Synthesizing Evidence of Diagnostic Accuracy

Sarahlouise White

Tim Schultz

Yeetey Akpe Kwesi Enuameh

Wolters Kluwer Health | Lippincott Williams & Wilkins

THE JOANNA BRIGGS INSTITUTE

AEQUAM SERVARE MENTEM

Synthesizing Evidence of Diagnostic Accuracy

Sarahlouise White
Tim Schultz
Yeetey Akpe Kwesi Enuameh

Lippincott-Joanna Briggs Institute
Synthesis Science in Healthcare Series: Book 6

Publisher: Anne Dabrow Woods, MSN, RN, CRNP, ANP-BC
Editor: Professor Alan Pearson, AM, Professor of Evidence Based Healthcare and Executive Director of the Joanna Briggs Institute; Faculty of Health Sciences at the University of Adelaide, SA 5005 Australia
Production Director: Leslie Caruso
Managing Editor, Production: Erika Fedell
Production Editor: Sarah Lemore
Creative Director: Larry Pezzato

Two Commerce Square
2001 Market Street
Philadelphia, PA 19103

ISBN 978-1-4511-6388-9

Printed in Australia

Series Editor: Professor Alan Pearson AM

This series of concise texts is designed to provide a "toolkit" on synthesizing evidence for healthcare decision-making and for translating evidence into action in both policy and practice. The series seeks to expand understandings of the basis of evidence-based healthcare and brings together an international range of contributors to describe, discuss and debate critical issues in the field.

Incredible developments have occurred in the synthesis and use of evidence in healthcare over the last several years, but the science and emerging practices that underpin evidence-based healthcare are often poorly understood by policy makers and health professionals. Several emerging and exciting developments have much to offer health professionals. Firstly, new, deeper understandings of the nature of evidence and of ways to appraise and synthesize it have led to the development of more sophisticated methodologies for synthesis science. Secondly, the realization that the rapid increase in the availability of high quality evidence has not been matched by increases in the translation of this evidence into policy and/or clinical action has spurred on developments in the science of knowledge implementation and practice improvement.

The burgeoning publications in this area – particularly books on evidence-based healthcare can go only so far in informing responsible and conscientious policy makers and health-care practitioners. This new series Lippincott/Joanna Briggs Institute, "Synthesis Science in Healthcare", is devoted to communicating these exciting new interventions to researchers, clinicians on the frontline of practice and policy makers.

The books in this series contain step-by-step detailed discussions and practical processes for assessing, pooling, disseminating and using the best available international evidence. In all healthcare systems, the growing consensus is that evidence-based practice offers the most responsible course of action for improving health outcomes. All clinicians and health scientists want to provide the best possible care for patients, families and communities. In this series, our aim is to close the evidence to action gap and make that possible.

About the Authors

Dr Sarahlouise White BSc (Hons) PhD is a Research Fellow in the Synthesis Science Unit at the Joanna Briggs Institute and has extensive experience in conducting both primary and secondary medical research. Dr White has been actively engaged in promoting evidence-based healthcare through the work of the Joanna Briggs Institute and Collaboration, in many countries.

Dr Tim Schultz is Technical Director at the Australian Patient Safety Foundation (APSF), and supervises research students enrolled in the Joanna Briggs Institute masters and PhD programs. These roles involve conducting and developing methodologies for systematic reviews, translational research, and evaluation of health services and patient safety initiatives. Dr Schultz co-ordinated the inaugural Aged Care Clinical Fellowship program run by the Joanna Briggs Institute, and continues involvement in the program as a facilitator. He has a BSc (Hons) and a PhD in comparative physiology, and a Graduate Diploma in Public Health from the University of Adelaide.

Dr Yeetey Akpe Kwesi Enuameh is the Director of the Affiliate Centre of the Joanna Briggs Institute at the Kintampo Health Research Centre in Kintampo (Ghana). His research interests include evidence based health care, health policy, infectious diseases and adolescent (sexual and reproductive) health. Dr Enuameh graduated as a Medical Doctor from the Vinnitsa State Medical University (Ukraine), has a MSc in Health Service Planning and Management and is currently a PhD candidate at Drexel University in Philadelphia (USA).

Contents

Introduction

In the provision of health care, practitioners and patients make numerous decisions and, in doing so, weigh up numerous types of information before taking action. This information includes: the results of well-designed research; information related to patients/clients and their relevant others; the practitioner's own experiences; the nature and the context (norms of the setting and culture) in which the care is being delivered. Due to the vastness of the available information and the speed at which the best available information is often required, the synthesis of the best available evidence to support decision making at the policy and practice levels has become increasingly important. The systematic review is one form of synthesized data, where the methods used to generate the data synthesis are systematic and transparent, allowing the reader to follow the key decision points of the review process.

Clinicians have long relied upon diagnostic tests for 'evidence' of the presence or absence of a disease or a condition. Similarly, policy makers must evaluate the value of a particular diagnostic test, compare it to any others, and decide which test should be made available or funded (Irwig et al., 1994). Both clinical and policy decision-makers require a thorough evaluation of the test and its ability to accurately determine who has, and who doesn't have, the disease or condition of interest. Additionally, information on the impact on healthcare providers, therapeutic impact, patient outcome and pecuniary costs and benefits of the technology should be systematically assessed (Guyatt et al., 1986). Diagnostic tests should be used only if they are cheaper, safer and at least as accurate as existing methods, if they eliminate the need for further investigations without reduction of accuracy, or if they lead to the institution of effective therapy (Guyatt et al., 1986).

Recent advances in technology and better understanding of the aetiology of disease have led to a huge expansion in the availability of new and exciting diagnostic tests that have the potential to significantly enhance patient outcomes across the world. For example, a fully automated molecular test for tuberculosis case detection and drug resistance that is suitable for developing countries was recently developed (Boehme et al., 2010). In addition to being more accurate than the standard 19th Century technique of smear microscopy, the new test reduces the time required for diagnosis from 6-8 weeks to 2 hours, allowing treatment to commence much earlier and thereby reduce mortality, secondary resistance and ongoing transmission that can occur in such immuno-compromised patients/subjects (Boehme et al., 2010, World Health Organization, 2010). Advances in the development of new diagnostic tests have been accompanied by an increased research effort to test the accuracy of new and existing techniques and technologies. For example, a MEDLINE search revealed approximately 2000 diagnostic test accuracy studies published in the period 1966-1970, which increased over eight times to about 17000 in the period 1996-2000 (Knottnerus and van Weel, 2002).

In general, assessing diagnostic test accuracy is achieved by comparing the results of using the novel (or index) test with those obtained using a reference (or standard) test on

the same population of patients or subjects. In many cases, the reference test is referred to as a 'gold standard', however, for the purposes of this text we will use the terminology 'reference test', to prevent potential misconceptions around the infallibility of such tests. Diagnostic accuracy is routinely reported using the following calculations: sensitivity – the percentage of patients with the disease that test positive with the index test and specificity – the percentage of patients without the disease that test negative with the index test. Other more detailed summary statistics from studies of diagnostic test accuracy will be discussed in chapter 3.

Meta-analysis of diagnostic test accuracy studies is required for a number of reasons. Firstly, results of diagnostic tests, such as laboratory, pathology or imaging studies, often vary greatly between different centres, patient populations and countries. There are many potential explanations for this variation – for example, sample sizes of the studies may be small, or the population sampled is not representative (Irwig et al., 1994). Additionally, diagnostic test results can differ along the spectrum of disease, or with test interpreters or the results of previous testing (Leeflang et al., 2008b). Therefore, calculation of a valid summary estimate based on all available data ,and the factors that affect the summary estimate, to inform clinical and policy decision makers is a key aim of meta-analysis (Irwig et al., 1995, Irwig et al., 1994) and a properly conducted systematic review of valid diagnostic studies sits atop the hierarchy of diagnostic evidence (Pai et al., 2004). Meta-analysis also allows for: examination of whether summary estimates are affected by study design characteristics (i.e. study quality) or characteristics of the patients/subjects or test (i.e. sub-group analysis), and identification of areas for future research (Deeks, 2001b, Irwig et al., 1994).

Methods to synthesize evidence from diagnostic test accuracy studies are now emerging and this text examines the methodological basis to the synthesis of diagnostic test accuracy data and describes the processes involved in the conduct of a diagnostic test accuracy systematic review. Although screening studies share some similarities with diagnostic studies and may report similar statistics, screening is typically applied to uncover very early signs of disease or the risk of disease, whereas diagnostic tests are generally applied to individuals with signs or symptoms of disease. Issues of meta-analysis of screening studies are discussed elsewhere (Gatsonis and Paliwal, 2006, Walter and Jadad, 1999).

Generally, systematic reviews and meta-analyses of diagnostic test accuracy studies have lagged behind reviews of effectiveness in terms of volume, methodological development, quality and uptake by clinicians (Deeks, 2001b, Irwig et al., 1995, Irwig et al., 1994, L'Abbe et al., 1987, Vamvakas, 1998, Vamvakas, 2001). As an illustration of the low volume of meta-analyses conducted, Chalmers and Lau (1993) found that nearly 9 times as many meta-analyses of RCTs ($n = 435$) than diagnostic studies ($n = 50$) were published in the medical literature over the time period 1950-1992 (Chalmers and Lau, 1993). Similarly, Irwig et al (1994) had found that in the years 1990 and 1991 only 11 meta-analyses of studies of diagnostic test accuracy were identified following an extensive search of the literature (Irwig et al., 1994). Over the period 1994-2000, a further 90 systematic reviews of diagnostic test accuracy were identified, although only two-thirds of these used meta-analysis (Honest and Khan, 2002).

The experience of the Cochrane Collaboration and the Joanna Briggs Institute is particularly insightful when considering the low profile of systematic reviews of diagnostic test accuracy.

In 2003, Cochrane decided to make provisions for including systematic reviews of diagnostic test accuracy in the Cochrane Library (Leeflang et al., 2008b). At the time of writing, there are over 4,500 Cochrane reviews in the Cochrane Library. However, this includes only three systematic reviews of diagnostic test accuracy, with the first – examining galactomannan detection for invasive aspergillosis in immunocompromized patients – published in 2008 (Brazzelli et al., 2009, Leeflang et al., 2008a, van der Windt et al., 2011). Although this number is set to grow with some 31 systematic reviews in progress as registered protocols, it is still dwarfed by the 2000 or so protocols of effectiveness reviews currently in progress (The Cochrane Collaboration, 2011). The same picture is reflected in the Joanna Briggs Institute Library of Systematic Reviews: with over 200 published systematic reviews, of which two are systematic reviews of diagnostic test accuracy (Raj and de Verteuil, 2011, White et al., 2011).

Leeflang et al (2008b) identified the slow pace of methodological development as one of two main reasons for the underuse of meta-analyses of diagnostic studies. Early work on aggregating the results of diagnostic studies lacked appropriate methodology to combine the results of different studies, and resulted in narrative summaries of results including editorial opinion pieces (see e.g. Johnson and Bungo, 1983, Thacker and Berkelman, 1986). Detrano et al's (1988) study sought to evaluate factors that affect the accuracy of exercise thallium scintigraphy for predicting angiographic coronary artery disease by reviewing relevant publications from 1977-1986 (Detrano et al., 1988). Although they presented the sensitivity and specificity for 56 individual studies and used the term 'meta-analysis' to describe their method, they calculated 'total' sensitivity and specificity by pooling the total numbers of patients across all studies. As we shall see later in this text (chapter 3), such separate pooling of sensitivity and specificity is not recommended in meta-analysis (Shapiro, 1995).

The slow pace of methodological development and the dissemination of this method to systematic reviewers is again illustrated through examining the experience of the Cochrane Collaboration and the Joanna Briggs Institute. Despite developing initial guidance for systematic reviews of diagnostic test accuracy in the late 1990s, and forming a working group to develop methodology, software, and a handbook; it was not until 2007 that the implementation of systematic reviews of diagnostic test accuracy was officially launched (Diagnostic Test Accuracy Working Group, 2011). While the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy does include an important completed chapter on meta-analysis and presenting results (Macaskill et al., 2010) at the time of writing many other chapters including interpreting results and selecting studies are still incomplete. Currently, the Joanna Briggs Institute guidance is currently underway and preliminary work has begun on computer software to the conduct of DTA systematic reviews.

Additionally, the value of many systematic reviews of diagnostic test accuracy studies is tarnished by the low quality of the included studies (Deeks, 2001b, Irwig et al., 1995, Irwig et al., 1994, Knottnerus and van Weel, 2002). In many cases, few high quality studies have been conducted, or it is not possible to properly assess the quality of included studies due to a lack of appropriate reporting (White and Schultz, 2011). A number of reviews have concluded that the conduct and reporting of primary diagnostic studies is poor and requires improvement to increase the validity of systematic reviews (Brazzelli et al., 2009, Deeks,

2001b, Irwig et al., 1994). To this end, there have been initiatives that have been developed which will be discussed in detail later in this text.

Finally, the complexity of methods required to conduct and interpret systematic reviews of diagnostic test accuracy studies (including critical appraisal and meta-analysis) is likely another reason why their uptake has been slow. Reviews may be difficult to understand for clinical researchers or may not provide enough appropriate information (Deville et al., 2002). It is hoped that this text can simplify the language and methods of systematic reviews of diagnostic studies to enhance their clinical utility and appeal to potential systematic reviewers.

SECTION 1

The Synthesis of Diagnostic Test Accuracy Evidence

Chapter 1:
The nature of diagnostic test accuracy

A diagnostic test can broadly be defined as any test that helps a clinician to identify the presence or absence of a particular condition in a presenting patient and thereby provides information with which to develop an appropriate treatment plan. The term "test" refers to any method used to obtain information on the health of a patient, ranging from history and physical examination, laboratory tests, functional tests, imaging tests and histopathology. The condition of interest or target condition can refer to a particular disease or to any other identifiable condition that may prompt further clinical action; such as further diagnostic testing, or the initiation, modification or cessation of treatment (Bossuyt et al., 2003).

Diagnostic tests are critical components of effective health care as they can allow distinction between similar conditions; however, for a diagnostic test to be useful there must be some beneficial treatment that results from detecting the problem or condition. The basic underlying assumption for utilizing diagnostic tests is that early detection leads to better care and ultimately better health outcomes for the patient (Courtney, 2004).

The Role of Diagnostic Test Accuracy Research

The world of diagnostic tests is dynamic, with new tests being developed, as well as the technology of existing tests being continuously improved. Improvements in the diagnostic test industry are driven by the demand for faster, more cost effective, less labor intensive tests– all of which without compromising patient safety or test accuracy. The ideal diagnostic test is one that aims to accurately identify patients with the target condition and exclude those without the condition within a timeframe and at a cost that allows effective decision-making (Borenstein et al., 2009, Macaskill et al., 2010, Zhuo et al., 2002). Often there are several tests for a particular condition, and with competing manufacturers claiming that their test is the best – how does a clinician chose which test to use?

Diagnostic test accuracy

Primary studies that examine the test performance (as determined by sensitivity and specificity) are referred to as diagnostic test accuracy (DTA) studies and these studies compare a "new"

test (or tests) to the best test (or method) that is currently available. The new test(s) is known as the index test and may be marketed as being more accurate, cheaper, faster or less invasive than currently available tests. The index test is compared to a test (or method) that is regarded as being the best in terms of accurately identifying the condition of interest. This test is called the reference test. The term "gold standard" is often used, however this term is misleading as it incorrectly implies that the reference test is perfect, which is unlikely to be the case (Virgili et al., 2009). The concept that no test is perfect is an important one as there are sure to be situations when an index test outperforms and subsequently replaces a reference test as diagnostic test methods and/or technology develop.

In studies of diagnostic test accuracy, the outcomes from one or more tests under evaluation are compared with outcomes from the reference test, both of which are measured in subjects who are suspected of having the condition of interest (Bossuyt et al., 2003). In this framework, the reference test is considered to be the best available method for establishing the presence or absence of the condition of interest and may consist of a single test or a combination of tests and can include laboratory tests, functional tests, imaging tests and pathology.

In the context of a DTA study, the term accuracy refers to the amount of agreement between data from the index test and the reference test. As the reference test is considered to correctly identify only condition positive patients, a positive result using this test, is considered as being a "true positive" and to occur for patients that truly have the condition of interest. The same applies with respect to a negative reference test result and patients who receive a negative reference test result are considered to be "true negatives" and to be condition-free.

The accuracy of the index test is reported relative to the reference test in terms of test sensitivity – how well does the index test correctly identify those as having the condition?, and test specificity – how well does the index test exclude those who do not have the condition of interest? Patients who receive a positive index test result but a negative reference test result are classified as being "false positive" and those who receive a negative index result and a positive reference test result are considered to be "false negative" due to the disagreement between the test results. Patient/subject classification is summarized in Table 1 below.

Table 1. Description of patient classification for diagnostic test accuracy studies

Patient classification	Description of test results
True Positive	Positive Index Test Result Positive Reference Test Result
True Negative	Negative Index Test Result Negative Reference Test Result
False Positive	Positive Index Test Result Negative Reference Test Result
False Negative	Negative Index Test Result Positive Reference Test Result

Table 2. A typical 2×2 table to classify patient test results and disease status

Test Outcome (Index test results)	Disease/condition status (Reference test results)		
	Disease/Condition positive	**Disease/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)

There is consensus (Borenstein et al., 2009, Macaskill et al., 2010, Zhuo et al., 2002) that test accuracy results should be presented as a 2 × 2 table, such as Table 2 which is adapted from Macaskill et al. (Macaskill et al., 2010). A patient (or subject) should appear in only 1 cell of the table. Once the table has been populated, the figures can then be used to assess the performance of the index test, using measures such as: sensitivity, specificity, likelihood ratios, and predictive values.

In order to generate a 2 × 2 table, the data should be dichotomized into positive and negative using a predetermined cut-off value, for each test. The value at which a test result becomes positive will have obvious implications for the number of patients in each category (the sensitivity and specificity for that test) so should it be chosen with caution. This is an important concept and is explored further in chapter 3. A cut-off value that is inappropriately high will wrongly exclude some patients that have the disease/condition and therefore falsely decrease test sensitivity. A cut-off value that is inappropriately low will wrongly classify some patients that do not have the disease/condition and therefore falsely decrease test specificity. In situations where there could be several logical cut-off points, it may be sensible to calculate the sensitivity and specificity at each of those thresholds and present them graphically as a receiver operator curve (ROC) to allow examination of how changing the threshold alters the findings (Borenstein et al., 2009, Zhuo et al., 2002).

The sensitivity and specificity of a diagnostic test can be calculated, using the patient/subject classifications from Table 2, using in the equations (Macaskill et al., 2010)

$$Sensitivity = \frac{a}{a+c}$$

$$Specificity = \frac{d}{b+d}$$

Once the test sensitivity and specificity have been calculated, these measures of test accuracy can be reported in a number of ways that convey relevant and practical information to the clinician about what the test result is likely to mean for the patient. Frequently used measures are predictive values (positive and negative) and likelihood ratios (positive and negative).

Positive and negative predictive values (PPV and NPV) are used to assess the usefulness of a result, once the test result is known. For example, a PPV of 95% indicates that 95% of patients who received a positive test result actually had the condition or disease of interest. A NPV of 60% indicates that 60% of patients that received a negative test result were truly disease/condition free. The equations used to calculate PPV and NPV are below.

Positive Predictive Value

$$PPV = \frac{TP}{TP + FP}$$

Or, by using the patient classifications from Table 2 this can be written as:

$$PPV = \frac{a}{a + b}$$

Negative Predictive Value

$$NPV = \frac{TN}{TN + FN}$$

Or, by using the patient classifications from Table 2 this can be written as:

$$NPV = \frac{d}{c + d}$$

A ratio of test sensitivity and specificity is often reported as a likelihood ratio (LR). The likelihood ratio describes the probability or likelihood that the test result would be expected in a person with the condition, compared with the probability or likelihood that the same result would be expected in a person without the condition (Deeks, 2001a). Likelihood ratios can be both positive and negative values.

A positive likelihood ratio (+ve LR) can be expressed as; people with the condition are X times more likely to receive a positive test result than those who are well (Grimes and Schulz, 2005).

A negative likelihood ratio (−ve LR) can be expressed as; people with the condition are X times more likely to receive a negative test result than those who are well (Grimes and Schulz, 2005).

Mathematically this can be expressed by the equations:

Positive Likelihood Ratio

$$+ve\ LR = \frac{Sensitivity}{1 - specificity}$$

Or, by using the patient classifications from Table 2 this can be estimated as:

$$+ve\ LR = \frac{a/(a + c)}{b/(b + d)}$$

Negative Likelihood Ratio

$$-ve\ LR = \frac{1 - sensitivity}{specificity}$$

Or, by using the patient classifications from Table 2 this can be estimated as:

$$-ve\ LR = \frac{c/(a+c)}{d/(b+d)}$$

Diagnostic Test Accuracy Evidence and Healthcare

The evidence base for diagnostic test accuracy is growing increasingly being used in healthcare to compare the performance of different tests in the diagnosis of specific conditions, where several tests exist for a condition. Over the past decade, estimates of publications related to diagnostic test accuracy have greatly increased. A cursory search of PubMed (June 2011) for "diagnostic test accuracy" revealed that between 2001 and 2011 over 5000 (5612) publications. Less than half of this total (1718) was available between 2001 and 2005, implying that most of the publications occurred in the later part of the decade. Such increases in diagnostic tests accuracy studies portray the quest to make available to the clinician the appropriate tools to serve the patient/subject, who in turn will be assured of receiving prompt quality treatment.

Direct comparison of diagnostic tests provides crucial information about the accuracy of the test in terms of correctly identifying "condition positive" patients/subjects (sensitivity) and correctly identifying "condition negative" patients/subjects (specificity). Correctly determining how well the test performs at identifying whether the patient has the condition of interest or not is important for clinicians to know as exaggerated and/or biased results from poorly designed or reported DTA studies could lead to a premature uptake of a test and lead to clinicians making incorrect diagnoses, followed by subsequent mistreatment (Bossuyt et al., 2003).

A rigorous evaluation of diagnostic tests before their introduction into clinical practice would not only reduce the number of unwanted clinical consequences (such as incorrect diagnoses and subsequent mistreatment), related to misleading estimates of test accuracy, but also reduce healthcare costs by preventing unnecessary testing. Studies to determine the diagnostic accuracy of a test are a vital part of this evaluation process (Bossuyt et al., 2003).

The drives for improvements in the diagnostic test industry include:

Faster Results

The need for a test to be able to provide results in a short timeframe to enable the clinician to confer a diagnosis so that a patient may begin a course of treatment more quickly, increasing the likelihood of positive patient outcomes. A short time between conduct of a diagnostic test and availability of the results is also important in situations where there may also be public health concerns, such as pandemic conditions.

Cheaper Tests

Finding less expensive methods to enable a clinician to diagnose a patient is beneficial for healthcare budgets in general – as long as test accuracy is not compromised. This becomes especially important in regions that are resource poor (such as developing countries) or in situations where many patients require testing (such as during a pandemic).

Easier Tests

There are less likely to be operational/technical errors with tests that are simple and easy to use. Ease of use is of benefit not only to those conducting and interpreting results of a diagnostic test, but also increases patient safety and confidence.

All of these drives seek to improve diagnostic tests without decreasing the accuracy or reliability of the results. Most diagnostic accuracy studies are, in principle, cross-sectional studies. Delayed- cross-sectional studies are those in which results of indicator tests are decided sometime after conclusions are made on the condition of the subject.

Two methods are used to recruit participants for a diagnostic accuracy study: (Deville et al., 2002)

a. Using a single set of inclusion criteria such as a specific diagnosis 'cohort type accuracy studies' or 'single-gate' studies.
b. The use of different sets of criteria for those with and those without the target condition 'case-control type accuracy studies' or 'two-gate' studies (Deville et al., 2002).

Chapter 2:
The Synthesis of Diagnostic Test Accuracy Evidence

Systematic reviews of diagnostic test accuracy (DTA) evidence offer the opportunity to generate a summary estimate of the sensitivity and specificity of a particular test, in comparison to a reference standard. Determining the overall accuracy of a diagnostic test by summarizing primary diagnostic studies can be complex and problematic. Several of the problems faced by review authors relate to methodology and reporting of diagnostic studies. In an attempt to improve the scientific rigor and completeness of reporting, The Cochrane Collaboration established the Standards for the Reporting of Diagnostic Accuracy (STARD) initiative (Bossuyt et al., 2003, Meyer, 2003) as a way of assessing study quality. Complete and accurate reporting of how the study was conducted allows the reader to detect the potential for bias within the study (internal validity) and the generalisability and applicability of the results (external validity) (Bossuyt et al., 2003).

The outcome of the STARD initiative was a 25-item checklist to aid authors of diagnostic studies when reporting study data. The checklist can also act as a useful source of data when assessing the methodological quality of this type of primary research. Other issues surrounding the assessment of diagnostic tests include the lack of studies that evaluate clinical outcomes of the test (i.e. what happens to the patient once the test results are known?), as well as a lack of data on use of the test in realistic clinical settings (i.e. not just the manufacturer's laboratories). The field of study that aims to assess methodological quality of diagnostic tests is a relatively new one and as such, is less developed than that which examines effectiveness of therapeutic interventions. There is also evidence to suggest (probably again because of its relative newness), that the quality of studies that report diagnostic test accuracy is not at the same standard as that of effectiveness studies (Leeflang et al., 2007). The statistical methods used to combine data from diagnostic test accuracy studies are also more complex than those used to examine the effectiveness of interventions, as both sensitivity and specificity are to be considered (Tatsioni et al., 2005).

Systematic Review and Meta-Analysis

Summarizing the findings of several studies on the same topic, using similar measures in a knowledge synthesis or systematic review is not a new concept. In the 1960's reviews became popular in many areas of research such as psychology, education and the social sciences (Chalmers et al., 2002). With the recognition of the importance of evidence-based decisions in healthcare, the review or knowledge synthesis has grown in popularity with clinicians and researchers alike as a way of summarizing salient information. There are many formats that a review of currently available literature may take and some of those most often used in healthcare are discussed in a recent paper (Grant and Booth, 2009).

Traditionally, summarizing a field of research to address whether a treatment was effective or not, was the domain of experts in that field who would read studies that addressed a particular

question, summarize the findings of those studies and then arrive at a conclusion regarding effectiveness (Chalmers et al., 2002, Leeflang et al., 2008b). This type of review is often called a narrative or a literature review and often provides little detail on how the information was gathered, assessed and summarized (Grant and Booth, 2009). In terms of transparency of process, a substantial improvement in the area of generating reviews, was the introduction of the systematic review. Since the 1990's the systematic review has become more popular (as determined by the increased numbers indexed by commercial databases) in healthcare research for its systematic approach toward literature searching, critical appraisal of studies to be included in the review, data extracted from included studies and methods of data synthesis, (Chalmers et al., 2002, Borenstein et al., 2009) together with its overall transparency. Due to the rigorous way in which systematic reviews are conducted and reported, they provide a solid evidence base on which to make healthcare decisions – attempting to bridge the gap between research and decision-making (Tricco et al., 2010).

The systematic and transparent approach to review writing offers advantages over the traditional approach, such as decreased subjectivity and transparency of how the conclusions were reached (Borenstein et al., 2009). In narrative reviews for example, different review authors may use different criteria about which studies to include in the review. Once included studies have been decided upon, different authors may place different credence or value upon different studies (based on items such sample size or method used). The process of how data is combined and how conclusions are drawn is most often not clearly explained in narrative reviews. With the systematic review approach, the criteria for which studies are to be included and excluded are clearly defined in advance. The data to be extracted and how that data will be analyzed are also clearly detailed in advance. These are clearly major advantages over a traditional literature, providing greater transparency and allowing a logical progression from the development of the review question to the synthesis of results. A systematic review may or may not include statistical combination (meta-analysis) of data from primary studies, depending on the data.

The term meta-analysis was first credited to Gene Glass in 1976 (Glass, 1976) and refers to systematic reviews where data is combined according to specific statistical rules and where relative weight or importance of a study occurs based on mathematical criteria, specified in advance (Borenstein et al., 2009). A meta-analysis allows calculation of a summary estimate or summary of the findings from all of the included studies. This summary provides a more precise estimate of the benefits or harms of the intervention for a given population.

For data to be combined in meta-analysis, certain criteria need to be fulfilled. The first is that the data is sensible to combine in terms of where it came from (e.g. similar participant population, settings and study design) and how it was collected (e.g. comparable methods and outcome measures). The second criterion is that is it useful to do so. Even if the study designs and patients are comparable, there may be no benefit of combining data from primary studies were the data gives results in opposing directions (Hatala et al., 2005).

To satisfy whether it is sensible to combine data statistically, clinical, methodological and statistical homogeneity should be demonstrated as defined below.

- Clinical Homogeneity – Are the participants of the included studies similar in terms of age, co-morbidities, disease state and medications? If so, then it may be appropriate to make a summary estimate and generalizations for this population.

- Methodological Homogeneity – Do the included studies utilize similar study designs and methods? Only studies that use similar methodologies can be combined as how the study is conducted determines how, when and what data is collected.
- Statistical Homogeneity – Do the included studies measure the same outcome measures, using similar scales? Statistical tests such as Chi square and I^2 can be used to determine the degree to which numerical data is homogenous.

Chapter 3:
Systematic reviews of diagnostic study data

As discussed previously, assessing the diagnostic accuracy of a test is complex as both sensitivity and specificity need to be considered when generating a summary estimate of diagnostic test accuracy. As a result, comparing and combining such studies in meta-analysis presents its own series of challenges, which have been topics of previous discussion (Hasselblad and Hedges, 1995, Leeflang et al., 2008b, Lijmer et al., 2002, Mallett et al., 2006, Tatsioni et al., 2005) and although there has been progress in some areas, several issues remain as discussed in the following sections.

The main objective of a DTA systematic review is to summarize the evidence on the accuracy of a test in comparison to an appropriate reference standard. Accuracy in this context is measured by sensitivity and specificity. Other descriptive statistics such as likelihood ratios and predictive values are also reported. Other objectives of diagnostic test accuracy systematic reviews are to critically evaluate the quality of primary diagnostic studies, check for heterogeneity in results across diagnostic studies and to explore sources of that heterogeneity.

A synthesis of data from well reported, high quality studies is an appropriate way of comparing the performance of individual diagnostic tests and can provide a useful way of drawing together studies that utilize the same test on a similar patient population. Zhuo et al (2002) summarize some uses of DTA systematic review data:

- Identify the number, quality and scope of primary DTA studies
- Summarize DTA over reported studies
- Determine whether there is heterogeneity among the accuracy values across studies
- Examine the relationship of DTA to study quality, patient characteristics and test characteristics
- Compare the DTA of related tests, increasing power to detect differences in accuracy over individual comparative studies
- Examine the relationship between test comparisons and study quality, patient characteristics and test characteristics
- Provide directions for future research

Challenges of undertaking systematic reviews of diagnostic test accuracy

Some challenges encountered in undertaking systematic reviews of diagnostic test accuracy data are discussed in the following sections:

Searching for and identifying relevant primary research

In areas of research where they are consistently applied, easily identifiable terms – such as randomized controlled trial, specific search filers for electronic databases have proved to be

useful at reducing the number of papers to screen, whilst not losing relevant papers (Leeflang et al., 2008b, Leeflang et al., 2006). Usefulness of such filters relies on the correct indexing of terms related to methodology and text words used in reporting the results. An extensive literature is available on designing search strategies to identify therapeutic intervention studies, however the corresponding body of literature on diagnostic test search strategies, is currently relatively small, therefore the recommendation is not to use search filters that are based on single search terms alone method (Gatsonis and Paliwal, 2006, Leeflang et al., 2006, Walter and Jadad, 1999).

Identifying primary research that reports test performance is one challenge that authors face when assessing diagnostic test accuracy (Garg et al., 2009, Haynes and Wilczynski, 2004, Ritchie et al., 2007). There is no unequivocal keyword or indexing term for an accuracy study in the literature databases. Use of the Medical Subject Heading (MeSH) "sensitivity and specificity" may look suitable but it is inconsistently applied by the major electronic databases (Garg et al., 2009, Haynes and Wilczynski, 2004, Leeflang et al., 2008b, Ritchie et al., 2007).

Recent work on search filters has shown that although indexing of diagnostic accuracy studies has improved in Medline, there are still no consistently applied search terms that will identify as many studies as a comparable unfiltered search (Haynes and Wilczynski, 2004). Until indexing systems properly code studies of diagnostic test accuracy, searching them will remain challenging and may require additional manual searches, such as screening reference lists, as filters often miss relevant articles and are unlikely to reduce the number of articles to be screened (Leeflang et al., 2008b, Mallett et al., 2006).

An additional factor that may also hamper study identification is that data on the diagnostic accuracy of a test may be hidden within studies that did not have test accuracy as their primary objective (Leeflang et al., 2008b, Leeflang et al., 2006).

Hand searching through publications, reference checking, and searching for unpublished reports, also helps, especially to assess the extent of publication bias. Finally, it is important to document search activity and the outcome of each search, in order to maintain transparency – a crucial feature in the credibility of the systematic review process (Gatsonis and Paliwal, 2006, Leeflang et al., 2006, Walter and Jadad, 1999). The overall consensus is that a search for diagnostic studies must be comprehensive, objective, and reproducible (Whiting et al., 2008b). Where possible, the search should cover not only journals but include other publications, such as conference proceedings and other Grey literature to ensure the maximum breadth and odds of identifying relevant studies.

Variations in study populations

Diagnostic tests performed in study populations located in different settings and with different characteristics could provide results that differ from each other by virtue of the differences inherent in those populations (Tatsioni et al., 2005). For example, the clinical manifestations of malaria (fever, lassitude/lethargy) among non-immune persons (for example children or persons not resident in a malaria endemic region) may be far more severe (and therefore perhaps more readily detectable) in comparison to those with immunity, such as adults residing

in a malaria endemic region (Castelli et al., 1999). As such when studying manifestations of clinical malaria, it would be appropriate to compare outcomes among similar populations of immune and less immune subjects in their defined settings, enhancing the homogeneity of outcomes whilst reducing the level and influence of heterogeneity. This would ultimately facilitate comparison and interpretation of studies (Tatsioni et al., 2005).

Publication bias

It has been well documented that studies with significant results are more likely to be reported than those with non-significant findings, this is known as publication bias (Mower, 1999). One way of assessing the size of publication bias within a research area is to compare the total number of studies undertaken (as determined by the number of studies approved by ethics review committees and research boards), with the number of studies that publish results. A second is to search for studies that are published in non-commercial journals and other forms of Grey Literature.

Diagnostic studies with poor test performance results that are not published may lead to exaggerated estimates of the true sensitivity and specificity of a test in a systematic review (Tatsioni et al., 2005). Attempts at addressing this publication bias has been made for randomized controlled trials, and several visual and statistical methods have been proposed to detect and correct for unpublished studies (Tatsioni et al., 2005). One solution to the problem of publication bias is the mandatory registration of all clinical trials before patient enrolment and for therapeutic trials; considerable progress has already been made in this area. Such a clinical trials registry has been suggested for clinical outcomes of diagnostic tests (Tatsioni et al., 2005) however this would prove to be problematic for diagnostic studies, as often there is no ethical review or study registration; therefore tracking of studies from registration to publication status is not currently possible (Leeflang et al., 2008b). DTA studies are often carried out on patient specimens that have served their purpose and are no longer required; therefore an ethical review is unnecessary.

Approaches to identify and minimize publication bias could include the following (Parekh-Bhurke et al., 2011, Whiting et al., 2008b):

1. Preventing publication bias before a systematic review (e.g. prospective registration of trials)
2. Reducing publication bias in the course of a systematic review (e.g. comprehensive search for primary studies from diverse sources in diverse languages, hand searching of journals, contacting experts and checking reference lists of relevant publications searching for unpublished or grey literature)
3. Detecting publication bias in the course of a systematic review (e.g. funnel plots, sensitivity analysis modeling)
4. Minimizing the impact of publication bias after a systematic review (confirmatory large scale trials and updating the systematic review)

Decisions on publication bias in DTAs should not be based on funnel plot based tests in view of the fact that they may provide misleading results and alternatives created for DTAs end up being poorly powered (Deeks, 2001b, Leeflang et al., 2008b).

Assessing methodological quality of diagnostic studies

The quality of a 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 reliability of a systematic review is greatly enhanced by the use of studies of the highest possible quality (Deeks, 2001b, Reitsma et al., 2009).

The terms "Assessing methodological quality", "assessing study validity" and "assessment of risk of bias" are synonymous and describe techniques 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).

The quote "the extent to which all aspects of a study's design and conduct can be shown to protect against systematic bias, non-systematic bias that may arise in poorly performed studies, and inferential error" (Lohr, 1999, Tatsioni et al., 2005) aptly describes factors contributing to the methodological quality of a study. Biases in study outcomes may be in the form of small study sample sizes that do not appropriately represent the population being studied, non-random sampling of the study sample and observers not blinded to the outcomes of the study (Tatsioni et al., 2005). The standards for reporting of diagnostic accuracy (STARD) 25-item checklist and the quality assessment of diagnostic accuracy studies (QUADAS) 14-item tool have been devised to improve on the quality of reporting of primary diagnostic studies and the assessment of primary DTA studies included in diagnostic systematic reviews respectively (Whiting, Harbord et al. 2008).

Diagnostic test accuracy studies with design deficiencies can produce biased results, leading to misinformation regarding the accuracy of the test under evaluation. Other factors resulting in bias within diagnostic test accuracy studies include differences in study populations and small sample sizes (Leeflang et al., 2008b). Table 3 is modified and expanded from (Leeflang et al., 2008b) and presents the major types of bias that can occur in studies and result from flawed or incomplete DTA reporting.

Attempts have been made to improve methodological quality and diagnostic test study reporting as mentioned previously. The development of tools to aid primary researchers address and avoid sources of bias such as those listed in the table, as it is only with complete and accurate reporting that research becomes transparent and its internal and external validity can be assessed. Current approaches to quality assessment of diagnostic studies include using a validated checklist instrument. This approach allows an author to work through a pre-defined list of items, with the overall aim of minimizing bias and improving the quality of their report. The checklist published by the STARD initiative (Bossuyt et al., 2003, Meyer, 2003) is one such instrument aimed at primary researchers.

For those interested in comparing findings from individual DTA studies (such as systematic review authors), assessing the quality of a diagnostic study requires consideration of several features of both study design and conduct, including factors such as definition of the research question and clinical context, specification of appropriate patient population, description of the diagnostic techniques under study and their interpretation, detailed accounting of how the reference standard information was defined and obtained, and any other factors that can affect the integrity of the study and the generalizability of the results (Gatsonis and Paliwal, 2006, Leeflang et al., 2006, Walter and Jadad, 1999).

Table 3. Majors types of bias that result from incomplete reporting in diagnostic test accuracy studies

Type of bias	When does it occur?	How does bias affect the diagnostic test accuracy assessment?	Steps that can be taken to address or prevent type of bias
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 enroll 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

(*continued*)

Table 3. Majors types of bias that result from incomplete reporting in diagnostic test accuracy studies (*Continued*)

Type of bias	When does it occur?	How does bias affect the diagnostic test accuracy assessment?	Steps that can be taken to address or prevent type of bias
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	When the patients' condition changes between administering the index and reference test	Under – or overestimation of accuracy, depending on the change in the patients' condition	Perform the reference and index with minimal delay. 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

A checklist approach can prove convenient for those comparing DTA studies, as the included studies are compared against a checklist of predefined items, noting their presence or absence, as a way of assessing key qualities. A significant development in this area was the quality assessment tool for diagnostic accuracy (QUADAS) (Whiting et al., 2003, Whiting et al., 2006). The QUADAS tool (Appendix I) is a rigorously constructed checklist of 14 items that can be used by investigators undertaking reviews of diagnostic test accuracy.

Assessing study quality using the QUADAS tool

Information regarding the methodological quality of a study may be collected in 5 stages (Reitsma et al., 2009):

1. Choosing the quality items to be assessed and developing guidance for the assessment based on the requirements of the review
2. Development of an assessment form
3. Testing and refinement the assessment form
4. Collection of data from the identified studies
5. Resolution of any disagreements or to seek clarification for data that is ambiguous or missing

Assessing the methodological quality of a study is usually performed by two independent reviewers and the process of resolving disagreements should be clearly defined in the protocol of the review, as discussed in chapter 4.

The response options to each QUADAS checklist item should be "Yes", "No" or "Unclear". "Yes" implying that the methodological feature is optimal; "No" meaning that the methodological feature is less than optimal with the potential of introducing bias or limiting its applicability. Other checklist items may be added to assess methodological quality, examples are shown in (Appendix II) depending on the nature of the review in question (Reitsma et al., 2009).

The use of a checklist allows an overall view of study quality to be gained and a quality score to be generated. This assessment can then be used to rank or exclude the study from the review. It was however observed that the assessment process of generating scores produced different rankings and different conclusions by different groupings. As such it is advised to use quality assessment as a means of investigating associations between the individual quality items of the QUADAS tool and estimates of diagnostic accuracy instead of arriving at quality decisions based on a summary quality score (Whiting et al., 2008b).

Presenting information on study quality

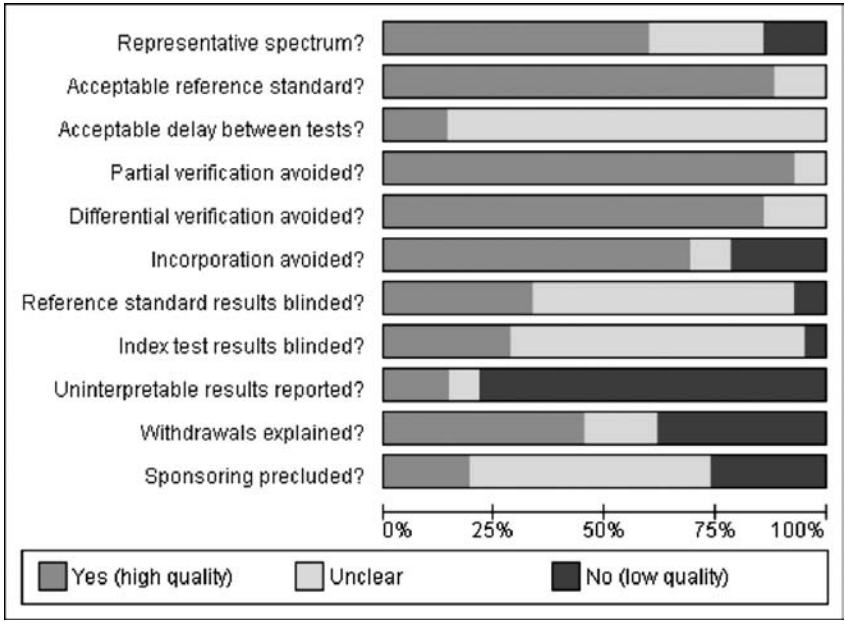
A summary graphic may be helpful to convey the methodological quality of each study. Below are two examples of how published DTA systematic reviews have graphically summarized the methodology quality of the included studies according to responses to the QUADAS checklist criteria.

Figure 1 presents each quality item on the QUADAS list has an overall value as interpreted from the study by the reviewer (Leeflang et al., 2008b, Reitsma et al., 2009).

Figure 2 illustrates the second example is more detailed and breaks down the individual areas of bias for each study.

The first example uses a figure to present the results as stacked bars representing each QUADAS item and percentages allotted on each item for "Yes", "No" and "Unclear" as assessed by the reviewers. Each study included in the review could also be assessed independently on each QUADAS item and presented in the form of a table as shown in the second example (Leeflang et al., 2008b, Reitsma et al., 2009). It is interesting to note that the literature regarding quality assessment of diagnostic studies is still relatively undeveloped, at

Figure 1: Each quality item on the QUADAS list has an overall value as interpreted from the study by the reviewer (Leeflang et al., 2008b, Reitsma et al., 2009)



least in comparison with therapeutic intervention studies (Leeflang et al., 2008b, Lijmer et al., 2002).

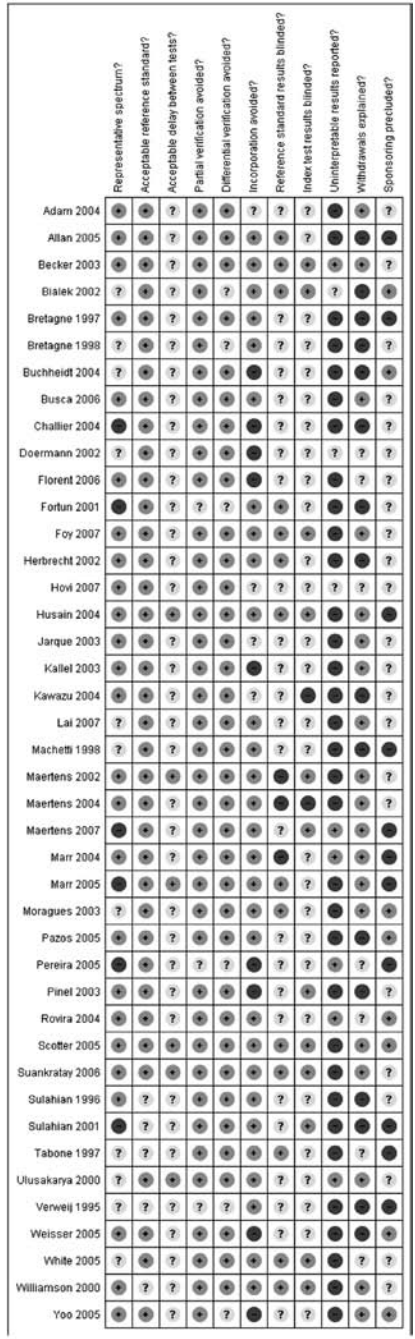
Incorporation of quality assessment results into systematic reviews is a matter of debate (Gatsonis and Paliwal, 2006, Leeflang et al., 2008b, Mallett et al., 2006) and will depend on the purpose of the systematic review and motivations of the authors. A simple and restrictive approach is to exclude studies with a poor quality score. This approach has the potential to yield few studies for inclusion in the systematic review (Borenstein et al., 2009).

One alternative is to rate studies according to their quality score and weight them in statistical analysis, however tests used with this method have been criticized as the ". . . statistical rationale for their use [being] shaky" (Gatsonis and Paliwal, 2006). The results of quality appraisal can be summarized to offer a general impression of the validity of the available evidence. Another problem with an overall quality score is that different items may generate different magnitudes of bias, even in opposing directions, making it very hard to attach sensible weights to each quality item (Leeflang et al., 2008b).

Another (perhaps more conservative) way of presenting the quality assessment is by tabulating the results of the individual checklist items for each study, allowing transparency and ease of comparison so that a reader can see all the information on which conclusions are based.

Still another proposed alternative is the conduct of sensitivity analysis, in which possible reasons why the study was poor quality and what overall effect that this made to the meta-analysis (Borenstein et al., 2009, Gatsonis and Paliwal, 2006, Zhuo et al., 2002). In the analysis phase, the results of the quality appraisal may guide explorations into those sources by using

Figure 2: Each quality item on the QUADAS list has an overall value as presented by the reviewer (Leeflang et al., 2008b, Reitsma et al., 2009)



methods such as subgroup analysis, or meta-regression analysis. In reality though, the number of included studies may be too small for meaningful for such investigations. Also, incomplete reporting by the researchers may hinder evaluation of study quality (Tatsioni et al., 2005).

In all, regardless of which approach is taken in assessing the quality of studies involved in a systematic review, there is the need for reviewers to transparently report in as much detail (to enable replication of the process) the exact approaches used. This will enable the consumers of the review to decide based on information provided on its relevance or otherwise to their specific needs.

Extracting and combining data from diagnostic test accuracy studies

A further challenge that faces authors of DTA systematic reviews is how to handle and sensibly combine the data from studies included in the review (Gatsonis and Paliwal, 2006, Hasselblad and Hedges, 1995, Leeflang et al., 2008b). Systematic reviews of diagnostic test accuracy are concerned with test results that can be presented in different formats (Borenstein et al., 2009, Macaskill et al., 2010, Zhuo et al., 2002) as summarized in Table 4.

Ideally, extracting data from primary research publications should involve two reviewers working independently. The reviewers should extract information of relevance to the review in addition to the actual test performance data – such as the type of participants involved in the study, and testing procedures (index and reference tests). Information on cut-off points for dichotomous outcomes as well as the ratings of ordinal outcomes. Consistency of data extraction can be optimized with the help of standardized data extraction forms that facilitate uniformity of information gathering. However when the test results are presented, it is important for their interpretation, that what constitutes test positive and test negative results are clearly defined.

Diagnostic test results are often defined on a continuous scale. On most occasions, a threshold/cutoff is defined below which the test result could be negative or above which it could be positive. With such a cutoff, results of a diagnostic test could be placed in a 2×2 table with the test result (in the left column) and the health condition (on the top row) being assessed as true positives, false positives, true negatives and false positives (Tatsioni et al., 2005). As has been described earlier in chapter 3, a change of the cutoff point would result in changes of both the sensitivity and specificity of the test. The challenge in defining a single figure to

Table 4. Typical formats for outcome data in DTA studies

Type of Outcome	Description
Dichotomous	Test results are reported as yes/no or positive/negative
Continuous	Test results are reported on a continuous scale or as a count (e.g. a concentration of a substance, number of features, mg, kg etc.)
Ordinal	Test results are reported as a set of numbered categories (e.g. 1 = definitely normal, 2 = probably normal, 3 = equivocal, 4 = probably abnormal, 5 = definitely abnormal

appropriately describe a diagnostic test result can be attributed to the fact that study outcomes are reported as two or more measures as with for example sensitivity and specificity, (Tatsioni et al., 2005, Leeflang et al., 2008b) positive and negative predictive values, likelihood ratios for the test results and receiver operator characteristic (ROC) curves.

Synthesizing results of diagnostic tests accuracy studies involves a series of approaches. When putting together several DTA outcomes, the first approach should include plotting sensitivity-specificity pairs for each included study as shown in Figure 6. The relationship between the sensitivity-specificity pair will define the appropriate approach to synthesizing outcomes. The following are some examples of how sensitivity-specificity relationships influence combining of study outcomes:

1. Likelihood ratios can only be combined when there is a linear relationship between sensitivity and specificity;
2. Outcomes should be summarized using diagnostic odds ratio when sensitivity-specificity assumes a curvilinear relationship, but not when the ROC curve is asymmetrical about the line of equal sensitivity and specificity;
3. Sensitivity or specificity alone should be combined in a situation when a change experienced in one does not influence the other.

Predictive values, likelihood ratios, summary ROC, diagnostic odds ratio (DOR) and meta-regression are some other approaches used in synthesizing diagnostic test accuracy studies depending on the initial relationships identified between sensitivity and specificity. Meta-analysis could be used to assess DTAs of the same condition, in which case the performance between tests should be described together with each test's individual performance.

Data extracted from primary studies are usually combined in a systematic review in the form of the meta-analysis (subject to the level of heterogeneity of the studies involved). When the level of heterogeneity does not allow for meta-analysis other approaches such as the use of a narrative summary and description may be used (Deville et al., 2002). The forms of heterogeneity are described in chapter 3. When meta-analysis is employed in combining data from studies of DTA, either a fixed effects model or a random effects model of statistical pooling could be used. The fixed effects model considers differences within study outcomes as due to random errors as studies are thought to be samples of one large study. With respect to the random effects model, the assumption is that real differences exist between study populations and procedures leading to other forms of error besides the random ones. Though homogenous studies are usually pooled with the fixed effects model, it is usually advised that most diagnostic studies be pooled with the random effects model in view of their predominantly low methodological quality (Deville et al., 2002).

Calculation of summary estimates of diagnostic test accuracy

Calculations of the outcome measures reported in DTA systematic reviews are summary measures of test accuracy: sensitivity, specificity, likelihood ratios and receiver operator characteristic (ROC) curve information. When combining diagnostic test accuracy data in meta-analysis, there are several factors to consider. Table 5 is derived from recent Cochrane Collaboration guidance (Macaskill et al., 2010) and summarizes the major challenges faced when combing data from primary research studies of diagnostic test accuracy, as compared

Table 5. Considerations when undertaking meta-analysis of DTA studies

Requirement	Explanation
Requirement for more data to be extracted for DTA reviews than for reviews of intervention effectiveness	Evaluating DTA requires knowing both the sensitivity and specificity of a test, whereas data from intervention studies generally only have a single measure to consider (e.g. difference in means or risk ratio)
Greater heterogeneity in data is expected in DTA reviews than for reviews of intervention effectiveness	Heterogeneity is to be expected in the results of a DTA meta-analysis, therefore a random effects model should be used to describe the variability across the included studies
DTA reviews require more sophisticated statistical methods than intervention reviews	A meta-analysis has to allow for the trade-off that occurs between sensitivity and specificity in studies that vary in the threshold used to define test positive and negatives
	Statistical methods that consider sensitivity, specificity, the relationship between them and the heterogeneity in test accuracy across studies are complex and therefore require specialized statistical knowledge and sophisticated computer software

with combining data from primary research examining the effectiveness of therapeutic interventions.

Briefly, the need to consider both sensitivity and specificity when assessing the accuracy of a diagnostic test makes data extraction and analysis more complex than for systematic reviews that examine therapeutic effectiveness.

Meta-analysis

In meta-analysis, the results of similar diagnostic studies are combined to calculate a summary estimate of a test's accuracy appropriately weighted for the sample size of the study. The science supporting meta-analyses of diagnostic studies has grown steadily since the 1980s. Whereas earlier reviewers lacked appropriate methods and tools for much analysis beyond simply reporting overall sensitivities and specificities for included studies, developments in the field now allow for sophisticated reporting and analysis. Some of these developments include: clarification on meta-analysis methods (Deeks, 2001b, Deville et al., 2002, Borenstein et al., 2009, Zhuo et al., 2002) the completion (in late 2010) of a chapter on 'Analyzing and Presenting Results' in the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy, (Macaskill et al., 2010) and the development of appropriate meta-analysis software such as RevMan (Cochrane Collaboration) and Comprehensive Meta-Analysis (CMA) (CMA, 2010) amongst others.

Inclusion of a meta-analysis in a systematic review of diagnostic studies is sufficient but not necessary. While completing meta-analysis is the aim of a systematic review of diagnostic test

accuracy, many published systematic reviews do not include meta-analysis and report results as a narrative summary (White and Schultz, 2011). Whether or not meta-analysis should be conducted depends on a number of factors, chiefly the number and methodological quality of the included primary studies and the heterogeneity of their findings of diagnostic test accuracy, as well as other features such as patient characteristics and methodologies as discussed in chapter 2.

Heterogeneity

When used in relation to meta-analysis, the term 'heterogeneity' refers to the amount of variation in the characteristics of included studies. For example, if three studies are to be included in a meta-analysis, does each of the included studies have similar sample demographics assess the same index test against the same reference test and utilize similar cut-off criteria for those tests? The methodological quality appraisal performed earlier in chapter 3 may serve as a guide in exploring for the sources of heterogeneity in the analysis phase of a systematic review (Leeflang et al., 2008b).

While some variation between studies will always occur due to chance alone, heterogeneity is said to occur if there are significant differences between studies, and under these circumstances meta-analysis is not valid and should not be undertaken. But how does one tell whether or not differences are significant? As mentioned earlier in chapter 3 when synthesizing several diagnostic test outcomes, it is essential to plot sensitivity-specificity pairs for each included study. The relationship between the sensitivity-specificity pair depicts how significant heterogeneity differences exist and helps define the appropriate approach to synthesizing outcomes (Tatsioni et al., 2005).

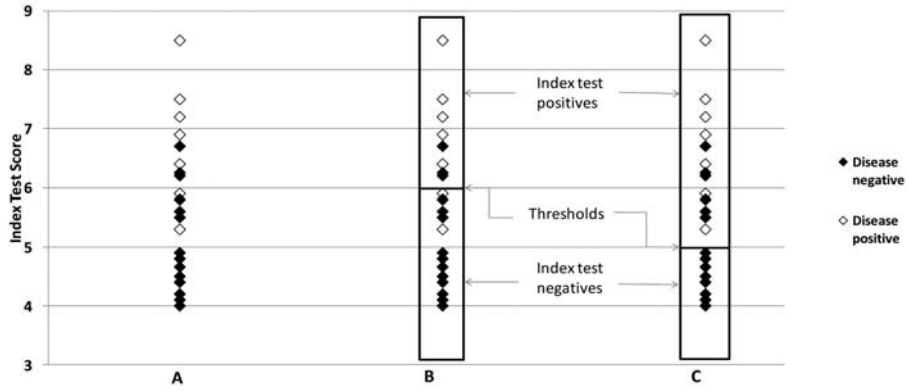
Positivity thresholds

As previously mentioned (Table 4) diagnostic test results may be binary (i.e. disease/condition is present or absent), ordinal (e.g. a 5 point scale such as definitely normal, probably normal, equivocal, probably abnormal, definitely abnormal) or continuous (e.g. measures of blood pressure or fasting blood glucose levels). Only binary results – derived by dichotomizing the results of an ordinal or continuous scale – are appropriate for meta-analysis (Macaskill et al., 2010, Borenstein et al., 2009, Zhuo et al., 2002). The continuous scale may be observed directly, such as blood pressure measured on a sphygmomanometer, or it may be latent, or unobserved, such as a radiologist's degree of suspicion about an abnormality on an x-ray (Gatsonis and Paliwal, 2006). In both cases, binary results are derived through the application of a threshold for test positivity. Results on either side of the threshold indicate the presence/absence of the disease/condition, thus allowing the summary of the test results in a 2×2 table, as shown in Table 2.

The impact of positivity thresholds on the sensitivity and specificity of diagnostic tests has been well documented and is graphically represented in Figure 3. Consider a diagnostic test which reports continuous data and a positivity threshold, which is set at a test result of 6. A score equal to or above this indicates presence of the disease/condition, and lower scores indicate absence of the disease/condition. (For the sake of consistency, this convention will be followed throughout this text – a larger test score will be assumed to indicate greater certainty about the presence of the disease/condition). If the positivity threshold is lowered, a

Figure 3: Graphical representation of hypothetical data to illustrate the impact of positivity thresholds on sensitivity and specificity measures. Condition A is the raw data, condition B has an index test with a threshold score of 6. In condition C the threshold is reduced to 5. Open symbols are true positives and closed symbols are true negatives (as determined by a reference test) (Gordis, 2000).

A



B

		Reference		
		+	−	
Index	+	5	3	8
	−	2	11	13
		7	14	21

Sensitivity = 5 / (2 + 5) = 0.71;
Specificity = 11 / (11 + 3) = 0.79

C

		Reference		
		+	−	
Index	+	7	6	13
	−	0	8	8
		7	14	21

Sensitivity = 7 / (7 + 0) = 1.0;
Specificity = 8 / (6 + 8) = 0.57

greater number of true positives (cell 'a' in Table 2) will be captured, thus increasing sensitivity (a/(a + c)). However, the number of false positives (cell 'b' in Table 2) will also be increased, thus reducing the specificity (d/(b + d)) of the test. The calculations for sensitivity and specificity are included in Figure 3.

In the left cell of Figure 3 (A), the results of a hypothetical diagnostic test accuracy study of 21 samples are shown with results ranging from 4 to 8.5. There are seven samples in which the disease is present and 14 samples in which the disease is absent, as determined from the reference standard. Using the standard 2 × 2 cross classification table, we know from this information that a + c = 7 and b + d = 14. In the middle cell of Figure 3 (B), a threshold for the index test results is set at a score of '6'; all 8 test results with a score greater than this are deemed by the index test to be diseased. Using the 2 × 2 table approach, the sum of 'a' and 'b' therefore equals 8, and it is apparent that there are three false positives (samples that are positive by the index test but negative by the reference test). Therefore 'b' equals 3, and consequently 'a' must equal 5. There are 13 samples below the threshold, and these are considered by the index test to be negative, therefore c + d = 13. Completion of the 2 × 2 table is relatively straightforward. Given that we know that a + c = 7, and a = 5, then c = 2.

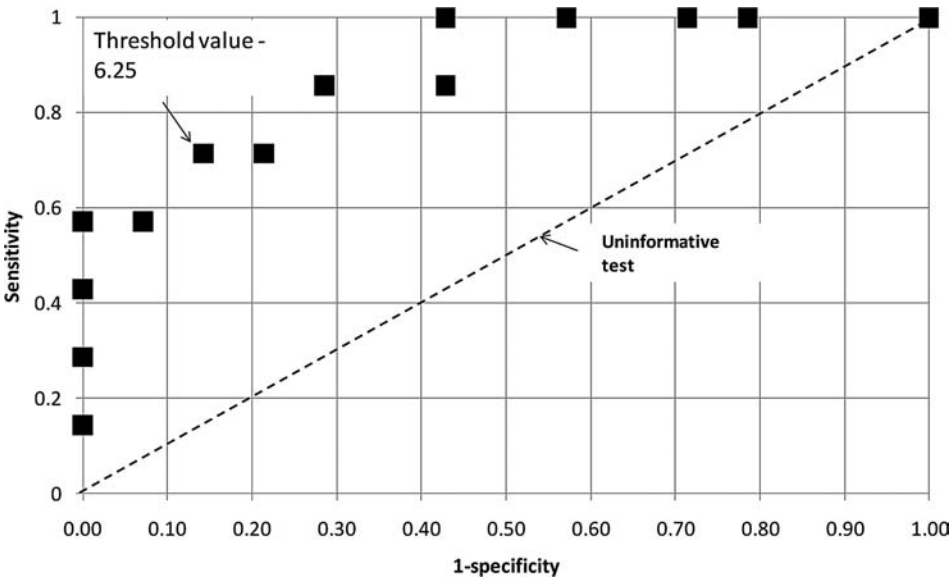
Similarly, if b + d = 14, and b = 3, then d = 11. The resulting sensitivity (0.71) and specificity (0.79) are calculated below.

If the threshold is lowered to a score of '5' as in Figure 3 (C), then all of the diseased samples will be considered disease positive by the index test, and the resulting sensitivity is 1.0. However, there are more false positives (n = 6) identified by the test, and these contribute to a substantial drop in the specificity to 0.57.

Although hypothetical, this example illustrates the difficulty in interpreting test results when a test is not 100% sensitive or specific. Additionally, as Gordis (2000) points out, when reviewing the results of a diagnostic test, the clinician does not have the benefit of the results of the reference test (Gordis, 2000).

Plotting a range of sensitivity and specificity values derived from varying the threshold allows determination of the most appropriate threshold for that particular test. Generally, the aim is to determine the threshold at which sensitivity and specificity are maximised. This plot is termed a receiver operator characteristic (ROC) plot, and involves plotting the sensitivity (y-axis) against 1-specificity (false positive rate) (x-axis). Ideally, a diagnostic test should have high sensitivity and high specificity. Such a test would have most points on an ROC plot in the top left of the plot, where both sensitivity and specificity are close to 1. Connecting the data points in the plot generates a ROC curve, which has a distinctive concave, shoulder-like curve. The greater the area under the curve (AUC), the better the diagnostic accuracy of the test. The AUC for a perfect test is 1. Conversely, the dashed diagonal line in Figure 4 indicates the results of an

Figure 4: Plot of hypothetical data from Figure 6 with pairs of sensitivity and specificity determined across 14 positivity thresholds ranging from 3.75 to 8.5. The plot of sensitivity (y-axis) versus 1-specificity is termed a ROC curve



uninformative test that does not distinguish between the presence or absence of the disease. The AUC of the uninformative test is 0.5.

The ROC plot data presented for our hypothetical diagnostic test indicates only reasonable sensitivity and specificity. Ultimately, the exact nature of the test and the condition, and the implications of potentially missing truly diseased samples (i.e. false negatives) or incorrectly labelling a sample as diseased (i.e. false positives), will dictate what threshold is recommended for use. (Gordis, 2000) The cost of false positives includes potential emotional distress to those who require secondary testing, and the additional costs of any subsequent secondary testing. Conversely, the missed diagnosis in the false negatives may represent a missed opportunity to catch a disease early in its progression, or can provide false reassurance to a clinician or a patient and deleteriously affect their healthcare. In this example, at a threshold of 6.25, the sensitivity is 0.71 and specificity is 0.86. This would seem to be a reasonable positivity threshold value for maximising both sensitivity and specificity.

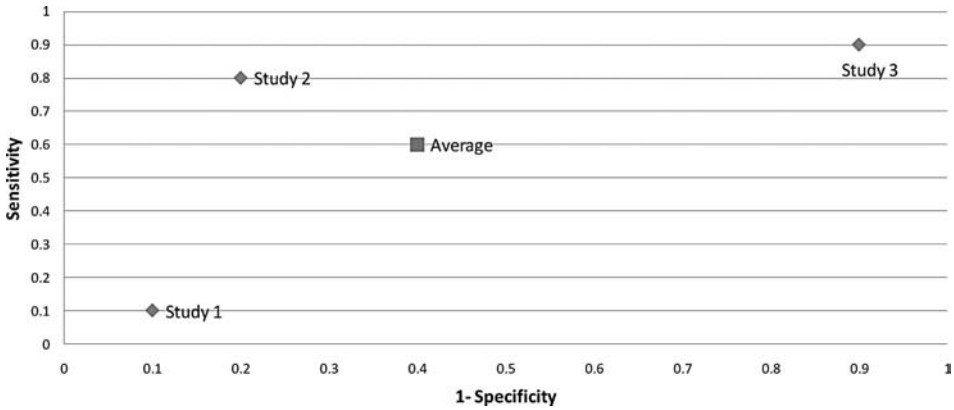
Variation between primary studies in positivity threshold

In the previous section, we have considered the case of variation in the positivity threshold impacting on the sensitivity and specificity of a single study. However, variation between studies in the choice of positivity threshold must also be considered prior to meta-analysis of DTA studies because a study that uses a higher threshold to diagnose a disease will likely find greater sensitivity at the expense of lower specificity. Although variation in thresholds between studies may be explicit and defined within the study (for example the number of colony forming units defining the presence/absence of a urinary tract infection or the presence/absence of diabetes based on fasting blood glucose levels), it may also be implicit, that is, not declared by set criteria (Irwig et al., 1995). Some examples of implicit variation include the assessment of radiographic abnormality by different readers, differences in observers, laboratories or equipment (Irwig et al., 1995, Reitsma et al., 2005). Unless accounted for when considering multiple studies, implicit or explicit variation in positivity thresholds can obscure true diagnostic test accuracy.

Irwig et al (1994) identify the implications of ignoring variation in positivity thresholds and simply pooling the results of several DTA studies to create simple or weighted averages of sensitivity and specificity. Chiefly, one can expect an underestimation of sensitivity and specificity and a negative correlation between sensitivity and specificity across primary studies. (Irwig et al., 1995, Irwig et al., 1994) Gatsonis and Paliwal (2006) (Gatsonis and Paliwal, 2006) provide an example (p. 274) in which three studies have sensitivity and specificity pairs of (0.1, 0.9), (0.8, 0.8) and (0.9, 0.1). These three pairs fit a classic ROC plot with a point close to the top left corner that indicates a high quality diagnostic test that accurately distinguishes between those with and without a condition or disease of interest (Figure 5). However, the average pair (0.6, 0.6) is very distant and clearly do not represent the data in any useful way. Therefore, calculating simple or weighted averages of sensitivity and specificity is not appropriate for drawing statistical conclusions of aggregated diagnostic test accuracy studies.

The issue of how to account for variation in positivity threshold between primary studies will be addressed in the following section on summary receiver operator characteristic (sROC) curves.

Figure 5: Plot of hypothetical data from three studies in ROC format illustrating that calculating an average is inappropriate meta-analysis of diagnostic test accuracy. The 'average' point is very distant from any of the actual data points that closely fit a ROC curve and therefore does not accurately summarise the three data points (from Gatsonis and Paliwal, 2006).



The six steps of meta-analysis

Deville et al (2002) summarised the literature on meta-analysis of diagnostic studies into six recommended steps:

(i) presentation of results of individual studies
(ii) searching for the presence of heterogeneity
(iii) testing the presence of an (implicit) cut-point (threshold) effect
(iv) dealing with heterogeneity
(v) deciding which model should be used if statistical pooling is appropriate, and
(vi) statistical pooling.

Each of these steps will be addressed in the following sections. Other steps also add value to a systematic review. For example, Deeks (2001) (Deeks, 2001b) illustrated the calculation of the post-test probability of a positive or negative test actually indicating the presence or absence of a disease. For example, the post-test probability of having endometrial cancer if endometrial thickness (as measured by endovaginal ultrasonography) is 5mm or less was calculated as 0.013. That is, 1.3% of women with endometrial thickness of 5mm or less will have endometrial cancer. Such information is particularly valuable for clinicians to aid interpretation and assist in clearly informing patients of the results of diagnostic tests. The use of summary estimates from systematic reviews strengthens the evidence base for such calculations.

Presentation of results of individual studies

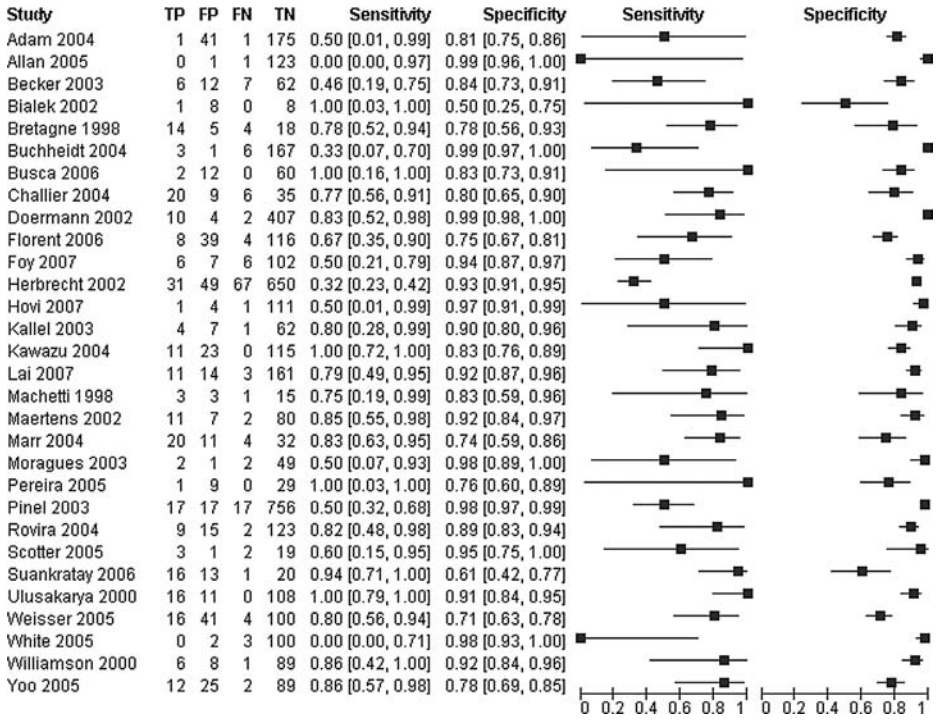
Presenting the results of individual diagnostic test accuracy studies gives the reader of the systematic review a better understanding of the outcomes and provides a first insight into potential heterogeneity of study findings (Deville et al., 2002). There are two main methods

of displaying the results of individual studies: forest plots and summary receiver operator characteristic (sROC) curves (Macaskill et al., 2010).

Forest plots

Forest plots are used in both meta-analyses of studies of effectiveness and diagnostic test accuracy to summarise the results of individual studies. Each study is presented with background information (author and year of publication) and the results of the diagnostic study, including numbers of true positives ('a' from 2 × 2 table), false positives ('b'), false negatives ('c') and true negatives ('d') (Leeflang et al., 2008b). The sensitivity and specificity, with 95% confidence intervals, are presented in graphical and tabular format within the forest plot. When used with diagnostic studies, the term 'coupled forest plot' may be used to emphasise the pairing of sensitivity and specificity measurements (Leeflang et al., 2008b, Macaskill et al., 2010). An example of a coupled forest plot, taken from the Cochrane Review of galactomannan detection for invasive aspergillosis (Leeflang et al., 2008a), is presented as Figure 6. Note

Figure 6: An example of a linked forest plot, taken from the Cochrane Review of galactomannan detection for invasive aspergillosis (Leeflang et al., 2008a). The data included in the figure are: the study identity, the number of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). Sensitivity and specificity, including 95% confidence intervals for each study are presented in tabular and graphical formats, with blue boxes marking the values and horizontal lines marking the confidence intervals.



for systematic reviews comparing multiple index tests, and/or reference tests, the test type and positivity threshold should be identified in the forest plot.

Summary ROC plots

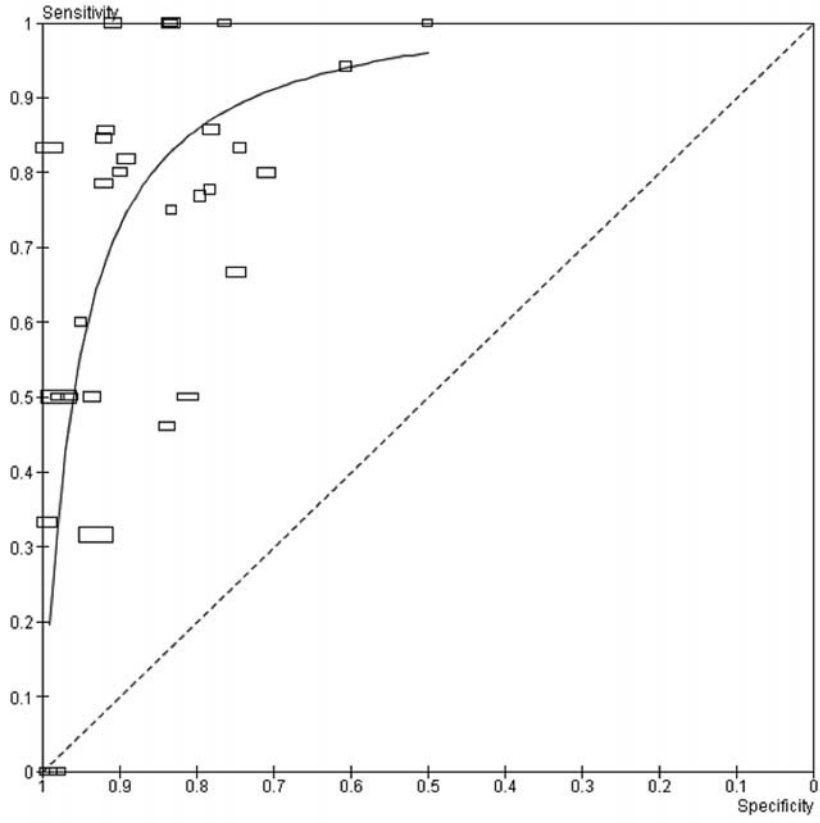
We have previously considered a ROC curve to plot the differential sensitivity and 1-specificity (false positive rate) of a single diagnostic study at different positivity thresholds or cut-offs (see chapter 3). Summary ROC (sROC) plots extend this example to include an estimate of sensitivity and specificity from each of a number of studies on the same plot. Summary ROC plots allow graphical representation of the results of a number of studies. The sROC approach to meta-analysis of diagnostic studies was first developed by Moses and Littenberg (Hasselblad and Hedges, 1995, Littenberg and Moses, 1993, Midgette et al., 1993, Moses et al., 1993, Walter, 2002). The aim is to find a smooth curve between the points on the sROC plot.

In a summary ROC plot, each study is represented by a single sensitivity-specificity point. The first stage of the process is calculation of the paired sensitivity and specificity for each study, which is conducted as per the standard 2 × 2 table and the equations shown in chapter 1. RevMan automatically calculates 95% confidence intervals from the populated 2 × 2 table. The size of the symbol, which is typically rectangular in shape, in the sROC plot is modified to depict the precision associated with each measure. The precision can be estimated using the standard errors (the inverse of the standard error of the logit(sensitivity) and logit(specificity)), or according to sample size (Macaskill et al., 2010). In RevMan sROC plots, the height of the square is directly related to the precision of the measurement of the number of diseased (and hence sensitivity) and the width is related to the precision of the measurement of the number of non-diseased (and hence the specificity). Larger squares indicate greater precision, and, usually, greater sample sizes.

An sROC curve, taken from the Cochrane Review of galactomannan detection for invasive aspergillosis (Leeflang et al., 2008b), is presented as Figure 7. Data for 30 studies at three positivity thresholds is included. As the width of the rectangles exceeds the height in most cases, it is apparent that at the study level, there are more non-diseased patients (the sum of false positives FP and true negatives TN) than diseased (the sum of true positives TP and false negatives FN). Phrased slightly differently, there is greater precision associated with the measurement of the specificity than the sensitivity. Comparing across studies, there is greater variability in the sensitivity (which ranged from 0-1) than the specificity (range 0.5-0.99). The distribution of study results on the sROC plot is consistent with a threshold effect, as there is a relatively close fit of the available data to the curve.

The method of calculating the sROC curve is based on regressing the log diagnostic odds ratio against a measure of the proportion of test results reported as positive (Leeflang et al., 2008b). The diagnostic odds ratio (DOR) is a measure of the discriminative power of a diagnostic test – the ratio of the odds of a positive test result among people with the disease/condition to the odds of a positive test result among people without the disease/condition (Deville et al., 2002). The value of the DOR ranges from zero to infinity, with higher values indicative of better discriminative performance (Glas et al., 2003). A value of 1 indicates that the test does not discriminate between people with and without the disease/condition. A score of less than 1 indicates more negative test results among those with the disease/condition and suggests

Figure 7: An sROC plot, showing the specificity-sensitivity pairs for 30 studies assessing the diagnostic accuracy of galactomannan detection for the diagnosis of invasive aspergillosis (IA) in 4792 patients (data from (Leeflang et al., 2008a)). The width of the rectangles is proportional to the number of patients with possible IA or without IA; the height of the rectangles are proportional to the number of patients with proven or probable IA. The solid line is the sROC curve, the dashed line indicates the results of an uninformative test. Data from three positivity thresholds are presented.



improper test interpretation (Glas et al., 2003). The DOR increases rapidly as sensitivity and specificity approach 1.

The diagnostic odds ratio can be calculated as the ratio of the positive likelihood ratio to the negative likelihood ratio (i.e. DOR = +ve LR / −ve LR). Alternatively, given the equations for +ve LR and –ve LR, DOR may be calculated as (TP/FP)/(FN/TN). Although the DOR is useful in summarising the results of a diagnostic test in a single statistic that describes how many times higher the odds are of obtaining a test positive result in a diseased person than a non-diseased person, it is not routinely reported for individual studies as it has little direct

clinical applicability (Macaskill et al., 2010). Nevertheless, it is mentioned here because of its importance in the calculation of sROC curves.

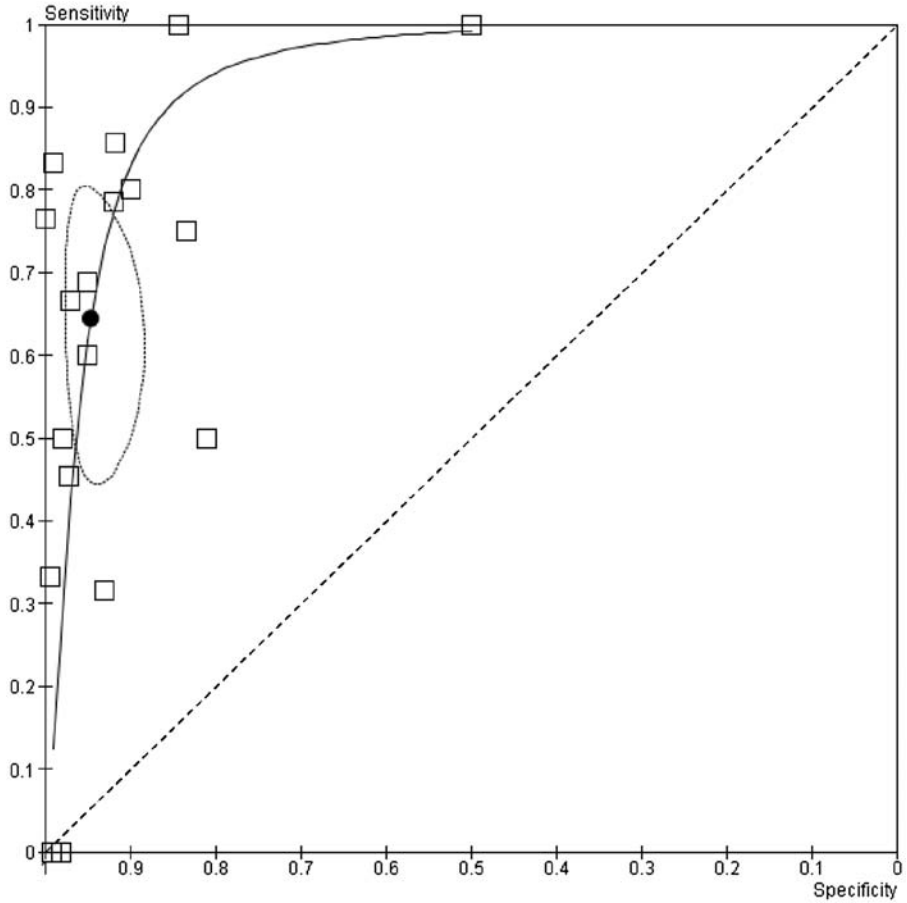
Plotting all the data to be considered for inclusion in meta-analysis in a sROC curve plot allows the reviewer to carefully consider the data and how best to analyse it. The Cochrane Collaboration proposes two choices of summary statistics for meta-analysis. Particularly in cases where the reviewer can be confident that all studies have used a common threshold (termed the average operating point), a reviewer can estimate the overall summary sensitivity and specificity, including 95% confidence intervals for that threshold (Leeflang et al., 2008b, Macaskill et al., 2010). Indeed, if there is little variation in positivity threshold, then the points in the sROC space will likely be tightly clumped, making it difficult to generate a meaningful sROC curve (Macaskill et al., 2010). Alternatively, if there are a range of thresholds (either implicit or explicit), it is more appropriate to estimate and report the entire sROC curve. Leeflang's summary (Leeflang et al., 2008b) illustrated both methods. However, given the positivity threshold effects previously noted for this data-set, it is most appropriate to conduct these meta-analytical methods on each of the threshold levels separately. For example, Figure 8 illustrates a sub-set of the data, with only a single positivity threshold plotted. Both sROC curves and mean specificity and sensitivity are presented here for comparative purposes. As indicated, the final decision about whether one or both of these methods are presented is determined by the reviewer who must decide if either better answers the research question (Macaskill et al., 2010). The two statistics may complement each other in providing clinically meaningful summaries, and in identifying the presence of effects; therefore, reviewers may reasonably decide to include both in a review. Additionally, if there is any doubt on the behalf of the reviewer, presenting and explaining both summary statistics will allow the reader to decide.

The methods for calculating these summary statistics are discussed later in this chapter.

Linked ROC plots

Systematic reviews of diagnostic test accuracy often aggregate multiple studies that have evaluated a single index test compared to the reference standard. Alternatively, a review may seek to compare a number of different index tests. If different tests are compared in a number of studies that examines different patient populations, or uses different laboratory analysts, substantial heterogeneity is likely and will confound comparisons (Leeflang et al., 2008b). Therefore, the preferred meta-analytic approach is to only include studies that have directly compared the accuracy of a pair of index tests in the same group of patients, or have randomised patients to each of the paired tests in the same study (Leeflang et al., 2008b). To illustrate the results of such studies, the Cochrane Collaboration recommends using linked ROC plots in which the results for each test (of a pair) are plotted using a different symbol and connected with a line (Macaskill et al., 2010). This allows an assessment of whether within-study differences for a pair of tests are consistent across the included studies, Linked sROC plots can be used to calculate either sROC curves, or an average sensitivity or specificity with 95% confidence intervals (Leeflang et al., 2008b). Therefore, two sets of summary statistics are generated and can be compared (Leeflang et al., 2008b).

Figure 8: sROC plot of 18 studies assessing the diagnostic accuracy of galactomannan detection for the diagnosis of invasive aspergillosis (IA) in 2777 patients with a positivity threshold at 1.5 ODI (optical density index) (data from Leeflang et al., 2008a). The solid line is the sROC curve, the dashed line indicates the results of an uninformative test. The filled circle marks the mean sensitivity and specificity, the dotted ellipse around the mean mark the 95% CI.



Searching for the presence of heterogeneity

Heterogeneity in included studies is a common finding in systematic reviews of DTA studies. For example, Willis and Quigley (2011) reported that 70% of 236 DTA meta-analyses published prior to 2009 identified heterogeneity (Willis and Quigley, 2011). Inherent variation between DTA studies, beyond that explicable by chance due to small sample sizes, has numerous causes. For example, differences in study patient populations, methods of conducting tests and interpreting results, the type of reference standard, or positivity thresholds all contribute to heterogeneity (Lijmer et al., 2002).

Flaws in study design and inclusion of biased studies ("methodological" or "artefactual" heterogeneity) (Irwig et al., 1995) can also be a significant factor contributing to heterogeneity (Begg, 2008, Lijmer et al., 2002). Begg (2005) illustrated how low quality DTA studies can have extremely high DOR and sensitivity-specificity (and thereby contribute to heterogeneity) if indeterminate findings are excluded from the analysis. For example, a review of ultrasound to screen for deep vein thrombosis in post-operative asymptomatic patients found significant heterogeneity. While some of the included studies reported indeterminate results, one study that had a high number of indeterminate results excluded these results, leading to very high DOR and sensitivity-specificity. Similarly, different units of analysis (the patient versus the limb), and clinical heterogeneity in terms of intervention received (hip and/or knee surgery versus craniotomy patients) significantly affected DTA and thereby increased heterogeneity (Begg, 2005). While heterogeneity is a major challenge to meta-analysis, the presence of heterogeneity should be further investigated as it increases the scientific value and clinical relevance of reviews (Thompson, 1994). So, how does one identify heterogeneity?

Plotting data on paired forest plots and sROC can provide a useful – though subjective – insight into the presence or absence of heterogeneity. Dinnes et al (2005) found that over half of the meta-analyses that they examined graphically represented the spread of study results, mostly through sROC plots of sensitivity and specificity. When viewing the sROC plots, heterogeneity is indicated by how closely the included data fits to an sROC curve, rather than how close the points are to each other in ROC space. Data that tightly fit a typical shoulder-shaped sROC curve indicates low heterogeneity and suggest a threshold effect. In the example cited in (section 0) of (Gatsonis and Paliwal, 2006) (Figure 1, p. 274) the sensitivity-specificity pairs of (0.1, 0.9), (0.8, 0.8) and (0.9, 0.1) would have low heterogeneity as they all fall close to a typical sROC curve (Figure 5, section 0). However, adding an additional data set to the ROC plot (0.6, 0.6), which is very distant from the sROC curve, onto the ROC plot, would strongly suggest the presence of heterogeneity for the latter results.

Some methods for identifying heterogeneity in meta-analysis of randomised controlled trials can be adapted for use in DTA studies (Lijmer et al., 2002). However, the I^2 is not routinely used in Cochrane DTA systematic reviews, as it does not account for phenomena such as positivity threshold effects. Dinnes et al (2005) found that statistical tests to identify heterogeneity were used in 41% of meta-analyses. The Chi-square test and Fisher's exact test to assess heterogeneity in individual aspects of test performance were most commonly used (Dinnes et al., 2005). However, the power of these tests tends to be low (Deville et al., 2002).

Testing the presence of a threshold effect

Although plotting DTA data in forest plots can provide a useful insight into the presence or absence of heterogeneity, the existence of threshold effects is not readily apparent from such plots. Deville et al. (2002) suggested calculating a Spearman correlation coefficient between sensitivity and specificity of all included studies to identify a threshold effect. A strong negative correlation between specificity and sensitivity is indicative of a threshold effect and suggests that the pairs of parameters represent the same DOR. However, Dinnes et al. (2005) found that only 16% of meta-analyses of DTA studies used the Spearman correlation coefficients to identify a threshold effect.

As previously mentioned, plotting the sensitivity-specificity pairs in ROC space is another useful method of identifying the existence of a positivity threshold effect. For example, Figure 7 includes data from three extrinsic positivity thresholds, and the resulting distribution of data in ROC space is a reasonable approximation of an sROC curve. The presence of a positivity threshold effect suggests that a sROC curve is an appropriate method of presenting the results and illustrating relationships, particularly if separate sROC curves are calculated for different thresholds. Nevertheless, the reviewer may decide to include both sROC curves and mean sensitivities and specificities to fully illustrate relationships within the data (Macaskill et al., 2010).

Dealing with heterogeneity

Heterogeneity between studies is an expected finding of a systematic review of diagnostic test accuracy (Macaskill et al., 2010). Begg (2005, 2008) strongly advocates for investigating reasons for heterogeneity on a case-by-case basis, with particular focus on potential for bias in study design. Harbord et al. (2008b) identify a case of a particular study type (diagnostic case-controls) contributing substantially to heterogeneity of meta-analysis of magnetic resonance imaging for the diagnosis of multiple sclerosis. Their response was to restrict the analysis to diagnostic cohort studies and therefore restrict meta-analysis to studies at low risk of bias. The uses and limitations of using sub-group analysis to explore reasons for heterogeneity in meta-analyses of randomised controlled trials also apply to meta-analyses of DTA studies (Macaskill et al., 2010). See Section 9.6 of the Cochrane Handbook for Systematic Reviews of Interventions for more information (Deeks and Altman, 2008). However, the generally poor quality of the conduct and reporting of studies of DTA (White and Schultz, 2011) can preclude sub-group analysis as a potential avenue for addressing heterogeneity. Sub-groups with very low numbers of studies will be prone to heterogeneity by virtue of the low number of studies included. Exploratory analysis can be conducted using RevMan and the more simple Moses and Littenberg models of sROC curves; however, this method is not valid for inferential statistics (Macaskill et al., 2010). In this case, each sub-group is analysed separately and the results compared across the sub-groups. Again, low numbers of included studies in some sub-groups will affect the feasibility of generating separate sROC curves (Macaskill et al., 2010).

In the presence of heterogeneity for DTA meta-analyses the Cochrane Collaboration advocate use of random effects meta-analysis. In fact, in Cochrane DTA reviews, heterogeneity is assumed and random effects models are fitted by default (Macaskill et al., 2010). If bivariate random-effects meta-analysis (see section 3.13) is conducted, the amount of heterogeneity observed can be enumerated (Macaskill et al., 2010). However, these values are difficult to interpret as they are expressed on log odds scales (Macaskill et al., 2010). Outliers may be excluded following careful examination and explanation (Deville et al., 2002). A Galbraith plot (Deville et al., 2002, Lijmer et al., 2002) can be used to identify outliers.

Lastly, if study heterogeneity is severe, reviewers may elect to refrain from pooling the results and conduct a narrative synthesis (Deville et al., 2002, Harbord et al., 2008b).

Model selection

The regression model proposed by Moses and Littenberg has been widely promoted for meta-analysis of DTA studies (Vamvakas, 1998, Deeks, 2001b, Deville et al., 2002, Reitsma

et al., 2005). However, more recent findings have identified flaws in this approach (Arends et al., 2008, Gatsonis and Paliwal, 2006). Briefly, the limitations of the model have been reported as: a failure to consider the precision of the study estimates, an inability to estimate between-study heterogeneity and the inclusion of measurement error in the regression model (Leeflang et al., 2008b, Willis and Quigley, 2011). According to Leeflang et al (p. 893) (2008b) these limitations "render estimates of confidence intervals (CIs) and probability (P) values unsuitable for formal inference".

Two newly proposed methods for fitting random effects to hierarchical models can improve the statistical analysis of between-study variation in sensitivity and specificity (Leeflang et al., 2008b, Willis and Quigley, 2011), although see (Begg, 2008). These methods are: the bivariate random effects model (BRM) (Reitsma et al., 2005) and the hierarchical sROC (HSROC) model (Rutter and Gatsonis, 2001). These methods model sensitivity-specificity pairs from each included study and give a valid estimation of the underlying sROC curve (HSROC) or the average sensitivity and specificity (BRM) (Leeflang et al., 2008b). Although they are often discussed as different ways to analyse data, the two methods give identical results when covariates are not fitted (Harbord et al., 2007). Addition of covariates to the models, or application of separate models to different sub-groups, allows examination for potential sources of heterogeneity, such as bias and sub-group variation (Leeflang et al., 2008b). Both methods require relatively sophisticated statistical software to calculate the model summary statistics (Begg, 2008, Harbord et al., 2008a, Harbord et al., 2008b). Once the models are calculated and entered into RevMan 5, publication-ready graphics can be easily produced. The main differences between the two methods will be discussed briefly in the next section.

Statistical pooling

Both BRM and HSROC methods are considered to be hierarchical models (Harbord et al., 2008b). BRM estimates average sensitivity and specificity and the unexplained variation in these parameters and the correlation between them, as well as confidence intervals. BRM is a random effects model, in which logit (sensitivity) and logit (specificity) are assumed to be related, while each following a normal distribution across the included studies (Jones et al., 2010, Leeflang et al., 2008b). In addition to providing a summary of average sensitivity and specificity, BRM provides a confidence region for the summary point and a prediction region in ROC space within which it is expected that the true sensitivity and specificity lies (Harbord et al., 2008b).

In comparison, the HSROC approach (Rutter and Gatsonis, 2001) expresses the relationship between logit-transformed sensitivity and specificity in each study in terms of accuracy (i.e. the log of the DOR) and threshold (Harbord et al., 2008b). The method allows for between-study variation in these variables, and a parameter that determines the shape of the summary curve (Harbord et al., 2008b). The results are usually expressed as a summary ROC curve.

Both methods allow for the inclusion of study-level covariates that may help to explain between-study heterogeneity. The addition of covariates is a form of meta-regression; the caveats previously identified by (Thompson and Higgins, 2002) about meta-regression for meta-analysis of studies of effectiveness should also be considered for diagnostic studies. For example, it may be difficult to determine which covariates should be examined and whether

they are truly associated with effects. If covariates are included, it is important that statistically rigorous methods are used to control against the risk of identifying spurious relationships (Higgins and Thompson, 2004).

Full details about how the BRM and HSROC are calculated are outside the scope of this text. General discussion about the need for BRM and HSROC is found in (Chappell et al., 2009, Harbord et al., 2007, Harbord et al., 2008b, Whiting et al., 2008a, Willis and Quigley, 2011).

For further details about BRM methodology, the reader is directed towards the following studies: Arends et al., 2008, Macaskill, 2004, Reitsma et al., 2005, Riley et al., 2007, Verde, 2010.

For further details about HSROC methodology, the reader is directed towards Macaskill, 2004, Rutter and Gatsonis, 2001, Walter, 2002, Wang and Gatsonis, 2008.

Additionally, the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy includes a clear summary of the two methods. In particular, readers are directed to section 10.5 'Model fitting', which illustrates how to enter parameter estimates from modelling conducted in statistical software (e.g. SAS, Stata) into RevMan. SAS commands and examples are included as an appendix to the Cochrane Handbook. RevMan uses the parameter estimates to calculate: a summary curve, summary operating point, a confidence region and prediction region for the summary point. The BRM can be fitted using a generalised linear mixed model and common routines are Proc NLMIXED (or Proc GLIMMIX) in SAS, and **xtmelogit** in Stata. Different commands are available for fitting models without covariates (Macaskill, Gatsonis et al. 2010).

The HSROC model is represented by a generalised non-linear mixed model. (Macaskill, Gatsonis et al. 2010). Inclusion of covariates in this model reduces the range of software available – typically this model is fitted using Proc NLMIXED in SAS, although the SAS METADAS macro also fits HSROC models with and without covariates (Macaskill, Gatsonis et al. 2010). The Stata **metandi** command can be used to fit HSROC models, but only in the absence of covariates (Macaskill, Gatsonis et al. 2010).

Systematic reviews of DTA in clinical practice

As observed in earlier sections, diagnostic tests are essential in the care of patients under clinical conditions. Systematic reviews of DTA are vital in arriving at best available evidence in the application of diagnostic tests in practice. Clinical history taking and examination were identified as accurately diagnosing up to 9 in 10 patients in a review by Willis & Quigley (2011).

This situation could be attributed to the development of newer diagnostic technologies at the expense of clinical tests; a situation that is bound to continue in the years to come (Willis and Quigley, 2011). There is also a predominance of tertiary settings over primary care and emergency settings in the analysis of tests. This is contrary to the fact that most patients are cared for in the primary care and emergency settings. This could be attributed to the fact that most resources for testing are present in the tertiary facilities as compared to the primary care and emergency settings (Willis and Quigley, 2011). To make results of tests more relevant to the target population, much more tests need to be carried out in the primary settings to make their outcomes more relevant to the predominant population being served.

The ability to interpret the findings of systematic reviews of DTA plays a vital role in the end user adapting the findings in practice. Unlike systematic reviews of therapeutic interventions that report a single statistic such as the odds ratio or risk ratio for example, reviews of DTA have a minimum of 2 statistics i.e. sensitivity and specificity (Willis and Quigley, 2011).

Despite improvements in reporting standards, several problems are still associated with meta-analysis in diagnostic research. There is still a lack of consensus on the best approach to aggregation of study results; in a study by Willis and Quigley, 2011, over two-thirds of studies used two or more statistical methods were used in aggregation (Willis and Quigley, 2011). The pooling of sensitivity/specificity and the use of SROC curve predominate the approaches to aggregation of study outcomes despite their deficiencies.

Likelihood ratios are alternative statistics for summarizing diagnostic accuracy with a higher relevance to clinical practice than other statistics (Deeks, 2004). Likelihood ratios are ratios of probability and can be treated the same way as risk ratios with confidence intervals (Deeks, 2004). Likelihood ratios can help adapt results of a study to future diagnostic/post-test probability of patients by multiplying the pre-test probability of a condition by the likelihood ratio identified in the study (Deeks, 2004). Using Fagan's nomogram and a straight-edged ruler, one can determine the post-test probability of a condition when the pre-test probability and likelihood ratio are known.

SECTION 2

Conducting a Systematic Review of Diagnostic Test Accuracy Evidence

Chapter 4:

Planning a systematic review of diagnostic test accuracy evidence

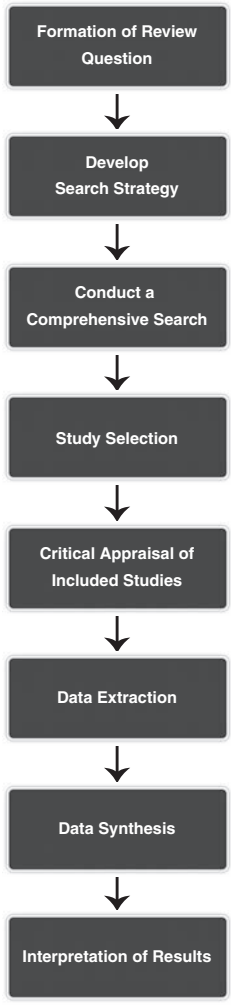
Conducting a systematic review of diagnostic test accuracy can be divided into eight main stages as shown in Figure 9. Each systematic review requires a detailed plan of how that particular review is going to be conducted. This plan is called the systematic review protocol and it is important because it pre-defines the objectives and methods of the systematic review. It is the systematic approach to the conduct and report of the review that allows transparency of process, which in turn allows the reader to see how the findings and recommendations were arrived at. The protocol details the criteria the reviewers will use to include and exclude studies, to identify what data is important and how it will be extracted and synthesized. A protocol provides the plan or proposal for the systematic review and as such is important in restricting the presence of reporting bias. As such, any deviations between the protocol and systematic review report should be discussed in the systematic review report.

The systematic review protocol

An apriori systematic review protocol is important because it pre-defines the objectives and methods of the systematic review. The protocol details each of the steps, as well as the criteria the reviewers will use to include and exclude studies, to identify what data is important and how it will be extracted and synthesized. Having a detailed plan for a systematic review is important in maintaining rigor. Acronyms can be useful when developing the protocol. The following is suggested as a way of ensuring all of the necessary elements are included in the protocol:

- **P** – population/participants
- **I** – index test
- **R** – reference test
- **A** – accuracy measures
- **T** – treatment outcomes
- **E** – expected outcomes of the test (e.g. triage, replacement, add-on etc.)

Figure 9: Important stages in the systematic review process.



Each of the PIRATE elements and stages of a diagnostic test accuracy protocol will be discussed in turn in the following sections. It is important to note that care should be taken when writing the protocol is as accurate as possible. If during the conduct of the review there are any deviations between the protocol and systematic review report, these should be discussed in the systematic review report.

Review title

In line with the STARD recommendations, a clear title allows the best chance of appropriate indexing by databases and the identification of the review as being one of diagnostic test accuracy. Furthermore, the title should be as descriptive as possible to allow potential readers

to determine whether the review addresses their particular needs. Incorporation of PIRATE elements into the title will aid its clarity.

The example below is taken from a recent JBI systematic review (White and Schultz, 2011).

> *The diagnostic test accuracy of laboratory tests for Influenza A H1N1 (Swine Flu) testing: a systematic review*

This example provides readers with a clear description that the review aims to assess the diagnostic test accuracy of laboratory tests (no restrictions are noted so it assumed that all tests and all populations are being considered).

Background

There should be a comprehensive, clear and meaningful background section to every systematic review. The background should provide sufficient detail to justify the conduct of the review, the choice of index and reference tests, as well as to justify the population of interest and outcome measures. Where complex or multifaceted tests/techniques are being described, it may be important to detail the whole of those tests or techniques for an international readership. Any topic-specific jargon or terms and specific operational definitions should also be explained.

The background should describe all of the elements that are going to be considered in the review. The mnemonic PIRATE may be a useful way to identify necessary areas to be covered by this section.

P – population/participants – Who is the population of interest for this review? In the above example, the population may be patients presenting with influenza-like illnesses, regardless of age or gender.

I – index test – What are the tests under evaluation in this review? In the above example, all laboratory tests used to confer a diagnosis of Influenza A H1 N1 (Swine Flu) will be considered.

R – reference test – To what test are the index tests going to be compared with? I.e. what is the best test currently available? In the example above, viral culture is used as the reference standard.

A – accuracy methods – How will accuracy be measured? Recommended values are sensitivity, specificity, likelihood ratios and predictive values.

T – test cut off point – How will the data be dichotomized? What constitutes positive and negative result should be clearly defined for both the index and the reference test.

E – expected test use – What is the anticipated role of the index test? The index test may be developed to replace the reference test, be an additional test or be used in assessing whether patients need to undergo further test (triage).

Systematic reviewers place 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 /patient preferences and experiences. The background should avoid making value-laden statements unless they are specific to papers that illustrate the topic and/or the need for a systematic review of the topic.

Review Objectives and Questions

The objectives guide and direct the development of the specific review criteria. Clarity in the objectives and specifically in the review question assists in developing a protocol, facilitates more effective searching, and provides a structure for the development of the full review report. The objectives should be stated in full and conventionally, a statement of the overall objective is made and elements of the review are then listed as review questions. For example:

Following on from the title, the review question(s), the review should explicitly state the research aims of the review. Using the above example (White and Schultz, 2011):

> *The aim of this review is to summarize and synthesize the best available evidence on the diagnostic test accuracy of currently available laboratory tests for Influenza H1N1 (Swine Flu), as compared with viral culture reference test.*

Developing a good review question is the critical first step in undertaking a sound systematic review. The review question structures key components of the review: objectives, inclusion/exclusion criteria and the search strategy. It is most important that reviewers spend enough time on question development to arrive at a clear and explicit question that is well thought through. Again, using the above example;

Specifically, the review question for this review is:

> *What is the diagnostic accuracy of available laboratory tests for identifying Influenza A (H1N1) in patients presenting with an influenza-like illness, compared with the reference test of viral culture?*

Criteria for inclusion/exclusion

The inclusion and exclusion criteria are important in determining the scope of the review. Each of the following elements should be addressed in the protocol.

Population/Participants

Who is the population of patients or subjects of interest for this review?

> *In the above example, the population may be described as being patients presenting with influenza-like illnesses, regardless of age or gender.*

The choice of population is important and should reflect those who will undergo the test in clinical practice. Extrapolation of test results to other populations may result in over or under estimation of test accuracy and should be avoided. Specific reference to population characteristics, either for inclusion or exclusion, should be based on clear scientific justification rather than based on unsubstantiated clinical, theoretical or personal reasoning, details of which should be presented in the background section.

Index Test

What are the tests under evaluation in this review?

> *In the above example, all laboratory tests used to confer a diagnosis of Influenza A H1 N1 (Swine Flu) will be considered.*

The review should decide at the protocol stage on what criteria the tests should be assessed as being similar enough to combine in meta-analysis. This is an important factor when interpreting the results of review. Using the given example of laboratory tests for H1N1, this review

considered all PCR based tests together, making no distinction between the differences in target primer sequences. Criteria on which the index test results are categorized as being positive and negative should also be decided at the protocol stage.

Reference Test

To what test are the index tests going to be compared with? I.e. what is the best test currently available? In the example above, viral culture is used as the reference standard.

This standard was chosen on the basis of being the test most frequently to confer a diagnosis in hospital based laboratories under non-pandemic conditions. As with the index tests, the review should also describe in detail the reference test/method that will be used for comparison. The review should also decide at the protocol stage on what criteria the reference test results are categorized as being positive and negative.

Accuracy Methods

How will diagnostic test accuracy be measured? Recommended measures include tests sensitivity, specificity, likelihood ratios and predictive values. The formulae and rationale for each of these values is discussed in chapter 2 of this book. Initial accuracy measures calculated are often sensitivity and specificity. Once calculated, these measures can be used to generate other clinically relevant statistics such as predictive values and likelihood ratios. The equations and implications for each of these statistics are also given in chapter 2.

It is important to consider other context of the test accuracy data, for example prevalence of a disease has been shown to influence estimates of diagnostic test accuracy, particularly positive and negative predictive values (Leeflang et al., 2009, Van den Bruel et al., 2006).

Test Cut-Off Point

On what basis will the test data be dichotomized? What constitutes positive and negative result should be clearly defined for both the index and the reference test. Where tests may have several logical cut-off points, it may be logical to present the accuracy findings as a summary receiver operator curve. Careful consideration should be given to cut-off point selection, as test sensitivity and specificity can be affected. In our given example, the cut-off point for the index test can be clearly defined on the basis of the number of amplification cycles needed to observe a PCR product visible with particular markers. Cut-off point choice for reference tests such as viral culture may be more difficult to determine as there is a larger element of clinical judgment and the cut-off point may be implicit.

Expected Test Use

What is the anticipated role of the index test? The protocol should include rationale for development or adoption of the index test. Potential drives for a new test may include: replacement of the existing reference test, to be an additional test or be used in assessing whether patients need to undergo further test (triage). Reasons for the above may be that the new test is cheaper, faster to generate results and/or require less resources (including technical skills), than the current best available (reference) test.

Types of Studies

In most types of systematic review the type(s) of study design to be considered for inclusion are detailed in this section. Diagnostic test accuracy studies generally use cross-sectional

study designs. Restricting a database search of the basis of study design may fail to identify relevant studies that contain accuracy data but where the primary focus of the study was not diagnostic test accuracy, therefore this practice is not recommended (Bayliss and Davenport, 2008, Kastner et al., 2009, Leeflang et al., 2008b, Leeflang et al., 2006, Ritchie et al., 2007).

An important feature of study designs considered for inclusion in a systematic review is that patient samples undergo both the reference and the index test and that the tests are independent of one another.

Search strategy

As previously mentioned in chapter 3, identifying diagnostic test accuracy data can be challenging. The aim of a systematic review is to identify all relevant international research on a given topic. This is done by utilizing a well-designed search strategy across a breadth of resources. There is not sufficient evidence to conclude that a particular number of databases, or that even whether particular databases provide sufficient topic coverage, therefore, literature searching should be based on the principal of inclusiveness with the widest reasonable range of databases included that are considered appropriate to the focus of the review. If possible, authors should seek the advice of a research librarian in the construction of a search strategy.

Systematic reviews are international sources of evidence; therefore particular nuances of local context should be informed by and balanced against the best available international evidence. The protocol should provide a detailed strategy that will be used to identify all relevant international research within an agreed time frame. The timeframe for the search should be clearly stated in the protocol to allow transparency in the review. The search timeframe may be influenced by such factors as the timing of technical developments of the test or technique under review of publication of existing systematic reviews, however, potentially relevant studies as well as seminal, early studies in the field may be excluded and should thus be used with caution, the decision preferably to be endorsed by topic experts and justified in the protocol.

The search strategy should also describe any limitations to the scope of searching in terms of resources to be accessed or publication languages. In addition to this, the protocol should also specify any limitations to the search if particular research methods/methodologies will be considered for inclusion in the review (e.g. cross-sectional observational, cohort studies), data ranges and publication languages.

Limitations to the search may vary depending on the nature of the topic being reviewed, or the resources available to the review team.

The protocol should list the databases that will be searched and the initial search terms that will be used. Systematic reviews of effectiveness of interventions will often include a hierarchy of studies that will be considered, however may not be necessary for diagnostic test accuracy reviews, as the majority of the studies may be expected to be cross-sectional studies.

Typically, a search strategy is conducted in three-phases:

Stage 1 Identification of keywords and search terms

A limited search may be undertaken of major databases (such as PubMed or CINAHL) 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, search terms or

search concepts that may help in the major database search for similar papers. This is done by analyzing words contained in the title, keywords and abstract.

Stage 2 Conducting the search across the specified databases

The second phase is to construct database-specific searches 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 slight differences in how articles are indexed and terms used to describe articles within in (descriptors).

Stage 3 Hand searching

The phase is to review the reference lists of all studies that are retrieved for appraisal to search for additional studies. Researchers who are considered experts in the field of interest may also be considered as a potential source of articles and/or unpublished data. Websites of societies or organizations may also be sources of information.

The comprehensiveness of searching and the documentation of the databases searched 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 include: 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.

Assessment criteria

The basis for inclusion and exclusion of studies in a systematic review needs to be transparent and clearly documented in the protocol, as this is a key step in establishing the credibility of the review. A systematic review aims to synthesize the best available evidence; therefore the review should aim to include the highest quality of evidence possible.

The checklist approach provides a convenient and reproducible way of assessing methodological quality. A checklist provides a list of predetermined criteria that can be scored as being met, not met or unclear or not applicable. The decision as to whether or not to include a study can be made based on meeting a pre-determined proportion of all criteria, or on certain criteria being met and it is possible to weight the different criteria differently, depending on the focus and scope of the review.

The aim of determining the quality of individual studies included within a systematic review is to identify potential sources of bias and to limit the effect of those biases on the estimates and conclusions of the systematic review. The Quality Assessment of Diagnostic Accuracy studies (QUADAS) checklist (Whiting et al., 2003, Whiting et al., 2006) provides a 14 item checklist with which to critically appraise diagnostic test accuracy studies. Individual items and the areas of bias they aim to address is discussed further in chapter 3. A copy of the

checklist to be used for appraising the methodological quality for inclusion of studies in a review should be appended to the protocol.

Data extraction

Data extraction refers to the process of sourcing and recording relevant results from the original (or primary) research studies that will be included in the systematic review. It is important that both reviewers use a standard extraction tool that they have practiced using and then consistently apply. The protocol must therefore describe how data will be extracted and include an appropriate data extraction instrument in appendices to the protocol. Strategies to minimize the risk of error when extracting data from studies include:

- utilizing a standardized data extraction form;
- pilot test extraction prior to commencement of review;
- train and assess data extractors; and
- have two people extract data from each study.

Reviewers should, where necessary, contact authors of publications and seek assistance in providing raw data. Data to be extracted generally includes details such as:

- **P**articipant characteristics
- **I**ndex test details
- **R**eference test details
- **A**ccuracy measures assessed
- **T**reatment outcomes used
- **E**xpected outcomes of the test

Information that may impact upon the generalizability of the review findings such as study method, setting and population characteristics should also be extracted and reported. Population characteristics include factors such as age, past medical history, co-morbidities, complications or other potential confounders. It may also be useful to identify and extract the study authors' conclusions and establish whether there is agreement with conclusions made by the reviewer authors.

Data synthesis

The protocol should also detail how the data will be combined and reported. A summary estimate of data combined in meta-analysis is considered to be the highest level of evidence, (Joanna Briggs Institute, 2011) however, for data to be combined statistically, there are set criteria that need to be fulfilled as previously discussed (Borenstein et al., 2009, Deeks, 2001b).

The protocol should detail on what basis the decision will be made on whether or not to combine data in a meta-analysis.

- Clinical – are the patient characteristics similar? (Such as age, co-morbidities, and treatments).
- Methodological – do the studies use the same study design and measure the same outcomes?
- Statistical – were outcomes measured in the same way, at the same time points, using comparable scales?

Statistical combination of diagnostic test accuracy data provides a summary estimate of the test performance, using transparent rules specified in advance (Borenstein et al., 2009). This allows an overall assessment of the test accuracy as compared to the reference test. Whilst the ultimate aim for a systematic review is to combine study data in meta-analysis, this is not always appropriate or it is important to combine the studies in an appropriate manner using methods appropriate to the specific type and nature of data that has been extracted. In the protocol, the methods by which studies will be combined should be described in as much detail as is reasonably possible.

The challenge that faces authors of diagnostic test accuracy systematic reviews is how to handle and sensibly combine the data from studies included in the review (Gatsonis and Paliwal, 2006, Hasselblad and Hedges, 1995, Leeflang et al., 2008b). This is because data is often reported in different formats (Borenstein et al., 2009, Macaskill et al., 2010, Zhuo et al., 2002) and summarized in Table 4.

The extracted data should include enough detail to be able to populate a 2×2 table.

The protocol should also describe a process for developing a narrative summary to anticipate the possibility that meta-analysis can not be conducted. Narrative summary should draw upon the data extraction, with an emphasis on the textual summation of study characteristics as well as data relevant to the test(s) of interest.

Conflict of Interest

The protocol should state if any conflicts of interests exist for the reviewers. A statement should be included that either declares the absence of any conflict of interest, or describes a specified or potential conflict of interest. Such a conflict may exist if a review author was determining the diagnostic test accuracy of a test that they were involved in developing, for example. Conflict of interest statements should adhere to the guidelines of the International Committee of Medical Journal Editors (ICMJE) for individual authors and project support (http://www.icmje.org/ethical_4conflicts.html). Additionally, the Committee on Publication Ethics (COPE) has extensive guidelines for conflict of interest statements that are intended to protect the authors as well as the readers, and review authors should ensure they are familiar with and adhere to the principals described within the COPE framework (http://www.publicationethics.org/).

Acknowledgements

The source of any financial grants and other assistance must be acknowledged, including any commercial or affiliations of the reviewers. The contribution of colleagues or Institutions should also be acknowledged.

References

The protocol should include any references that it cites. Requirements for the use of specific referencing styles differ depending on where the systematic review is submitted for publication. The review authors should clarify which style to use with the journal.

Appendices

Appendices should be placed at the end of the protocol and be numbered in the order in which they appear in the text. At a minimum this will include critical appraisal and data

extraction tools. If there are bulky tables or swathes of text which break the flow of the protocol, such items may be moved to the appendix section.

Searching for diagnostic test studies

The comprehensiveness of searching and the documentation of the resources searched 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 include: 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.

Systematic literature searching for diagnostic test accuracy evidence presents particular challenges. Additionally, inconsistency of thesaurus terms between databases means reviewers need to be cognizant of the limitations in each database they use. Some work has been undertaken to examine how well different searching strategies identify diagnostic test research and this work suggests a combination of broad thesaurus terms and specific method terms be used to construct search strategies (Bayliss and Davenport, 2008).

Search filters

Search filters are pre-tested strategies that identify articles based on criteria such as specified words in the title, abstract and keywords. They can be of use to restrict the number of articles identified by a search from the vast amounts of literature indexed in the major medical databases. Search filters look for sources of evidence based on matching specific criteria – such as certain predefined words in the title or abstract of an article. Search filters have strengths and weaknesses:

(i) Strengths: they are easy to implement and can be pre-stored or developed as an interface

(ii) Limitations: database-specific; platform-specific; time-specific; not all empirically tested and therefore not reproducible; assume that articles are appropriately indexed by authors and databases.

Key to terms used in searching:

ab = words in abstract

exp = before an index term indicates that the term was exploded

hw = word in subject heading

mp = free text search for a term

pt = publication type

sh = subject heading

ti = words in title

tw = textwords in title/abstract

? = in middle of term indicates use of a wildcard

/ = MeSH subject heading (and includes all subheadings being selected)

$ = truncation symbol

adj = two terms where they appear adjacent to one another (so adj4, for example, is within four words)

There have been several publications examining the advantages and disadvantages of using narrow search filters to identify diagnostic test accuracy publications, (Bayliss and Davenport, 2008, Kastner et al., 2009, Leeflang et al., 2008b, Leeflang et al., 2006, Ritchie et al., 2007, Whiting et al., 2008b) however the current recommendation is to avoid using narrow filters as they will continue to miss relevant papers until there are improvements in indexing of diagnostic test accuracy publications.

Search terms that are used to index diagnostic test accuracy studies in major databases include:

- Sensitivity
- Specificity
- Diagnostic test
- Accuracy
- Likelihood ratio
- Predictive value

Such terms can be used in conjunction with the area/test of interest in constructing a search strategy.

Generic medical/science databases

The two major databases that index healthcare research are PubMed and CINAHL and they are usually the first stop when searching for healthcare research papers.

The main component of PubMed is the MEDLINE (Medical Literature Analysis and Retrieval System Online) database, which is the U.S. National Library of Medicine's (NLM) main bibliographic database. In addition to MEDLINE, PubMed also provides access to other resources that are beyond the scope of MEDLINE, such as general chemistry and biology articles. Approximately 5,200 journals published in the US and more than 80 other countries are indexed for MEDLINE. A distinctive feature of MEDLINE is that the records are indexed with NLM's controlled vocabulary, the Medical Subject Headings (MeSH®).

In addition to MEDLINE citations, PubMed also contains:

- In-process citations which provide a record for an article before it is indexed with MeSH and added to MEDLINE or converted to out-of-scope status.
- Citations that precede the date that a journal was selected for MEDLINE indexing (when supplied electronically by the publisher).
- Some old MEDLINE citations that have not yet been updated with current vocabulary and converted to MEDLINE status.

- Citations to articles that are out-of-scope (e.g., covering plate tectonics or astrophysics) from certain MEDLINE journals, primarily general science and general chemistry journals, for which the life sciences articles are indexed with MeSH for MEDLINE.
- Some life science journals that submit full text to PubMed Central® and may not yet have been recommended for inclusion in MEDLINE although they have undergone a review by NLM, and some physics journals that were part of a prototype PubMed in the early to mid-1990's.
- Citations to author manuscripts of articles published by National Institute of Health (NIH)-funded researchers.

One of the ways users can limit their retrieval to MEDLINE citations in PubMed is by selecting MEDLINE from the Subsets menu on the Limits screen.

CINAHL (Cumulative Index to Nursing & Allied Health Literature) is available via the EBSCO host and indexes journal articles, selected pamphlets, selected books, dissertations, selected conference proceedings, standards of practice for nursing specialties, audiovisuals, educational software, and selected poems and cartoons. Foreign language nursing & allied health journal articles are included when English abstracts are provided. Over 4,500 journal were indexed as of July 2010 – indexing can cover as far back as 1937. A list of journals with dates of coverage is linked from EBSCO Publishing. Searchable cited references have been included since 1994 for over 1,300 journals.

CINAHL Plus with Full Text provides full text of over 770 journals and 275 books/monographs, plus legal cases, clinical innovations, critical paths, drug records, research instruments and clinical trials.

Grouping Terms Together Using Parentheses

Parentheses (or brackets) may be used to control a search query. Without brackets, a search is executed from left to right. Words that you enclose in parentheses are searched first. Why is this important? Brackets allow you to control and define the way the search will be executed. The left phrase in parentheses is searched first; then based upon those results the second phrase in brackets is searched.

Grey or Gray Literature, Deep Web searching

Developing a Search Strategy for Grey literature

Since the mid-1980s and particularly since the explosion of the Internet and the opportunity to publish all kinds of information electronically, there has been an 'information revolution'. This revolution is making it increasingly impossible for people to read everything on any particular subject. The research field of diagnostic tests is no exception. There is such a huge amount of data being written, published and cited that Internet search engines and medical specialist databases such as MEDLINE and CINAHL, cannot hope to catalogue or index everything. There are therefore valuable sources of evidence that can prove useful when doing systematic reviews, but many have not been 'captured' by commercial electronic publishers.

Grey (or Gray – alternative spelling) literature includes documents such as:

- technical reports from government, business, or academic institutions
- conference papers and proceedings

- preprints
- theses and dissertations
- newsletters
- raw data such as census and economic results or ongoing research results

When building a search strategy for grey literature, it is important to select terms specifically for each source. In using mainstream databases, such as Mednar (including Google Scholar), it is best to draw from a list of keywords and variations developed prior to starting the search.

A consistent and systematic process, using the same keywords and strategy is recommended, as it is easy to become swamped in the volume of information.

As when searching commercial databases, it is important to create a strategy, compile a list of keywords, wildcard combinations and identify organizations that produce grey literature. If controlled vocabularies are used, record the index terms, qualifiers, keywords, truncation, and wildcards.

Searching the medical grey literature can be time-consuming because there is no 'one-stop shopping' database or search engine that indexes materials the way, for example as CINAHL does for nursing and allied health or MEDLINE does for the biomedical sciences.

MedNar

The MedNar database http://mednar.com/mednar/ indexes diagnostic test accuracy grey literature articles. MedNar is a federated search engine therefore non-indexing, designed for professional medical researchers to quickly access information from a multitude of credible sources. Researchers can take advantage of Mednar's many tools to narrow their searches, drill down into topics, remove duplicate titles, as well as rank and cluster search results.

WorldWideScience.org

WorldWideScience.org http://worldwidescience.org/index.html is a global science gateway that allows access to national and international scientific databases and portals. The WorldWideScience Alliance is a multilateral partnership and consists of participating member countries and provides the governance structure for WorldWideScience.org. This resource is relatively new (established June 2007) and includes OpenSIGLE, and links to resources in Chinese, Indian, African, and Korean languages.

It should be remembered that access to bibliographic databases may depend on subscriptions and the search interface may also vary depending on the database vendor (for example Ovid, EBSCO, ProQuest, etc) or whether you access MEDLINE via the free PubMed interface.

A research librarian is a valuable assert when determining which databases are accessible (to the particular institution) and relevant (to the particular review topics).

The following search engines are suggestions for finding health-based scientific literature:

- http://www.scirus.com
- http://www.metacrawler.com
- http://www.disref.com.au/
- http://www.hon.ch/Medhunt/Medhunt.html
- http://www.medworld_stanford.edu/medbot/

- http://http://sumsearch.uthscsa.edu/cgi-bin/SUMSearch.exe/
- http://www.intute.ac.uk/healthandlifesciences/omnilost.html
- http://www.mdchoice.com/index.asp
- http://www.science.gov/
- http://http://www.eHealthcareBot.com/
- http://http://medworld.stanford.edu/medbot/
- http://http://omnimedicalsearch.com/
- http://http://www.ingentaconnect.com/
- http://http://www.medical-zone.com/

There are numerous health information gateways or portals on the Internet containing links to well organized websites containing primary research documents, clinical guidelines, other sources and further links. For example:

- World Health Organization, http://www.who.int/library/
- Canadian Health Network, http://www.canadian-health-network.ca/customtools/homee.html
- Health Insite, http://www.healthinsite.gov.au/
- MedlinePlus, http://www.nlm.nih.gov/medlineplus
- National Guidelines Clearinghouse, http://www.guideline.gov/index.asp
- National Electronic Library for Health (UK), http://www.nelh.nhs.uk/
- Partners in Information Access for the Public Health Workforce, http://phpartners.org/guide.html

Theses and Dissertations

ProQuest Dissertations and Theses Database

With more than 2.3 million entries, the ProQuest Dissertations & Theses (PQDT) database is one of the most comprehensive collection of dissertations and theses in the world. Graduate students customarily consult the database to make sure their proposed thesis or dissertation topics have not already been written about. Students, faculty, and other researchers search it for titles related to their scholarly interests.

Dissertation Abstracts Online (DIALOG)

This is a substantial subject, title, and author guide to virtually every American dissertation accepted at an accredited institution since 1861. Selected Masters theses have been included since 1962. In addition, since 1988, the database includes citations for dissertations from 50 British universities that have been collected by and filmed at The British Document Supply Centre. Beginning with DAIC Volume 49, Number 2 (Spring 1988), citations and abstracts from Section C, Worldwide Dissertations (formerly European Dissertations), have been included in the file. Abstracts are included for doctoral records from July 1980 (Dissertation Abstracts International, Volume 41, Number 1) to the present. Abstracts are included for masters theses from Spring 1988 (Masters Abstracts, Volume 26, Number 1) to the present.

Individual, degree-granting institutions submit copies of dissertations and theses completed to University Microfilms International (UMI). Citations for these dissertations are included in the

database and in University Microfilms International print publications: Dissertation Abstracts International (DAI), American Doctoral Dissertations (ADD), Comprehensive Dissertation Index (CDI), and Masters Abstracts International (MAI). A list of cooperating institutions can be found in the preface to any volume of Comprehensive Dissertation Index, Dissertation Abstracts International, or Masters Abstracts International.

Universities, colleges, institutes, collaborative research centers (CRCs) provide a range of relevant resources and web links already listed. For example, theses or dissertations are generally included on universities' library pages as they are catalogued by library technicians according to subject heading, author, title, etc.

University library pages may also have links to other universities' theses collections, for example:

Other resources to consider

Academic Libraries' Online Public Access Catalogues (OPACS)

These catalogues provide access to local and regional materials, are sources for bibliographic verification, index dissertations as well as government and technical reports.

Other researchers or systematic reviewers working in the topic area

Potentially other researchers often already have reference lists that they are prepared to share or names of others working in the same/related fields, for example authors of Cochrane or Joanna Briggs Institute protocols that are not yet completed. This is especially useful for clinicians because they often know (or at least know of!) experts who work in their specific area of interest.

Conference Proceedings

Conference series in the area of interest are also useful sources and can generally be accessed through academic or national libraries. Many national libraries collect grey literature created in their countries under legal deposit requirements. Their catalogues are usually available on the Internet. Some also contain holdings of other libraries of that country, as in the case of the Australian National Library's Libraries Australia:

http://librariesaustralia.nla.gov.au/apps/kss.

WORLDCAT is a service that aims to link the catalogues of all major libraries under one umbrella. http://www.worldcat.org/

Newspapers/News websites

The media often reports recent medical or clinical trials and newspaper sites on the Internet may report who conducted a study, where, when, the methodology used, and the nature of the participants to assist in locating an original source.

Finding grey literature on government websites

Generally, most health-related government-sponsored or maintained websites will go to the trouble of showing:

- how or if their documents are organized alphabetically, topically or thematically;
- how individual documents are structured, i.e. contents pages, text, executive summary, etc.;
- database-type search strategies to find them;
- links to other web sites or other documents that are related to the documents that they produce;
- when their collection of grey literature has been updated; and
- documents in PDF or Microsoft Word downloadable form.

Other useful tactics include:

Setting up 'auto alerts' if possible on key databases to learn about new relevant material as it becomes available.

Joining a relevant web discussion group/list and post questions and areas of interest; contacts may identify leads to follow up.

Grey literature is increasingly referenced in journal articles, so reference lists should be checked via hand-searching. Hand searching is recommended for systematic reviews because of the hazards associated with missed studies.

Hand searching is also a method of finding recent publications not yet indexed by or cited by other researchers.

With millions of resources available on the Internet, it is difficult to find relevant and appropriate material even if you have good search skills and use advanced search engines. Issues of trust, quality, and search skills are very real and significant concerns - particularly in a learning context. Academics, teachers, students and researchers are faced with a complex environment, with different routes into numerous different resources, different user interfaces, search mechanisms and authentication processes.

Documenting a search strategy

One of the major strengths of a systematic review is the systematic approach to identifying relevant studies. An important factor in this process is documenting the search and the findings of the search. Commonly, electronic databases are used to search for papers, many such databases have indexing systems or thesauruses, which allow users to construct complex search strategies and save them as text files. 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, decisions made and consider the comprehensiveness and exhaustiveness of the search strategy for each included database. It is therefore, crucial to ensure that the results of searching each resource utilized is captured and documented.

Any planned restrictions to the search such as timeframe, number of databases searched and languages should be detailed in the systematic review protocol and the implications of these restrictions should be discussed in the discussion section of the review. Each electronic database is likely to use a different system for indexing key words within their search engines. Hence the search strategy will be tailored to each particular database. These variations are important and need to be captured and included in the systematic review report.

Managing references

Bibliographic programs such as Endnote (available from http://www.endnote.com/) can be extremely helpful in keeping track of database searches and many databases facilitate this process by generating output files that are compatible with such software – either directly (as a RIS file) or indirectly as a text (.txt) file.

Summary on Searching

In a protocol, reviewers are required to state the databases to be searched, the initial key words that will be used to develop full search strategies and other resources to be accessed (e.g. experts in the field, topic-specific websites/conferences etc)

The search strategy should also describe any limitations to the scope of searching in terms of timeframe or publication languages. Each of these factors may vary depending on the nature of the topic being reviewed or the resources available. Limiting by date may be used however, may exclude seminal early studies in the field and should thus be used with caution; the decision preferably endorsed by topic experts and justified in the protocol.

When conducting the search for diagnostic test accuracy evidence, current evidence suggests that search filters may fail to identify potentially relevant articles, mainly due to inconsistencies in how diagnostic research is currently indexed. There is insufficient evidence to suggest a particular number of databases should be included in a search for diagnostic test accuracy studies. Thus, literature searching should be based on the principal of inclusiveness, with the widest reasonable range of databases included that are considered appropriate to the focus of the review.

Assessing the Methodological Quality of Diagnostic Test Accuracy Studies

Selecting studies

When the search for evidence is complete (or as the search progresses in some cases) reviewers decide which papers found should be retrieved and then subjected to critical appraisal. This initial process is referred to as the selection of papers for appraisal. All selected papers are then subjected to critical appraisal to determine methodological quality.

Study selection is an initial assessment that occurs following the review search addressing the simple question: “should the paper be retrieved?” Studies in a review will also undergo another ‘round’ of selection in the next systematic step in the review process. This second round of assessment asks a different question: “should the study be included in the review?” which is the critical appraisal. Study selection is performed with the aim of selecting only those studies that address the review question and that match the inclusion criteria documented in the protocol of your review. Two assessors, to limit the risk of error, should perform the process. Both assessors will scan the lists of titles, and if necessary abstracts, to determine if the full text of the reference should be retrieved. Sometimes it will be difficult or impossible to determine if the reference matches the inclusion criteria of the review on the basis of the title or abstract alone; in this case the full text should be retrieved for further clarification. It is best to err on the side of caution in this process. It is better to spend a bit more time here, in careful consideration, rather than risk missing important and relevant evidence related to the

review question. The entire process must be transparent and clear so that if an independent person were to apply the same inclusion criteria to the same list of citations, they would arrive at the same result of included studies.

Assessment of methodological quality/critical appraisal

Once the search is complete and articles that are deemed to meet the pre-defined inclusion/exclusion criteria have been identified, the next stage of conducting a systematic review is to determine the methodological quality of those studies. A systematic review aims to synthesize the best available evidence; therefore the review should aim to include the highest quality of evidence possible.

The aim of determining the quality of individual studies included within a systematic review is to identify potential sources of bias and to limit the effect of those biases on the estimates and conclusions of the systematic review. Using pre-defined criteria, such as that provided by the QUADAS checklist (Whiting et al., 2006) predetermined can be scored as being met, not met or unclear or not applicable. The decision as to whether or not to include a study can be made based on meeting a pre-determined proportion of all criteria, or on certain criteria being met, as detailed in the protocol. The QUADAS checklist (Whiting et al., 2003, Whiting et al., 2006) is a 14 item checklist tool that can be used to critically appraise diagnostic test accuracy studies.

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 to the needs of the review in terms of the specific study characteristics such as cut-off point and definitions of what constitutes positive and negative results. The reviewers should be clear on what constitutes acceptable levels of information to allocate a positive appraisal compared with a negative, or response of "unclear". This discussion should take place before independently conducting the appraisal. A copy of the QUADAS checklist items can be found in Appendix I.

Extracting Data from Diagnostic Test Accuracy Studies

The protocol should detail the data to be extracted from the included studies. There is no recommended data extraction instrument, although the STARD checklist has been used as a data extraction tool (White and Schultz, 2011). In addition to the 2 × 2 table (Table 2), other data to be extracted may include elements such as that detailed below in Table 6.

Table 6. Data extraction tool

Participants	Index Test	Reference Test	Sensitivity	Specificity	Setting

Data synthesis

The protocol should detail how the extracted data is to be synthesized. If the data is heterogeneous it should be presented in narrative summary and potentially sources of heterogeneity should be discussed (e.g. clinical, methodological or statistical) as well as on what basis it was determined inappropriate to combine the data statistically (such as differences in populations, study designs or by Chi square or I^2 tests).

Where meta-analysis is appropriate, the model used (BMR or HSROC) and the statistical methods and the software used should be described.

Chapter 5:

Systematic Review Reports of Diagnostic Test Accuracy Evidence

The systematic review protocol details how the review will be conducted, what tests are outcomes of interest and how the data will be presented. The systematic review report should be the follow up to an approved protocol - any deviations from the protocol need to be clearly detailed in the report, to maintain transparency. The following headings offer a detailed framework for the necessary sections of a report. The report should also be consistent with the PRISMA statement and include a completed PRISMA checklist (PRISMA, 2011).

Title of Systematic Review

The title should be the same as detailed in the protocol.

Review Authors

The names, contact details and any affiliations should be listed for each reviewer. Generally, each systematic review of diagnostic test accuracy has a minimum of two reviewers. The name and contact details (physical address and email address) of the corresponding author should be included in this section.

Executive Summary/Abstract

The executive summary or abstract should be a short summary (approximately 500 words) of the systematic review. The purpose of the executive summary is to allow abstraction and indexing of the review so it is important that it accurately reflects the purpose, basic procedures, main findings and principal conclusions of the study.

Background

In the systematic review report, as for the protocol, the background section should describe the test under evaluation, the reference test, the target population, and outcomes that are documented in the literature. The background should be an overview of the main elements and the acronym PIRATE may be of assistance, as previously discussed in section 1. The background section should provide sufficient detail to justify why the review was conducted and the choice of the various elements. It is often as important to justify why other elements related to the area under review are not included. The Joanna Briggs Institute suggests that a background should be a minimum of 1000 words.

Review objectives

The review objectives should be explicitly stated as per the protocol and should be presented cognizant of the mnemonic PIRATE to help cover as much of the area of the review as possible.

Criteria for Considering Studies for this Review

Population/types of participants

The report should provide details about the type participants included in the review. Useful details include: age – mean and the range, condition/diagnosis or health care issue, administration of medication. Details of where the studies were conducted (e.g. rural/urban setting and country) should also be included. Again the decisions about the types of participants should have been explained in the background.

Types of tests

The details of both the index and reference test included in the review should be presented in this section as detailed in the protocol.

Types of accuracy measures

This section of the review report should be a list of the accuracy measures considered such as sensitivity, specificity, ROCs, likelihood values and predictive values as detailed in the protocol.

Types of studies

As per the protocol section, the types of studies that were considered for the review should be included. There should be a statement about the target study type and whether or not this type was not found. The types of study identified by the search and those included should be detailed in the report.

Search Strategy

This section of the review report should provide an overview of the search strategy. Often an example of a search strategy specific to a particular database is included as an appendix to the review. This section should detail search activity (e.g. databases searched, initial search terms and any restrictions) for the review, as predetermined in the protocol. 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, decisions made and consider the comprehensiveness and exhaustiveness of the search strategy for each included database. Any restrictions to the search such as timeframe, number of databases searched and languages should be reported in this section of the report and any limitations or implications of these restrictions should be discussed in the discussion section of the review.

Assessment of methodological quality

A description of how methodological assessment was determined should be included, with reference to the critical appraisal tool(s) used. This section should also provide a description of which components of the QUADAS items were used in developing the appraisal tool; and explanations as to why these items were selected. A copy of the tool(s) should be included in the appendix section. Specific study characteristics such as cut-off point and definitions of what constitutes positive and negative results should also be included in this section.

Data extraction

This section of the report should include details of the types of data extracted from the included studies, as predetermined in protocol. Information that may impact upon the generalizability of the review findings such as study method, setting and population characteristics should also be extracted and reported. Population characteristics include factors such as age, past medical history, co-morbidities, complications or other potential confounders. Mention should also be made of the standardized data extraction instrument and a copy placed in the appendix.

Data synthesis

This section should describe how the extracted data was synthesized. If the data was heterogeneous and is presented as a narrative summary, potentially sources of heterogeneity should be discussed (e.g. clinical, methodological or statistical) as well as on what basis it was determined inappropriate to combine the data statistically (such as differences in populations, study designs or by Chi square or I^2 tests). Where meta-analysis was used, the statistical methods and the software used (such as The Cochrane Collaborations RevMan) should be described.

Review Results

Description of studies

The type and number of papers identified by the search strategy and the number of papers that were included and excluded should be stated. The major stages of study selection are suggested below:

- Numbers of studies identified
- Numbers of studies retrieved for detailed examination
- Numbers of studies excluded on the basis of title and abstract
- Numbers of full text articles retrieved
- Numbers of studies excluded on the basis of full text
- Numbers of appraised studies
- Numbers of studies excluded following critical appraisal and an overview of reasons for exclusion
- Numbers of studies included in the review

Details of all of the full text articles that were retrieved but were not included in the review should be provided in a table of excluded studies, as an appendix to the review. The reasons for exclusion should also be detailed.

Methodological quality of the included studies

The overall methodological quality of the included studies should be discussed in this section, as determined by results of the critical appraisal. The checklist items form the basis of determining the quality of the included papers. Graphs and figures describing the outcomes of methodological quality assessment of studies involved in the review should be included in the text of the report or as an appendix, as appropriate.

Results – findings of the review

This section should be organized in a meaningful way based on the objectives of the review and the criteria for considering studies. There is no standardized international approach to how review findings are structured or how the findings of reviews ought to be reported. It would be logical however, to present findings in the same order as the review questions and/or review objectives. The audience for the review should be considered when structuring and presenting the review findings. The role of tables and appendices should not be overlooked. Adding extensive detail on studies in the results section may "crowd" the findings, making them less accessible to readers, hence using tables, graphs and in-text reference to specific appendices is encouraged.

Given there is no clear international standard or agreement on the structure or key components of this section of a review report, and the level of variation evident in published systematic reviews, the advice here is general in nature. In general, findings are discussed textually and then supported with meta-graphs, tables, figures as appropriate. The focus should be on presenting information in a clear and concise manner. Any large or complex diagrams/tables/figures should be included as appendices so as not to break the flow of the text.

Interpretation of the results

This section should focus the discussion of the results in light of the objectives of the review. The section should debate how the review findings will influence the course of diagnosis in the area in which the review was performed. Contemporary evidence should be brought in where necessary to make comparisons or direct the debate on relevance of the review results. The effects of the review findings on the field of diagnosis related to the DTA under review, as well as its influence on patients/subjects and other relevant issues should be discussed under this heading.

Conclusions

The discussion should include an overview of the results and it should address issues arising from the conduct of the review including limitations and issues arising from the results of the review. Conclusions should focus on the implications for practice and for research. These should be detailed and must be based on the documented results, not author opinion. Where evidence is of a sufficient level, appropriate recommendations should also be made. Recommendations must be clear, concise and unambiguous.

Implications for practice

Implications for practice should be detailed and based on the documented results, not reviewer opinion. In qualitative reviews, recommendations are declamatory statements that are steeped in context. Therefore, generalizability occurs between cases rather than across broad populations. Recommendations must be clear, concise and unambiguous.

Implications for research

All implications for research must be derived from the results of the review, based on identified gaps in the literature or on areas of weakness, such as methodological weaknesses.

Implications for research should avoid generalized statements calling for further research, but should be linked to specific issues.

Appendices

Again, as in the initial protocol, the final review report should include references and appendices. The references should be appropriate in content and volume and include background references and studies from the initial search. The appendices should include:

- Critical appraisal form(s)
- Data extraction form(s)
- Table of included studies
- Table of excluded studies with justification for exclusion

Conflict of interest

A statement should be included in every systematic review report that either declares the absence of any conflict of interest or describes a specified or potential conflict of interest. Conflict of interest statements should adhere to the guidelines of the International Committee of Medical Journal Editors (ICMJE) for individual authors and project support (http://www.icmje.org/ethical_4conflicts.html). Additionally, the Committee on Publication Ethics (COPE) have extensive guidelines for conflict of interest statements that are intended to protect the authors as well as the readers, and review authors should ensure they are familiar with and adhere to the principals described within the COPE framework (http://www.publicationethics.org/).

Acknowledgements

The role of institutions, sponsors, persons and any other significant contributions made towards the accomplishment of the systematic review should be acknowledged in this section.

References

Arends, L. R., Hamza, T. H., Van Houwelingen, J. C., Heijenbrok-Kal, M. H., Hunink, M. G. & Stijnen, T. 2008. Bivariate random effects meta-analysis of ROC curves. *Med Decis Making,* 28, 621–38.

Bayliss, S. E. & Davenport, C. 2008. Locating systematic reviews of test accuracy studies: how five specialist review databases measure up. *Int J Technol Assess Health Care,* 24, 403–11.

Begg, C. B. 2005. Systematic reviews of diagnostic accuracy studies require study by study examination: first for heterogeneity, and then for sources of heterogeneity. *Journal of Clinical Epidemiology,* 58, 865–866.

Begg, C. B. 2008. Meta-analysis methods for diagnostic accuracy. *J Clin Epidemiol,* 61, 1081–2; discussion 1083–4.

Boehme, C. C., Nabeta, P., Hillemann, D., Nicol, M. P., Shenai, S., Krapp, F., Allen, J., Tahirli, R., Blakemore, R., Rustomjee, R., Milovic, A., Jones, M., O'brien, S. M., Persing, D. H., Ruesch-Gerdes, S., Gotuzzo, E., Rodrigues, C., Alland, D. & Perkins, M. D. 2010. Rapid molecular detection of tuberculosis and rifampin resistance. *New England Journal of Medicine,* 363, 1005–1015.

Borenstein, M., Hedges, L., Higgins, J. & Rothstein, H. 2009. *Introduction to meta-analysis,* Chipenham, Wiley.

Bossuyt, P., Reitsma, J., Bruns, D., Gatsonis, C., Gatsonis, C., Irwig, L., Lijmer, J., Moher, D., Rennie, D. & De Vet, H. 2003. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *Annals of Internal Medicine,* 138, 40–44.

Brazzelli, M., Sandercock, P. A. G., Chappell, F. M., Celani, M. G., Righetti, E., Arestis, N., Wardlaw, J. M. & Deeks, J. J. 2009. Magnetic resonance imaging versus computed tomography for detection of acute vascular lesions in patients presenting with stroke symptoms. *Cochrane Database of Systematic Reviews* [Online]. Available: http://www.mrw.interscience.wiley.com/cochrane/clsysrev/articles/CD007424/frame.html.

Castelli, F., Caligaris, S., Gulletta, M., El-Hamad, I., Scolari, C., Chatel, G. & Carosi, G. 1999. Malaria in migrants. *Parassitologia,* 41, 261–265.

Chalmers, I., Hedges, L. & Cooper, H. 2002. A brief history of research synthesis. *Evaluation and the Health Professional,* 25, 12–37.

Chalmers, T. C. & Lau, J. 1993. Meta-analytic stimulus for changes in clinical trials. *Stat Methods Med Res,* 2, 161–72.

Chappell, F. M., Raab, G. M. & Wardlaw, J. M. 2009. When are summary ROC curves appropriate for diagnostic meta-analyses? *Stat Med,* 28, 2653–68.

CMA. 2010. *Statistical software for meta-analysis* [Online]. Comprehensive meta-analysis. Available: www.Meta-Analysis.com.

Courtney, M. 2004. *Evidence for nursing practice*, Elsevier Churchill Livingstone.

Deeks, J. 2001a. Systematic reviews of evaluations of diagnostic and screening tests. *In:* Egger, M., Davey Smith, G. & Altman, D. (eds.) *Systematic reviews in healthcare: Meta-analysis in context.* BMJ publishing group.

Deeks, J. J. 2001b. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. *BMJ,* 323, 157–62.

Deeks, J. 2004. Diagnostic tests 4: likelihood ratios. *British Medical Journal,* 329, 168–169.

Deeks, J. J. H., J. P. T. & Altman, D. G. 2008. Chapter 9. Analysing data and undertaking meta-analyses. *In:* Higgins, J. P. T. & Green, S. (eds.) *Cochrane Handbook for Systematic Reviews of Interventions.* Chichester John Wiley & Sons.

Detrano, R., Janosi, A., Lyons, K. P., Marcondes, G., Abbassi, N. & Froelicher, V. F. 1988. Factors affecting sensitivity and specificity of a diagnostic test: the exercise thallium scintigram. *Am J Med,* 84, 699–710.

Deville, W. L., Buntinx, F., Bouter, L. M., Montori, V. M., De Vet, H. C., Van Der Windt, D. A. & Bezemer, P. D. 2002. Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC Med Res Methodol,* 2, 9.

Diagnostic Test Accuracy Working Group. 2011. *Diagnostic Test Accuracy Working Group* [Online]. Diagnostic Test Accuracy Working Group. Available: http://srdta.cochrane.org/more-about-us.

Dinnes, J., Deeks, J., Kirby, J. & Roderick, P. 2005. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. *Health Technol Assess,* 9, 1–113, iii.

Garg, A. X., Iansavichus, A. V., Wilczynski, N. L., Kastner, M., Baier, L. A., Shariff, S. Z.,

Rehman, F., Weir, M., Mckibbon, K. A. & Haynes, R. B. 2009. Filtering Medline for a clinical discipline: diagnostic test assessment framework. *BMJ,* 339, b3435.

Gatsonis, C. & Paliwal, P. 2006. Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer. *AJR,* 187, 271–281.

Glas, A., Lijmer, J., Prins, M., Bonsel, G. & Bossuyt, P. 2003. The diagnostic odds ratio: a single indicator of test performance. *Journal of Clinical Epidemiology* 56.

Glass, G. 1976. Primary, secondary and meta-analysis of research. *Education Research,* 10, 3–8.

Gordis, L. 2000. *Epidemiology,* Philadelphia, WB Saunders.

Grant, M. & Booth, A. 2009. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal,* 26, 91–108.

Grimes, D. & Schulz, K. 2005. Refining clinical diagnosis with likelihood ratios. *Lancet,* 365, 1500–1505.

Guyatt, G. H., Tugwell, P. X., Feeny, D. H., Haynes, R. B. & Drummond, M. 1986. A framework for clinical evaluation of diagnostic technologies. *CMAJ,* 134, 587–94.

Harbord, R. M., Deeks, J. J., Egger, M., Whiting, P. & Sterne, J. A. 2007. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics,* 8, 239–51.

Harbord, R., Whiting, P., Egger, M., Deeks, J. J., Shang, A., Bachmann, L. M. & Sterne, J. A. C. 2008a. Response to commentary: dealing with heterogeneity in meta-analyses of diagnostic test accuracy. *Journal of Clinical Epidemiology,* 61, 1083–1084.

Harbord, R. M., Whiting, P., Sterne, J. A., Egger, M., Deeks, J. J., Shang, A. & Bachmann, L. M. 2008b. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. *J Clin Epidemiol,* 61, 1095–103.

Hasselblad, V. & Hedges, L. V. 1995. Meta-analysis of screening and diagnostic tests. *Psychol Bull,* 117, 167–78.

Hatala, R., Keitz, S., Wyer, P. & Guyatt, G. 2005. Tips for learners of evidence-based medicine: 4. Assessing heterogeneity of primary studies in systematic reviews and whether to combine their results. *CMAJ,* 172, 661–665.

Haynes, R. & Wilczynski, N. 2004. Opitmal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. *British Medical Journal*, doi10.1136/bmj.38068.557998.EE.

Higgins, J. P. T. & Thompson, S. G. 2004. Controlling the risk of spurious findings from meta-regression. *Stat Med,* 23, 1663–82.

Honest, H. & Khan, K. S. 2002. Reporting of measures of accuracy in systematic reviews of diagnostic literature. *BMC Health Serv Res,* 2, 4.

Irwig, L., Macaskill, P., Glasziou, P. P. & Fahey, M. 1995. Meta-analytic methods for diagnostic test accuracy. *J Clin Epidemiol,* 48, 119–130.

Irwig, L., Tosteson, A. N. A., Gatsonis, C., Lau, J., Colditz, G., Chalmers, T. C. & Mosteller, F. 1994. Guidelines for meta-analyses evaluating diagnostic tests. *Annals of Internal Medicine,* 120, 667–676.

Joanna Briggs Institute 2011. *The Reviewers manual,* Adelaide.

Johnson, R. L. & Bungo, M. W. 1983. The diagnostic accuracy of exercise electrocardiography–a review. *Aviat Space Environ Med,* 54, 150–7.

Jones, C. M., Ashrafian, H., Darzi, A. & Athanasiou, T. 2010. Guidelines for diagnostic tests and diagnostic accuracy in surgical research. *J Invest Surg,* 23, 57–65.

Kastner, M., Wilczynski, N. L., Mckibbon, A. K., Garg, A. X. & Haynes, R. B. 2009. Diagnostic test systematic reviews: bibliographic search filters ("Clinical Queries") for diagnostic accuracy studies perform well. *J Clin Epidemiol,* 62, 974–81.

Knottnerus, J. A. & Van Weel, C. 2002. General introduction: evaluation of diagnostic procedures. *In:* Knottnerus, J. A. (ed.) *The Evidence Base of Clinical Diagnosis.* London: BMJ Books.

L'abbe, K. A., Detsky, A. S. & O'rourke, K. 1987. Meta-analysis in clinical research. *Ann Intern Med,* 107.

Leeflang, M. M., Scholten, R. J., Rutjes, A. W., Reitsma, J. B. & Bossuyt, P. M. 2006. Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. *J Clin Epidemiol,* 59, 234–40.

Leeflang, M., Reitsma, J., Scholten, R., Rutjes, A., Dinisio, M., Deeks, J. & Bossuyt, P. 2007. Impact of adjustment for quality on results of meta-analyses of diagnostic accuracy. *Clinical Chemistry,* 53, 164–172.

Leeflang, M., Debets-Ossenkopp, Y. J., Visser Caroline, E., Scholten, R. J. P. M., Hooft, L., Bijlmer, H. A., Reitsma, J. B., Bossuyt, P. M. M. & Vandenbroucke-Grauls, C., M. 2008a. Galactomannan detection for invasive aspergillosis in immunocompromized patients. *Cochrane*

Database of Systematic Reviews [Online]. Available: http://www.mrw.interscience.wiley.com/cochrane/clsysrev/articles/CD007394/frame.html.

Leeflang, M. M., Deeks, J. J., Gatsonis, C. & Bossuyt, P. M. 2008b. Systematic reviews of diagnostic test accuracy. *Ann Intern Med,* 149, 889–97.

Leeflang, M. M. G., Bossuyt, P. M. M. & Irwig, L. 2009. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. *Journal of Clinical Epidemiology,* 62, 5–12.

Lijmer, J. G., Bossuyt, P. M. & Heisterkamp, S. H. 2002. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Stat Med,* 21, 1525–37.

Littenberg, B. & Moses, L. E. 1993. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Med Decis Making,* 13, 313–21.

Lohr, K. 1999. Assessing "best evidence": issues in grading the quality of studies for systematic reviews. *The joint commission journal on quality improvement* 25, 470–479.

Macaskill, P. 2004. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *J Clin Epidemiol,* 57, 925–32.

Macaskill, P., Gatsonis, C., Deeks, J. J., Harbord, R. M. & Takwoingi, Y. 2010. Chapter 10: Analysing and Presenting Results. *In:* Deeks, J. J., Bossuyt, P. M. & Gatsonis, C. (eds.) *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.* The Cochrane Collaboration.

Mallett, S., Deeks, J. J., Halligan, S., Hopewell, S., Cornelius, V. & Altman, D. G. 2006. Systematic reviews of diagnostic tests in cancer: review of methods and reporting. *BMJ,* 333, 413.

Meyer, G. 2003. Guidelines for reporting information in studies of diagnostic accuracy: The STARD initiative. *Journal of Personality Assessment,* 81, 191–193.

Midgette, A. S., Stukel, T. A. & Littenberg, B. 1993. A meta-analytic method for summarizing diagnostic test performances: receiver-operatingcharacteristic-summary point estimates. *Med Decis Making,* 13, 253–7.

Moses, L. E., Shapiro, D. & Littenberg, B. 1993. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med,* 12, 1293–316.

Mower, W. 1999. Evaluating Bias and Variability in Diagnostic Test Reports. *Annals of Emergency Medicine,* 33, 85–91.

Pai, M., Mcculloch, M., Enanoria, W. & Colford, J. M. J. 2004. Systematic reviews of diagnostic test evaluations: what's behind the scenes? *Evid Based Med,* 9, 101–103.

Parekh-Bhurke, S., Kwok, C. S., Pang, C., Hooper, L., Loke, Y. K., Ryder, J. J., Sutton, A. J., Hing, C. B., Harvey, I. & Song, F. 2011. Uptake of methods to deal with publication bias in systematic reviews has increased over time, but there is still much scope for improvement. *J Clin Epidemiol,* 64, 349–57.

PRISMA. 2011. *PRISMA- Transparent reporting of systematic reviews and meta analysis* [Online]. Available: http://www.prisma-statement.org/.

Raj, A. & De Verteuil, R. 2011. Systematic review of the diagnostic accuracy of the single, two and three field digital retinal photography for screening diabetic retinopathy *The Joanna Briggs Institute Library of Systematic Reviews,* 9, 491–537.

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

Reitsma, J. B., Glas, A. S., Rutjes, A. W., Scholten, R. J., Bossuyt, P. M. & Zwinderman, A. H. 2005. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol,* 58, 982–90.

Riley, R. D., Abrams, K. R., Sutton, A. J., Lambert, P. C. & Thompson, J. R. 2007. Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Med Res Methodol,* 7.

Ritchie, G., Glanville, J. & Lefebvre, C. 2007. Do published search filters to identify diagnostic test accuracy studies perform adequately? *Health Info Libr J,* 24, 188–92.

Rutter, C. M. & Gatsonis, C. A. 2001. A hierarchical regression approach to metaanalysis of diagnostic test accuracy evaluations. *Stat Med,* 20, 2865–2884.

Shapiro, D. E. 1995. Issues in combining independent estimates of sensitivity and specificity of a diagnostic test. *Acad Radiol,* 2.

Tatsioni, A., Zarin, D., Aronson, N., Samson, D., Flamm, C., Schmid, C. & Lau, J. 2005. Challenges in systematic reviews of diagnostic technologies. *Annals of internal medicine,* 142, 1048–1055.

Thacker, S. B. & Berkelman, R. L. 1986. Assessing the diagnostic accuracy and efficacy of selected antepartum fetal surveillance techniques. *Obstet Gynecol Surv,* 41, 121–41.

The Cochrane Collaboration. 2011. *The Cochrane Library* [Online]. The Cochrane Collaboration. Available: http://www.thecochranelibrary.com/view/0/index.html.

Thompson, S. 1994. Why sources of heterogeneity in meta-analysis should be investigated. . *British Medical Journal,* 309, 1351–1355.

Thompson, S. G. & Higgins, J. P. T. 2002. How should meta-regression analyses be undertaken and interpreted? *Stat Med* 21, 1559–73.

Tricco, A. C., Tetzlaff, J. & Moher, D. 2010. The art and science of knowledge synthesis. *Journal of Clinical Epidemiology,* In Press, Corrected Proof.

Vamvakas, E. C. 1998. Meta-analyses of studies of the diagnostic accuracy of laboratory tests: a review of the concepts and methods. *Arch Pathol Lab Med,* 122, 675–86.

Vamvakas, E. C. 2001. Applications of meta-analysis in pathology practice. *Am J Clin Pathol,* 116, s47–s64.

Van Den Bruel, A., Aertgeerts, B. & Buntinx, F. 2006. Results of diagnostic accuracy studies are not always validated. *Journal of Clinical Epidemiology,* 59, 559.e1–559.e9.

Van Der Windt, D. A. W. M., Simons, E., Riphagen, I. I., Ammendolia, C., Verhagen, A. P., Laslett, M., Devillé, W., Deyo Rick, A., Bouter Lex, M., De Vet Henrica, C. W. & Aertgeerts, B. 2011. Physical examination for lumbar radiculopathy due to disc herniation in patients with low-back pain. *Cochrane Database of Systematic Reviews* [Online]. Available: http://www.mrw.interscience.wiley.com/cochrane/clsysrev/articles/CD007431/frame.html.

Verde, P. E. 2010. Meta-analysis of diagnostic test data: a bivariate Bayesian modeling approach. *Stat Med,* 29, 3088–102.

Virgili, G., Conti, A., Murro, V., Gensini, G. & Gusinu, R. 2009. Systematic reviews of diagnostic test accuracy and the Cochrane collaboration. *Internal Emerg Medicine,* 4, 255–258.

Walter, S. D. 2002. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Stat Med,* 21, 1237–56.

Walter, S. D. & Jadad, A. R. 1999. Meta-analysis of screening data: a survey of the literature. *Stat Med,* 18, 3409–24.

Wang, F. & Gatsonis, C. A. 2008. Hierarchical models for ROC curve summary measures: design and analysis of multi-reader, multimodality studies of medical tests. *Stat Med,* 27, 243–56.

White, S. & Schultz, T. 2011. *Diagnostic test accuracy of Influenza A H1N1 (Swine Flu) testing; A systematic review.* Masters in Clinical Sciences, University of Adelaide.

White, S., Schultz, T. & Tufanaru, C. 2011. Diagnostic test accuracy of Influenza A H1N1 (Swine Flu) testing; A systematic review. *The Joanna Briggs Institute Library of Systematic Reviews,* (In Press).

Whiting, P., Harbord, R., De Salis, I., Egger, M. & Sterne, J. 2008a. Evidence-based diagnosis. *J Health Serv Res Policy,* 13 Suppl 3, 57–63.

Whiting, P., Rutjes, A., Reitsma, J., Bossuyt, P. & Kleijnen, J. 2003. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology,* 3, 1–13.

Whiting, P., Westwood, M., Burke, M., Sterne, J. & Glanville, J. 2008b. Systematic reviews of test accuracy should search a range of databases to identify primary studies. *J Clin Epidemiol,* 61, 357–364.

Whiting, P., Weswood, M., Rutjes, A., Reitsma, J., Bossuyt, P. & Kleijnen, J. 2006. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Medical Research Methodology,* 6, 1–8.

Willis, B. H. & Quigley, M. 2011. Uptake of newer methodological developments and the deployment of meta-analysis in diagnostic test research: a systematic review. *BMC Med Res Methodol,* 11, 27.

World Health Organization. 2010. WHO endorses new rapid tuberculosis test. Available: http://www.who.int/tb/features_archive/new_rapid_test/en/index.html.

Zhuo, X., Obuchowski, N. & Mcclish, D. 2002. *Statistical methods in diagnostic medicine,* New York, Wiley.

Appendices

Appendix I The QUADAS Critical Appraisal Checklist

1. Was the spectrum of patients representative of the patients who will receive the test in practice? (Representative spectrum)
2. Were selection criteria clearly described?
3. Is the reference standard likely to correctly classify the target condition? (Acceptable reference standard)
4. Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests? (Acceptable delay between tests)
5. Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis? (Partial verification avoided)
6. Did patients receive the same reference standard regardless of the index test result? (Differential verification avoided)
7. Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)? (Incorporation avoided)
8. Was the execution of the index test described in sufficient detail to permit replication of the test?
9. Was the execution of the reference standard described in sufficient detail to permit its replication?
10. Were the index test results interpreted without knowledge of the results of the reference standard? (Reference standard results blinded)
11. Were the reference standard results interpreted without knowledge of the results of the index test? (Index test results blinded)
12. Were the same clinical data available when test results were interpreted as would be available when the test is used in practice? (Relevant clinical information)
13. Were un-interpretable/intermediate test results reported? (Uninterpretable results reported)
14. Were withdrawals from the study explained? (Withdrawals explained)

Appendix II Potential additional quality items

1. Were cut-off values established before the study was started?
2. Is the technology of the index test unchanged since the study was carried out?
3. Did the study provide a clear definition of what was considered to be a “positive” result?
4. Had test operators had appropriate training?
5. Was treatment withheld until both the index test and reference standard were performed?
6. Were data on observer variation reported and within an acceptable range?
7. Were data on observer variation reported and within an acceptable range?
8. Were objectives pre-specified?
9. Was the study free of commercial funding?