to patient identity and past reference test results).

5. If a threshold was used, was it pre-specified?

   Diagnostic thresholds may be chosen based on what gives the optimum accuracy from the data, or they may be pre-specified. When no diagnostic threshold is applied (i.e. the results of a test is based on the observation of a specific characteristic which is either there or not) this question will be answered NA.

6. Is the reference standard likely to correctly classify the target condition?

   The reference test should be the gold standard for the diagnosis of the condition of interest. Additionally, the reporting of the study should describe its conduct in sufficient detail that the reviewers can be confident that it has been correctly and competently implemented.

7. Were the reference standard results interpreted without knowledge of the results of the index test?

   The points made for criteria 4 apply equally here. The results of the reference test should be interpreted by someone who is blind to the results of the index test. The index test may not have been conducted at the point that the reference test is carried out, if so the answer to this question will be 'Yes'. If the person who interprets the reference test also interpreted the index test then it is assumed that this question will be answered 'No' unless there are other factors in play (for instance, the interpretation of the results may be separate from their collection, in which case the interpreter may be blinded to patient identity and past index test results).

8. Was there an appropriate interval between the index test and the reference standard?

   The index test and the reference test should be carried out close enough together that the status of the patient could not have meaningfully changed. The maximum acceptable time will vary based on characteristics of the population and condition of interest.

9. Did all patients receive the same reference standard?

   The reference standard by which patients are classed as having or not having the condition of interest should be the same for all patients. If the results of the index test influence how or whether the reference test is used (i.e. where an apparent false negative may be detected the study design may call for a 'double check') this may result in biased estimates of sensitivity and specificity. Additionally, in some studies two parallel reference tests may be used (on different patients) and the results then pooled. In either case the results should be 'No'.

10. Were all patients included in the analysis?

    Loses to follow up should be explained and there cause and frequency should be considered in whether they are likely to have had an effect on the results (Subjectivity may exist in this context, overall low tolerance should be applied in deciding to answer 'No' to this question, but a single withdrawal from a large cohort should not necessarily force a negative response). However, if a patients' results being difficult to interpret results in their data being excluded from the analysis this will result in an exaggerated estimate of DTA, and this question should definitely be answered 'No'.

This tool is based on and largely informed and taken from the QUADAS-2 approach.

Whiting, Penny F., et al. "QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies." Annals of internal medicine 155.8 (2011): 529-536.

Appendix II: 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 centres)	
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			

Appendix III: Meta-analysis equations and models

The Moses-Littenberg model

The models are explained below and their formulas are issued from Macaskill et al. (2010),[19] unless otherwise specified.

The method underlying the Moses-Littenberg model is based on a linear regression describing the variation of the test accuracy in function of the positivity threshold. It can be written as:

$$D = \propto + \beta S + error$$

The test accuracy is defined by the logit of the diagnostic odds ratio (D) following:

$$D = logit(sensitivity) - logit(1 - specificity)$$

The estimation of the positivity threshold (S) is:

$$S = logit(sensitivity) + logit(1 - specificity)$$

The linear regression model, describing the variation of the test accuracy in function of the positivity threshold, can be written as:

$$D = \propto + \beta S + error$$

This equation provides, through least squares method, values for α and β, which are then used to estimate sensitivity values for chosen specificities, with:

$$E(sensitivity) = 1/[1 + \exp(-\frac{[\propto + (1+\beta) logit(1 - specificity)]}{1-\beta})]$$

Usually, the chosen values of specificities are the one issued from the literature.

The bivariate model

The number of test positives in study i is defined according to:

$y_{Ai} \sim \text{Binomial}(n_{Ai}, \pi_{Ai})$ for sensitivity,

and the number testing negative following:

$y_{Bi} \sim \text{Binomial}(n_{Bi}, \pi_{Bi})$ for specificity,

with n_{Ai} and n_{Bi} respectively the number of diseased/control subjects in the study i and π_{Ai} and π_{Bi} respectively the probability of a positive/negative test in the respective group of the study i.

For the variability between studies, the logit-transformed sensitivity is treated with a normal distribution characterised by a mean μ_A and a variance σ_A^2. Similarly, the normal distribution of the logit-transformed specificity is defined by the mean μ_B and a variance σ_B^2. The correlation between these two components is integrated in a bivariate normal model, written as:

$\begin{pmatrix}\mu_{A,i}\\ \mu_{B,i}\end{pmatrix} \sim \text{Normal}\left(\begin{pmatrix}\mu_A\\ \mu_B\end{pmatrix}, \Sigma\right)$ where $\Sigma = \begin{pmatrix}\sigma_A^2 & \sigma_{AB}\\ \sigma_{AB} & \sigma_B^2\end{pmatrix}$. The term σ_{AB} expresses the covariance between logit sensitivity and specificity.

The HSROC Model

The model of Rutter and Gatsonis is based on hierarchical regression to estimate variations at the within studies level as well as at the between studies one.

At the within studies level, binomial distributions are assumed for the number of positive individuals in the diseased (y_{i1}) and control groups (y_{i2}) of study i. They are written as: $y_{ij} \sim \text{Binomial}(n_{ij}, \pi_{ij})$ with n_{ij} the sample size of tested individuals and π_{ij} the probability of a positive test. Accordingly, the probability of a positive test is determined simultaneously for diseased and control groups, following: $\text{logit}(\pi_{ij}) = (\theta_i + \alpha_i dis_{ij})\exp(-\beta dis_{ij})$

with θ_i modelling the positivity threshold, α_i the diagnostic accuracy of study i, dis_{ij} the status (diseased vs. control) for a patient of the i[th] study, and β called the scaled parameter permitting variation of accuracy with threshold. This engenders ROC curve with potential asymmetry. β is considered as a fixed effect.

The between-studies variability is treated with two normal distributions. One for the threshold, characterised by a mean Θ (capital theta) and a variance σ_β^2, and the other for diagnostic accuracy with parameters of mean Λ (capital lambda) and variance σ_α^2.

Based on the above-mentioned estimated parameters, a SROC curve can be plotted at given values of specificity.

$$sensitivity = 1/\{1 + \exp[-(\Lambda e^{-0.5\beta} + logit(1 - sensitivity)e^{-\beta})]\}$$

This equation is issued from Macaskill (2004).

As expected, when $\beta = 0$, the SROC curve will be symmetric.

Appendix IV: Examples of databases

Databases of published literature

Nursing and allied health

- *Allied and Complementary Medicine (AMED):*

(http://www.ebscohost.com/academic/AMED-The-Allied-and-Complementary-Medicine-Database)

- *British Nursing Index (BNI):*

(www.bniplus.co.uk/)

- *Cumulative Index to Nursing and Allied Health (CINAHL):*

(www.cinahl.com/)

Primary care:

- *Essential Evidence Plus (formerly Patient Oriented Evidence that Matters (InfoPOEMs)):*

(www.essentialevidenceplus.com/)

Social science, psychology and psychiatry

- *Applied Social Sciences Index and Abstracts (ASSIA):*

(http://www.proquest.com/products-services/ASSIA-Applied-Social-Sciences-Index-and-Abstracts.html)

- *PsycINFO:*

(www.apa.org/psycinfo/)

- *Sociological Abstracts:*

(http://proquest.libguides.com/SocAbs)

Biology and chemistry

- *Biological Abstracts / BIOSIS Previews:*

(http://thomsonreuters.com/biosis-previews/)

- ***Chemical Abstracts:***

(www.cas.org/)

- ***Database of the International Federation of Clinical Chemistry and Laboratory Medicine***

Committee for Evidence-based Laboratory Medicine (IFCC C-EBLM database)

(contact j.watine@ch-rodez.fr)

International health

- ***Global Health:***

Available via: (www.cabi.org)

In addition to subject-specific databases, general search engines include:

- ***Google Scholar (free on the internet):***

(scholar.google.com/advanced_scholar_search)

- ***Turning Research into Practice (TRIP) database (evidence-based healthcare resource)***

(free on the internet): (www.tripdatabase.com/)

"Citation searching"

Citation searching is an important and effective adjunct to database searching and hand searching. Information about these citation indexes is available at: **Cochrane handbook**

- ***Science Citation Index:***

scientific.thomson.com/products/sci/

- ***Social Sciences Citation Index:***

scientific.thomson.com/products/ssci/

- ***Web of Science:***

scientific.thomson.com/products/wos/

- ***Web of Knowledge:***

isiwebofknowledge.com/

- ***Scopus:***

http://www.elsevier.com/online-tools/scopus

Theses specific databases

- ***ProQuest Dissertations & Theses Database:***
www.proquest.co.uk/products_pq/descriptions/pqdt.shtml
- ***Dissertation Abstracts Online (DIALOG)***

- ***Index to Theses in Great Britain and Ireland***
www.theses.com/
- ***DissOnline: indexes 50,000 German dissertations:***
www.dissonline.de/

Grey literature databases

- ***MedNar***
mednar.com/mednar
- ***OpenSIGLE***
http://www.greynet.org/opensiglerepository.html
- ***National Technical Information Service (NTIS)***
www.ntis.gov/
- ***WorldWideScience.org***
worldwidescience.org/index
- ***Open Grey***
http://www.opengrey.eu/

References

1 White S, Schultz, T., Enuameh, YAK., ed. Synthesizing evidence of diagnostic accuracy. Philadelphia, USA: 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(7305):157-62.

3 Leeflang MM, Deeks JJ, Takwoingi Y, Macaskill P. Cochrane diagnostic test accuracy reviews. Syst Rev.. 2013;2:82.

4 Leeflang MM. Systematic reviews and meta-analyses of diagnostic test accuracy. Clin Microbiol Infect. 2014;20(2):105-13.

5 Sackett DL, Haynes RB. The architecture of diagnostic research. Bmj. 2002;324(7336):539-41.

6 Habbema J, Eijkemans R, Krijnen P, Knottnerus J. Analysis of data on the accuracy of diagnostic tests. In: Knottnerus J, Buntinx F, eds. The Evidence Base of Clinical Diagnosis: Theory and methods of diagnostic research. 2nd ed. London: BMJ Publishing Group 2009:118-45.

7 Leeflang MM, Rutjes AW, Reitsma JB, Hooft L, Bossuyt PM. Variation of a test's sensitivity and specificity with disease prevalence. CMAJ. 2013;185(11):E537-44.

8 Lau J, Ioannidis JP, Schmid CH. Quantitative synthesis in systematic reviews. Ann Intern Med. 1997;127(9):820-6.

9 Whiting P, Rutjes AW, Dinnes J, Reitsma J, Bossuyt PM, Kleijnen J. Development and validation of methods for assessing the quality of diagnostic accuracy studies. Health Technol Assess. 2004;8(25):iii, 1-234.

10 Leeflang MM, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. J Clin Epidemiol. 2009;62(1):5-12.

11 Altman DG, Bland JM. Diagnostic tests 1: Sensitivity and specificity. Bmj. 1994;308(6943):1552.

12 Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity. Continuing Education in Anaesthesia, Critical Care & Pain. 2008;8(6):221-3.

13 Mulligan EP, Harwell JL, Robertson WJ. Reliability and diagnostic accuracy of the Lachman test performed in a prone position. J Orthop Sports Phys Ther. 2011;41(10):749-57.

14 Altman DG, Bland JM. Diagnostic tests 2: Predictive values. Bmj. 1994;309(6947):102.

15 Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Stat Med. 1997;16(9):981-91.

16 McGee S. Simplifying likelihood ratios. J Gen Intern Med. 2002;17(8):647-50.

17 Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation. 2007;115(5):654-7.

18 Metz CE. Basic principles of ROC analysis. Semin Nucl Med. 1978;8(4):283-98.

19 Macaskill P, Gatsonis C, Deeks J, Harbord R, Takwoingi Y. Chapter 10: Analysing and presenting results. In: Deeks J, Bossuyt P, Gatsonis C, eds. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy: The Cochrane Collaboration 2010.

20 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29-36.

21 Erol B, Gulpinar MT, Bozdogan G et al. The cutoff level of free/total prostate specific antigen (f/t PSA) ratios in the diagnosis of prostate cancer: A validation study on a Turkish patient population in different age categories. Kaohsiung J Med Sci. 2014;30(11):545-50.

22 Reitsma J, Whiting, P., Vlassov, V., Leeflang, M. & Deeks, J. Chapter 9: Assessing methodological quality. . In: J D, ed. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. York: The Cochrane Collaboration. 2009.

23 Bossuyt P, Reitsma, J., Bruns, D., Gatsonis, C., Gatsonis, C., Irwig, L., Lijmer, J., Moher, D., Rennie, D. & De Vet, H. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. Croat Med J. 2003;138:40-4.

24 Meyer G. Guidelines for reporting information in studies of diagnostic accuracy: The STARD initiative. J Pers Assess. 2003;81:191-3.

25 Whiting P, Rutjes, AWS., Westwood,ME., Mallett, S., Deeks, JJ., Reitsma,JB., Leeflang, MMG., Sterne, JAC., Bossuyt, PMM., the QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155::529e36.

26 Gatsonis C. Do we need a checklist for reporting the results of diagnostic test evaluations? The STARD proposal. Acad Radiol. 2003;10(6):599-600.

27 Campbell JM, Kluger M, Ding S, Carmody DP, Hakonsen SJ, Jadotte YT, White S, Munn Z. Diagnostic test accuracy: Methods for Systematic Review and Meta-analysis. IJEBH. 2015.

28 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.

29 Littenberg B, Moses LE, Rabinowitz D. Estimating Diagnostic-Accuracy from Multiple Conflicting Reports - a New Meta-Analytic Method. Med Decis Making. 1990;38(2):A415-A.

30 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(14):1293-316.

31 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(4):347-75.

32 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(11):1095-103.

33 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(12):1331-2; author reply 2-3.

34 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 Technol Assess. 2005;9(12):1-113, iii.

35 de Vet H, Eisinga A, Riphagen I, Aertgeerts B, Pewsner D. Chapter 7: Searching for Studies. In: Deeks J, Bossuyt P, Gatsonis C, eds. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy: The Cochrane Collaboration 2008.

36 Beynon R, Leeflang MM, McDonald S et al. Search strategies to identify diagnostic accuracy studies in MEDLINE and EMBASE. The Cochrane database of systematic reviews. 2013;9:MR000022.

37 Schünemann H BJ, Guyatt G, Oxman A, editors. 7. The GRADE approach for diagnostic tests and strategies. In: The GRADE Working Group, ed. GRADE handbook for grading quality of evidence and strength of recommendations 2013.

38 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(7):760-8.

39 Atkins D, Best D, Briss PA et al. Grading quality of evidence and strength of recommendations. Bmj. 2004;328(7454):1490.

40 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.