UNIVERSITY OF SOUTHAMPTON

FACULTY OF MEDICINE

Community Clinical Sciences

The influence of spectrum (case-mix) on diagnostic test accuracy

by

Jacqueline Dinnes

Thesis for the degree of Doctor of Philosophy

June 2008

UNIVERSITY OF SOUTHAMPTON
ABSTRACT
FACULTY OF MEDICINE
COMMUNITY CLINICAL SCIENCES
Doctor of Philosophy
The influence of spectrum (case-mix) on diagnostic test accuracy
By Jacqueline Dinnes

This thesis assessed the degree to which the technique of meta-analysis can provide insight into spectrum effects through comparing study results between studies (or between subgroups within studies). *Chapter 1* introduced the concept of diagnostic accuracy as the means by which diagnostic tests can be evaluated and also introduced the idea that diagnostic tests can operate differently according to spectrum-related characteristics. It was hypothesised that meta-analysis may provide the best available tool to identify the extent to which various the sources of heterogeneity, including spectrum, can affect test accuracy. *Chapter 2* explained four methods of meta-analysis that allow for variability in threshold and for variation in DOR with threshold. Only the so-called 'advanced' models possess the characteristics of an 'optimal' meta-analytic method.

Chapter 3 reported a methodological review of how heterogeneity has been examined in existing systematic reviews of diagnostic test accuracy. Less than optimal methods of meta-analysis that do not allow for threshold effects have been commonly employed. Spectrum-related variables were commonly investigated as potential sources of heterogeneity and 'statistically significant' results often reported. The few reviews using the advanced models of meta-analysis showed overall improved systematic review methods and were more likely to have considered spectrum-related characteristics.

Chapter 4 reported a detailed case study comparing the four meta-analytic methods on a large dataset of tests for the detection of tuberculosis. The main observations arsing from these analyses were further explored in *Chapter 5* using data obtained from a large sample of previously published systematic reviews of diagnostic tests and using only spectrum-related covariates. The main findings were as follows:

1. On average, weighting the Moses model by the inverse variance of the log of the DOR (SE(lnDOR)) underestimated the results of the unweighted Moses model by around 30%, with considerable disagreement between models. This underestimation is likely due to bias in the SE(lnDOR) and hence it is likely that the weighted model results are misleading. The circumstances that lead to biased SE(lnDOR) are common in diagnostic test meta-analyses therefore this form of weighting is not recommended.
2. The unweighted Moses model results were more similar to those of the HSROC model than those of the weighted Moses model, however it cannot be relied upon to approximate the results of the 'optimal' HSROC model.
3. The BVN model and the HSROC model produce almost identical results for the primary data analyses (this was investigated only in Chapter 4)
4. For the HSROC model, allowing for differences in the distribution of test results between diseased and nondiseased by covariate (shape differences) sometimes affects the conclusions that would be drawn from an analysis and sometimes not. Although the magnitude of differences between groups may vary between models, the inclusion of a shape interaction term does not necessarily change the strength of evidence for differences in accuracy. It is not clear whether potential differences in the distributions of test results (differences in shape) should be routinely modelled or whether the more simple parallel curve approach will generally suffice. The optimal approach for the investigation of heterogeneity requires further investigation
5. Finally, strong evidence of effects from spectrum-related characteristics on at least one model parameter were identified by the parallel or crossing curve HSROC model for over half of the investigations conducted in Chapter 5 (32/50). This could have considerable implications for the use of tests in practice.

The advanced methods of meta-analysis show promise in enabling the detection of clinically important spectrum effects. However, one of the ongoing challenges in the investigation of heterogeneity, and especially spectrum, in systematic reviews are limitations in the primary study data.

Acknowledgements

Completion of this thesis has taken more than seven years and I have several key people to thank for helping me get to the end of it. My interest in the methodology of evaluating diagnostic tests began during my time at the Centre for Reviews and Dissemination at the University of York and I have my then employer, Jos Kleijnen, to thank for stimulating my interest in this area. The persuasiveness of Norman Waugh encouraged me to move to the Wessex Institute for Heath Research and Dissemination at Southampton University in 2000. It is thanks to Norman that I got into the field of systematic reviews six years previously and also primarily due to him that I was able to enrol for my PhD, at a time when no-one else in the department were studying for PhDs. His continued support and faith in me during the years that we have worked together are very valuable to me.

I am indebted to my supervisers, Paul Roderick and Jon Deeks, for their time, patience, guidance and constant encouragement over the course of this thesis. With their input the thesis took on a whole new, and ultimately more challenging, direction. Paul's clinical perspective really helped to make me think through the implications of some of the more 'technical' findings. My move from the Wessex Institute to his department and his support of my later move to Scotland were instrumental in allowing me to finish the thesis and I sincerely thank him for all his support.

Without Jon's expertise and in-depth knowledge of the field I would not have been able to accomplish half of the work contained here and not to nearly the standard that has been achieved. I am very lucky to have worked with him over the years and hope that will continue in the future. I would also like to thank Michele, Jessie and Oscar for their hospitality during my visits from Scotland!

I gratefully acknowledge the help of my other co-authors on various publications related to the thesis: Jo Kirby for the review of previously published reviews, Christine Clar and Andrea Gibson for help with data extraction of the TB studies, and Susan Mallett and Sally Hopewell for data extraction of reviews for reanalysis. I would also like to thank the National Institute for Health Research's Research Capacity Development initiative for funding my fellowship over the last four years.

I have had several office mates through the years and would particularly like to mention Colin Green for saving my sanity and latterly, Julie, Chris and Chris (and Trisha next door), for providing a stimulating and supportive environment in which I rediscovered my faith and enjoyment in academia. I doubt I would have completed the work without their cheerful encouragement.

My friends and family are in a state of shock that I have finally finished this work. My family have always supported me in everything I do and it is partly for them that I have persevered this far. In recent months mum and dad have gone above and beyond their usual to look after our little Isla and allow me many extra days of PhD time. This was a real godsend and I am very lucky to have them. My husband Jon, I think, will be almost more relieved than I to have this piece of work finished. He always manages to make me see the lighter side of life when all is doom and gloom but he will be glad that grumpy Jac will (largely) be gone and he will finally "get his wife back".

List of abbreviations

AFB	acid fast bacilli
AUC	area under the curve
BCG	Bacille-Calmette-Guerin
BMI	body mass index
BVN	bivariate normal
CAD	coronary artery disease
CDSR	Cochrane Database of Systematic Reviews
CI	confidence interval
CONSORT	Consolidated Standards of Reporting Trials
CT	computed tomography
DARE	Database of Abstracts of Reviews of Effects
DNA	deoxyribonucleic acid
DOR	diagnostic odds ratio
ECG	electrocardiogram
ELISA	enzyme-linked immunosorbent assay
ES	effect size
ESS	effective sample size
FN	false negative
FP	false positive
FPR	false positive rate
GP	general practitioner
HIV	human immunodeficiency virus
HRT	hormone replacement therapy
HSROC	hierarchical summary ROC
HTA	Health Technology Assessment
INAHTA	International Network of Agencies of Health Technology Assessment
IPD	individual patient data
IQR	inter-quartile range

LJ	Lowenstein-Jensen
InDOR	log of the diagnostic odds ratio
LR	likelihood ratio
M.	mycobacterium
MDR-TB	multidrug resistant tuberculosis
MEDION	a database of diagnostic systematic reviews updated by a group of Dutch and Belgian researchers
mmHg	millimetres of mercury
MRI	magnetic resonance imaging
MTD®)	Gen-Probe Amplified Mycobacterium tuberculosis Direct Test
NAAT	nucleic acid amplification test
OR	odds ratio
PCR	polymerase chain reaction
PE	pulmonary embolism
PET	positron emission tomography
PPD	purified protein derivate
PV	predictive value
QA	quality assessment
QUADAS	Quality Assessment of Diagnostic Accuracy Studies
RCT	randomised controlled trial
RDOR	relative diagnostic odds ratio
ROC	receiver operating characteristic
ROR	ratio of diagnostic odds ratios
RROR	ratio of relative diagnostic odds ratios
SE	standard error
SROC	summary receiver operating characteristic
STARD	STAndards for Reporting Diagnostic Accuracy
TB	tuberculosis
TN	true negative
TP	true positive

TPR	true positive rate
TST	tuberculin skin test
UTI	urinary tract infection
WHO	World Health Organisation

1 Introduction

The subject of this thesis is to assess the degree to which the technique of meta-analysis can provide insight into spectrum effects (or patient case-mix) through comparing results between studies (or between subgroups within studies).

This chapter introduces the concepts of diagnostic accuracy and patient spectrum, and the possible influence of spectrum on indices of accuracy in both primary studies and systematic reviews, with or without meta-analysis. It ends with a summary of why it is important to investigate spectrum effects in systematic reviews and a recognition of the limitations in doing so.

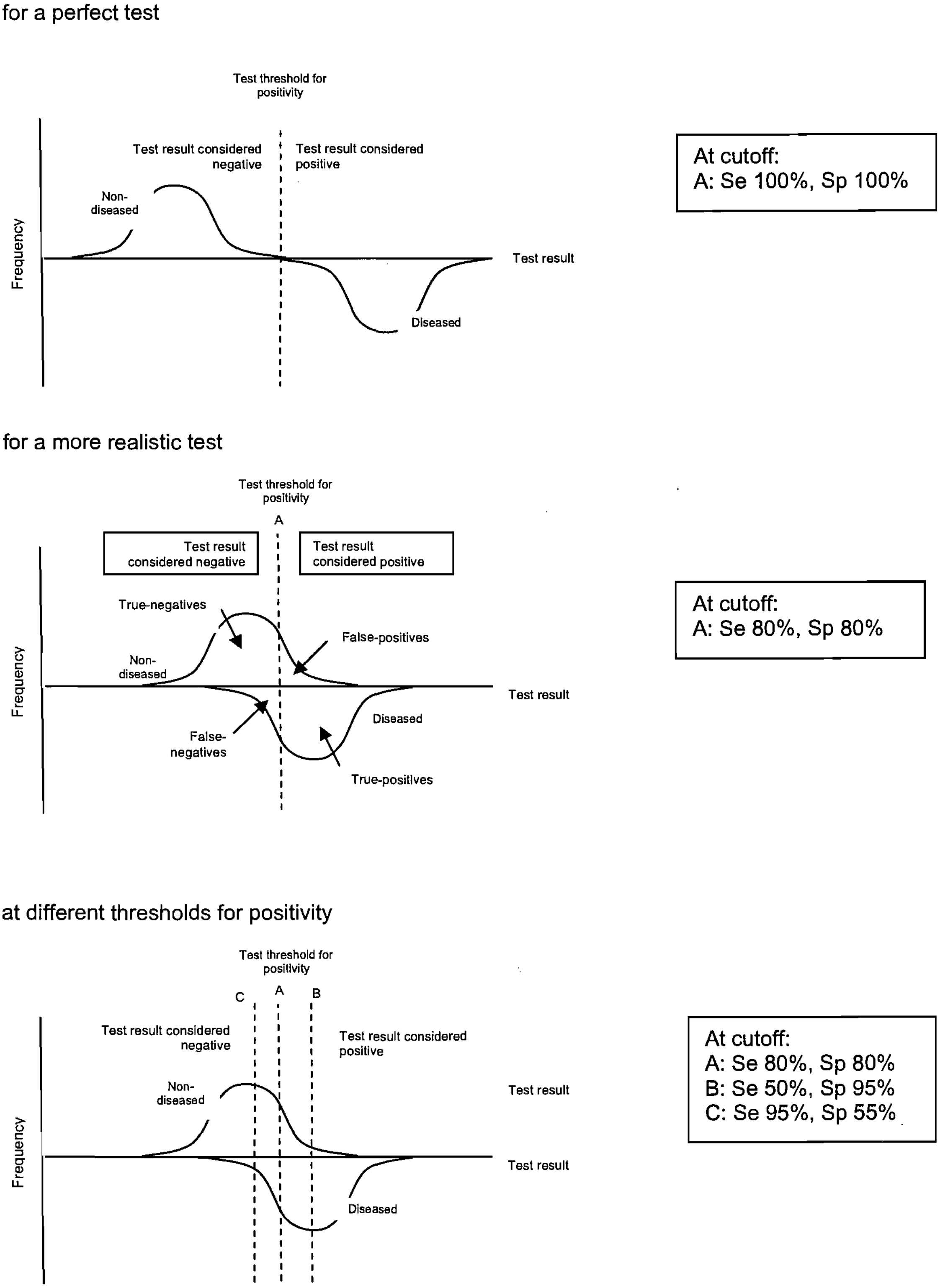
1.1 Diagnosis

Diagnosis is a fundamental element of patient care. It can sometimes be established by clinical examination or history taking but it usually depends on additional laboratory, radiology or pathology tests. The diagnostic process is important for establishing the presence of specific disorders, for informing or monitoring patient prognosis and therapy and in reassuring clinicians and/or patients when disorders are ruled out. Diagnostic tests or strategies can be applied to differentiate diseased from nondiseased (diagnosis); mild from severe disease (for prognosis or therapy decisions); or in a screening situation (which can be seen either as a means of identifying high risk groups, or of identifying disease at an earlier stage). Diagnoses can also be made in situations where there are no effective treatments. The most obvious are for some genetic tests that are applied in order to give peace of mind to the patient, either by excluding the inherited condition, or enabling them to prepare themselves and their families for the onset of disease at a later date.

With the advancement of technology, an increasingly wide variety and number of tests have become available. Their roles may be to replace an existing test, to rule out patients who need not progress to further testing (triage) or they may be performed in addition to one or more existing tests.[1] The value of new tests (or new applications of existing tests) and the contribution that they can make to clinical care requires careful evaluation.

Although there is an increasing interest in the evaluation of diagnostic tests and strategies in terms of their impact on patient management and outcomes,[2,3] there are practical difficulties in designing studies to evaluate these outcomes, not least in terms of the large sample sizes needed to detect the effect on patient outcomes from both test and subsequent treatment. The majority of studies therefore focus on estimating diagnostic test accuracy, whereby the results of one (or more) tests for the detection of a given disorder are compared with the results of some reference standard in a group of patients suspected of having the target disorder. The resulting indices of test accuracy that may be estimated are outlined in Appendix 1 and discussed in more detail in Chapter 2. The most commonly used are

Figure 1 Hypothetical distribution of test results (adapted from Griner and colleagues, 1981[4])



a. for a perfect test

Test threshold for positivity
Test result considered negative
Test result considered positive
Non-diseased
Frequency
Test result
Diseased

At cutoff:
A: Se 100%, Sp 100%

b. for a more realistic test

Test threshold for positivity
A
Test result considered negative
Test result considered positive
True-negatives
False-positives
Non-diseased
Frequency
Test result
Diseased
False-negatives
True-positives

At cutoff:
A: Se 80%, Sp 80%

c. at different thresholds for positivity

Test threshold for positivity
C
A
B
Test result considered negative
Test result considered positive
Non-diseased
Test result
Frequency
Test result
Diseased

At cutoff:
A: Se 80%, Sp 80%
B: Se 50%, Sp 95%
C: Se 95%, Sp 55%

Se - sensitivity; Sp – specificity

sensitivity (proportion of diseased participants who have positive test results) and specificity (proportion of nondiseased participants who have negative test results), but other indices such as predictive values and likelihood ratios are also presented in reports of primary studies.

In order to better understand sensitivity and specificity, it helps to think about the distribution of test results in diseased and nondiseased graphically (Figure 1). Although often interpreted dichotomously, most tests can be perceived as having a continuous distribution. For example, the results of imaging tests tend to be divided into four or five categories ranging from definitely positive through to definitely negative. Biochemical tests measure the level of a given indicator in the blood or urine, producing results that can range, for example for creatinine kinase from zero to over 480 units.[5] Similarly, blood pressure is measured to the nearest 1 mmHg.

Figure 1a demonstrates the distribution of test results for a 'perfect' test, i.e. one that could discriminate between diseased and nondiseased people with 100% accuracy. There is no overlap in the distributions of test results and hence no false-positive or false-negative results. A more realistic picture is presented in Figure 1b, where there is some overlap in the distribution of test results for diseased and nondiseased persons. Positive test results in both diseased and nondiseased are on the right-hand side of the threshold line and negative results are on the left. Figure 1c demonstrates the trade-off between sensitivity and specificity with changing threshold; as the threshold decreases (moves to the left in this example) sensitivity increases (fewer false negative results) and specificity decreases (more false positive results) and vice versa.

The distribution of test results in diseased and nondiseased people and the relationship between them are subject to a variety of biases and effect modifiers which can affect test accuracy, just as for randomised controlled trials and non-randomised or observational studies for the evaluation of therapeutic interventions. The investigation of one of these – so-called spectrum effects – is the main focus of this thesis.

1.2 Diagnostic test accuracy and patient spectrum

The term 'spectrum' was coined by Ransohoff and Feinstein[6] to represent the pathologic, clinical and co-morbid patient characteristics (both for diseased and nondiseased) that might affect a test's sensitivity and/or specificity. In other words, it refers to the case-mix of patients included in a study. For the diseased group, pathologic features are defined as those relating to the extent, location and, for certain diseases such as cancer, the cell-type of disease. The clinical component refers to features such as the chronicity and severity of symptoms. A test may be positive in patients with more extensive or severe disease and not in those with localised or less severe disease. The co-morbid component refers to co-existing conditions, not directly related to the disease under investigation, but that may share the same underlying determinants that may make a test falsely negative.[6] For example, leukocyte esterase-based dipstick tests for the detection of urinary tract infection can be falsely negative in patients with immune suppression as the test relies on the presence of white blood cells. The Ransohoff and Feinstein definition does not specifically refer to the potential impact from demographic variables, however there may be circumstances in which age or gender for example might affect test accuracy, e.g. exercise testing for heart disease.[7] It is also possible however that these may be proxies for true spectrum-related characteristics that are difficult to precisely identify, characterise and record.

For nondiseased patients, or the comparator group, the relevant features to look for are those that might lead to false-positive diagnoses. Generally these relate to the presence of co-morbid conditions whose pathologic or clinical features might be sufficiently similar to those in the diseased group as to cause false-positive diagnoses. For example Ransohoff and Feinstein[6] give the example of a study of a radiolabelled dye marker for diagnosing the patency of the cystic duct in cholecystitis – patients with severe liver disease may give false-positive results if the liver does not excrete dye properly, but such patients were not included in the study leading to falsely elevated specificity.

Sensitivity and specificity (for any given threshold) are often considered to be fixed test properties so that what are assumed to vary between studies with different prevalences of disease are the predictive values.[8] Re-consideration of the contingency tables from which accuracy indices are calculated (Appendix 1) demonstrates the basis for this assumption. Whilst predictive values are calculated across the rows of the 2x2 table, sensitivity and specificity are calculated on the columns. If the overall relative number of patients with and without disease should change (change in prevalence), the proportion of each who test positive need not necessarily change.

The source of the confusion is perhaps the fact that test accuracy is viewed within a probability framework – sensitivity being the probability that a patient with disease will have a positive test result and specificity the probability that a patient without disease will have a negative test result. This implies that the results of a diagnostic test are random, i.e. if a test has been shown to have 70% sensitivity and it is applied to a randomly selected group of patients with the target disorder in question, 70% would test positive.[9] It does not take a broad stretch of the imagination to see that this is not true of most tests and diseases – not all patients with (or without) disease will have the same chance of testing positive (or negative). In other words "homogeneity of risk" is unusual, therefore where patients are not all equally likely to have a positive result, sensitivity and specificity will be strongly affected by the case mix of patients recruited to a given study. As Rutjes and colleagues so concisely stated, "diagnostic accuracy is not a feature of a test itself but a description of how the test behaves in a particular clinical population".[10]

1.2.1 Patient spectrum, disease prevalence and variation in sensitivity and specificity

Ransohoff and Feinstein were amongst the first to propose that each of the components of spectrum could affect the results of a test in both diseased and nondiseased patients[6] and that, in some cases, problems in the choice of spectrum for any one component could invalidate a study's results. A key factor is the possible impact from disease severity on sensitivity and of conditions mimicking the target disorder on specificity. If the sensitivity of a test is related to the severity of disease, a test that is highly sensitive in patients with severe disease may be less discriminatory or even useless in those with mild to moderate disease.[6,10] As one might expect a higher proportion of more severe disease in higher prevalence studies, it follows that sensitivity may appear to increase with increasing prevalence. Specificity is affected by the range of alternative diagnoses in patients without the target disorder that could cause false positive results. Specificity may fall in studies with higher prevalence due to a higher proportion of patients with diseases most closely resembling the target disorder. If people without the disease in question share some of same underlying characteristics or have similar clinical features to those with the disease, they become more difficult to separate at the gatekeeper primary care level, i.e. the false positive rate would be higher and specificity lower.

Table 1 Hypothetical example of how spectrum might affect accuracy with constant prevalence

a. Spectrum of diseased

		Number of patients	
Stage of disease	**Test sensitivity by stage of disease**	**General practice** **(n = 100)**	**Hospital** **(n = 100)**
Early	0.50	80	20
Intermediate	0.75	15	30
Advanced	1.00	5	50
Observed sensitivity		0.56	0.83

b. Spectrum of nondiseased

		Number of patients	
Alternative diseases	**Test specificity by alternative disease**	**General practice** **(n = 100)**	**Hospital** **(n = 100)**
Alternative disease X	0.30	30	75
Alternative disease Y	0.95	65	25
Healthy	0.99	5	0
Observed specificity		0.76	0.46

Deeks[11] provides a theoretical example to show how differences in the distribution of diseased and nondiseased characteristics can occur without any difference in prevalence. Table 1 shows two studies of the same test, one conducted in general practice and one in a hospital setting. The general practice study has a higher proportion of milder cases of disease compared to the hospital study, despite the same overall prevalence of disease. The sensitivity of the test is therefore lower in general practice than when used on a hospital sample.

Similarly, if the likelihood of a false positive result is greater when certain alternative diagnoses are present (for example Alternative disease X in Table 1b), and these alternative diagnoses are more likely to be present in a hospital sample compared to a general practice sample - perhaps because GPs find it particularly difficult to distinguish them from the target disorder - the overall specificity of a test will vary according to the study setting.

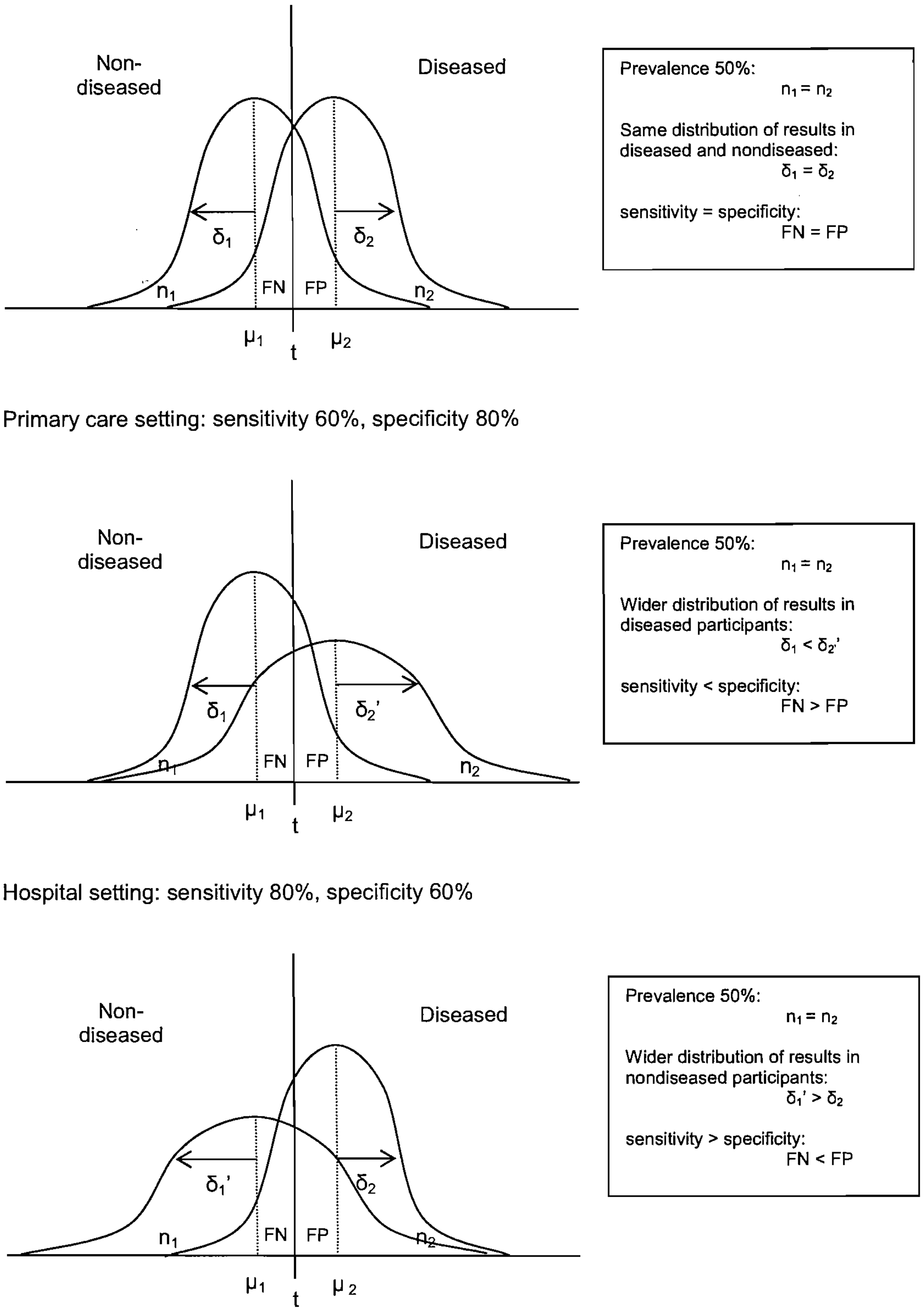
In order for test sensitivity and specificity to remain constant across different prevalences of disease, the mix of disease severity and symptoms must be the same regardless of disease prevalence.[12] Test results must then differ between the diseased group and the nondiseased group (usually being higher in the diseased group), but not within the diseased group nor within the nondiseased group (i.e. the distribution of results in each group should be constant). In other words the distributions of test results in diseased and in nondiseased should be constant in both average (location) and spread (shape).[13] This is unlikely in practice unless the test is not affected by disease severity or the presence of alternative diagnoses/conditions.

Variation in test results due to differing responses to a test that in turn result from variation in spectrum-related characteristics is shown schematically in Figure 2. Figure 2a presents the distribution of test results for a hypothetical test. The prevalence of disease is 50%, therefore the two bell-shaped curves representing the test results in nondiseased and diseased participants cover the same area (number of nondiseased, n_1 is equal to the number of diseased participants n_2). The mean value for the test results in nondiseased and diseased are represented by μ_1 and μ_2; the distance between them indicating how good (discriminating) the test is. The distribution of test results in diseased and nondiseased is the same, therefore the standard deviation in test results in nondiseased (δ_1) is equal to the standard deviation in test results in diseased participants (δ_2), where δ is the standard deviation of the mean test result (μ). The sensitivity and specificity of the test are also equal (number of false-negative results, FN, is equal to the number of false positive results, FP).

Figure 2b and Figure 2c show how the same test might perform in two different settings, a primary care setting and a hospital setting. In the primary care, ‘gatekeeper’, setting one might expect a lower prevalence of disease and lower proportions of both participants with

Figure 2 Variation in distribution of results with constant prevalence

a. Hypothetical scenario



Prevalence 50%:
$n_1 = n_2$

Same distribution of results in diseased and nondiseased:
$\delta_1 = \delta_2$

sensitivity = specificity:
FN = FP

b. Primary care setting: sensitivity 60%, specificity 80%

Prevalence 50%:
$n_1 = n_2$

Wider distribution of results in diseased participants:
$\delta_1 < \delta_2'$

sensitivity < specificity:
FN > FP

c. Hospital setting: sensitivity 80%, specificity 60%

Prevalence 50%:
$n_1 = n_2$

Wider distribution of results in nondiseased participants:
$\delta_1' > \delta_2$

sensitivity > specificity:
FN < FP

Where:
n – number of participants, δ – standard deviation (distribution of results), μ – mean test result, t – threshold for positivity, FN – false negatives, FP – false-positives

more advanced disease and of nondiseased particpants with characteristics closely resembling the diseae in question compared to the hospital setting. For the sake of simplicity of presentation, the two figures show scenarios with the same prevalence of disease. The test, however, is better at picking up more advanced cases of disease therefore the sensitivity of the test is lower in primary care than when used in a hospital setting. Figure 2b shows that the distribution curves for diseased and nondiseased still cover the same area (n_1=n_2) but the shape of the curves differs. The curve for the diseased group is flatter and wider, indicating that although the mean of the test results (μ_2) does not necessarily change, fewer participants have test results near to it.[a] The distribution of results (δ_2') has increased such that there are more participants with test results in the tails of the distributions and therefore a higher number of false negative results. The curve for the diseased participants in the hospital-based study (Figure 2c) remains similar to that shown in Figure 2a as there are a higher proportion of patients with advanced disease and the test is therefore shown to be more accurate.

The distribution curves for nondiseased participants in Figure 2b and c show how this variation in the mix of nondiseased participants translates into variation in the distribution of test results between general practice and hospital settings and therefore into a variation in specificity between settings.

Variation in spectrum therefore can manifest as a variation in prevalence leading to a misconception that sensitivity and specificity vary with prevalence. In fact variation in sensitivity and specificity is not actually due to a direct relationship with prevalence but, as shown here, is related to the distribution of disease severity or symptoms in diseased patients and of conditions similar to that of the target disorder in nondiseased participants that in turn leads to variation in test results.[11,12] This indirect relationship of sensitivity and specificity with prevalence is generally related to study setting, as prevalence would not be expected to remain constant across different study settings.

The referral process and patient spectrum

A key factor affecting the different prevalences and spectrums of disease across different settings is the referral process. As Sackett[5] has pointed out, major clinical centres of excellence will have a particularly distorted sample of patients with a given condition in comparison with the general population of such patients. Given its reputation and expertise, particularly problem cases are more likely to be referred there and are also more likely to be kept track of once referred in comparison to less “interesting” patients.

Knottnerus and colleagues[14] outlined three factors that affect whether or not a patient is referred from general practice to a more specialist setting. They did not include factors such

[a] Note that the distribution curves are not necessarily symmetrical about the mean but may be skewed, most likely towards the centre of the data, i.e. the left-hand tail for the diseased group and right-hand tail for the nondiseased group will contain more results than the tails for the extreme positive or negative results.

as patient anxiety or pressure for second opinion that can also influence referral decisions. The first of the factors outlined is the degree of 'symptomatology'. This can range from asymptomatic to "fully developed" and can also affect patients without the target disorder, e.g. patients with no coronary artery disease may still suffer from chest pain. The second is the suspicion of disease severity – there may be clinical signs or other reasons (outwith symptoms), for example family history, that lead a clinician to suspect that the disease is present. Finally the result of a diagnostic test applied by the GP will influence the probability of referral – an abnormal result will increase the probability as might a normal test result if the clinician has strong clinical grounds for suspecting the presence of a given disease.

The overall impact of referral of patients with more symptoms in whom suspicion of disease severity is higher is that a relatively large proportion of patients who are referred will have abnormal test outcomes (both true positive and false positive) in a referral setting, leading to increased sensitivity, decreased specificity or both.[14]

Where referral is influenced by the result of the same test applied by the GP or other referring physician, specificity will fall further. This was shown by Rozanski and colleagues[15] for exercise radionuclide ventriculography for the detection of coronary artery disease in angiographically normal patients. Evaluations of the test conducted in an earlier time period found much higher test specificity compared to evaluations conducted at a later time due to changes in the patient population. When radionuclide ventriculography was first used, it was evaluated on more severe cases and relatively healthy controls. As the apparent high accuracy of the test became better known and its use became more widespread, an abnormal response to radionuclide ventriculography then became a powerful decision criterion for referral to coronary angiography such that lots of patients undergoing angiography had already had an abnormal response to radionuclide ventriculography. The subsequent commissioning of later studies to evaluate ventriculography that selected only patients who had undergone angiography would produce

a. sensitivity estimates approaching 100% (as all patients with disease would have a positive response to both ventriculography and angiography) and
b. falling specificity (as a large proportion of patients without disease would also have an abnormal ventriculography result in the study setting, given that a previous abnormal ventriculography result had referred them for angiography in the first place).

Philbrick and colleagues[16] have further outlined the way in which patients can be selected for inclusion in a research study after referral, using a sample of patients who underwent exercise tests. They begin with the available population of 205 patients, i.e. those who were present, or who had been referred to the appropriate medical centre at the time of the research. It is worth noting that members of the true clinical population may not have consulted their physician, may have obtained their health care in another setting, or may have refused the invitation to undergo the test. This may be related to factors such as symptom severity, socio-economic status or practice habits of their physicians.[16]

The available population was reduced by 128 patients, either because they had conditions considered likely to cause false-positive or false-negative results (n=98), or because their exercise test results were uninterpretable (n=30). A further 71 patients did not undergo the reference test leaving only six patients who underwent angiography. This is a key problem for retrospective studies where samples are selected on the basis that the reference test has been received – it is highly likely that the decision to refer patients for an often invasive and or unpleasant test is influenced by the clinician's degree of suspicion that the target disorder is present and also potentially, by the result of the index test itself.

Referral bias for whatever reason severely affects that centre's ability to generalise its studies results to other settings.

1.2.2 Spectrum effects…a bias or effect modifier?

As more authors[7,16-19] have recognised the potential effect from spectrum, the term 'spectrum bias' has been used to describe scenarios where the accuracy indices obtained in one study cannot be assumed to apply to other patients in other contexts and also to where test accuracy has been seen to vary according to subgroups of patients within the same study. However as Mulherin and Miller point out,[20] the term 'bias' implies that there has been some systematic error in the study design that to a smaller or greater extent invalidates a study's results. In fact variations in accuracy between subgroups or between studies due to spectrum-related covariates can be true variations, i.e. the test really does perform differently in different groups, and the use of the term 'bias' is therefore something of a misnomer.

If differences in participants arise intentionally, for example by a deliberate selection of certain participants during the recruitment process, it seems reasonable for these differences to be referred to as 'effect modfiers'. If spectrum differences arise unintentionally because of the features of the study design, such as use of a case-control design, so that they give you the wrong answer to the question that you are asking, they should be described as 'bias'.

1.2.3 Dealing with "spectrum" in primary studies

The proposed solution to what is usually termed 'spectrum bias' is often to recruit an 'appropriate' or 'representative' sample whose characteristics reflect the reality of clinical practice, i.e. a broad sample of mild and severe, treated and untreated disease, plus individuals with different but commonly confused disorders. This is similar to the scenario for RCTs where in order to get a picture of an intervention's effectiveness rather than efficacy, pragmatic entry criteria are used in order to include a wider range of participants. This is recommended in a number of tools for assessing the quality of diagnostic test studies and in textbooks on primary study design.[21]

However, regardless of the presence of an appropriate spectrum of patients in a study, the sensitivity and specificity estimates that are produced will still be 'average' estimates that are potentially very unrealistic for certain parts of the spectrum. It is important also to consider conducting subgroup analyses according to plausible covariates. Just as in the field of therapeutic intervention evaluation where appropriate use of subgroup analyses to identify intervention effectiveness according to patient characteristics is recommended[22-25] and the generalisability of study results is often of concern, the same approach has been proposed for diagnostic test accuracy studies.[26-28] However, subgroup analyses appear to be much less common among diagnostic test studies (see section 1.3.3 below), possibly because studies are too small to allow any subgroup effects to be detected or because the likelihood of variation in accuracy by clinically defined subgroups is not well recognised.

The Standards for Reporting of Diagnostic Accuracy (STARD) statement[29,30] published in 2003 covers spectrum in some detail. Similar to the successful CONSORT initiative for RCTs,[31] the STARD initiative aims to improve the quality of reports of diagnostic accuracy. Of the 25 criteria listed in the STARD checklist, six relate at least partially to patient spectrum. These can be broadly classified into three groups

1. who the study participants are,
2. how they were recruited, and
3. what the impact on accuracy is.

The following items relate to who the study participants are, or to the description of the spectrum composition:

i) item 3: describe the study population - the inclusion/exclusion criteria, setting and locations where data were collected
ii) item 4: describe participant recruitment - was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or reference standard?
iii) item 15: report clinical and demographic characteristics of the study population (e.g. age, sex , spectrum of presenting symptoms, comorbidity, current treatments, recruitment centres)
iv) item 18: report the distribution of severity of disease in those with the target condition and of other diagnoses in those without the target condition

These criteria primarily allow the reader to judge the generalisability of the study's results and its applicability to their own setting and patients. The latter item is described as key for the consideration of 'spectrum bias' as the most notable examples involved differences in the severity of the target condition: "test sensitivity is often higher in studies with a higher proportion of patients with more advanced stages of the target condition......[and] in the presence of comorbid conditions, false positive or false-negative test results may occur more often".[30]

The following item relates to how the study participants were recruited:

v) item 5: describe participant sampling: was the study population a consecutive series of participants defined by the selection criteria in items 3 and 4? If not specify how participants were further selected

This item again allows the reader to judge how generalisable the study's findings are. If participants appear to have been highly selected, the resulting sample is unlikely to be representative of the patient population. On the other hand it would of course be possible to have a prospective study using consecutive recruitment that still studied the wrong patients, so this item cannot be considered independently of the preceding ones.

A final item relates to what the impact from spectrum on accuracy is by considering analysis of data in pertinent subgroups:

vi) item 23: report estimates of variability of diagnostic accuracy between subgroups of participants, readers or centres, if done.

Bossuyt and colleagues[30] point out that since variability in study results should always be expected in diagnostic test accuracy studies, pre-planned subgroup analyses should always be performed in order to explore possible sources of heterogeneity in results.

The extent to which these items have been considered by primary test accuracy studies is examined in section 1.3.3 below.

1.2.4 Examples of spectrum affecting sensitivity and specificity

In order to better picture the potential impact from spectrum on sensitivity and specificity, it helps to again think about the distribution of test results in diseased and nondiseased graphically. The plots in Figure 3 describe a series of hypothetical examples of how spectrum might affect the distribution of test results. Prevalence is kept at a constant 50% throughout (the distribution curves for diseased and nondiseased cover the same area) as is the distribution of test results (represented by the shape of the curves), in order to make the graphs simpler to interpret. In reality both the prevalence of disease and the distribution of results between diseased and nondiseased will differ between subgroups when stratified by a spectrum related covariate, as demonstrated in Figure 2. Real clinical examples where differences in patient spectrum have impacted on sensitivity and/or specificity are presented in Table 2.

Figure 3b and c demonstrate a scenario often reflected in theoretical explanations of spectrum effects: sensitivity or specificity increasing at the expense of the other. This might be expected in the presence of a variable that similarly affected test results in both diseased and nondiseased groups. The resulting effect is akin to a shift in threshold, the two distribution curves remain the same distance apart, and the threshold appears to shift up or down.

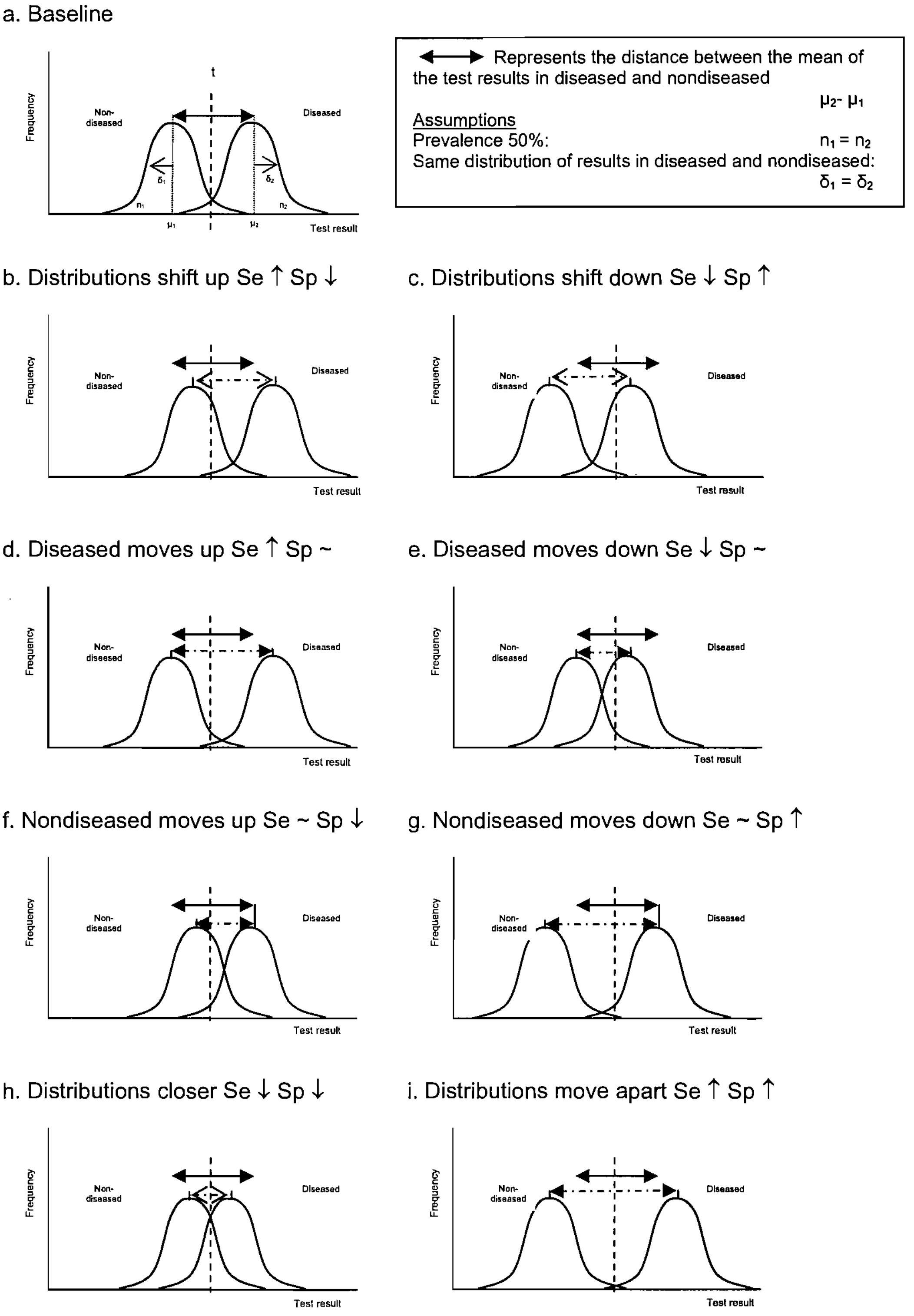
Ginsberg and colleagues[32] have shown a similar scenario to that in Figure 3 b and c when evaluating the D-dimer test to detect pulmonary embolism (PE). As the prior probability of having PE (based on clinical assessment of signs, symptoms and risk factors and the likelihood of a diagnosis other than PE) increased, sensitivity increased (from 79% in the low probability group to 93% in the high probability group) and specificity decreased (from 76% to 45%). Similarly, Hlatky and colleagues[7] found that the sensitivity of the exercise electrocardiography test for detection of coronary disease was highest and specificity lowest in patients with typical angina and sensitivity was lowest and specificity highest in those with non-anginal symptoms (Table 2).

When Hlatky and colleagues[7] compared test accuracy in patients with atypical anginal symptoms and those with typical angina, however, there was very little difference in specificity estimates but big differences in sensitivity. This pattern, depicted in Figure 3d and e was also shown by the data in the study by Mulherin[20] who examined the use of an enzyme immunoassay for the detection of chlamydia in younger versus older women and by Lachs and colleagues[17] who investigated dipstick tests for urinary tract infection stratified by prior probability of disease (Table 2). The authors of the latter study state that although a variety of “classic” UTI symptoms were found in both groups, patients classified as higher probability had a higher prevalence of dysuria, frequency, urgency, double voiding, gross haematuria and costovertebral angle tenderness, and furthermore that one would expect sensitivity to be higher in those with urgency, dysuria and haematuria.[17]

The scenario of constant sensitivity but differences in specificity (Figure 3 f and g), i.e. where test results are only affected by a spectrum-related variable in patients without disease, was demonstrated by the Ginsberg data when patients with a low probability of PE were compared to those with a moderate probability; sensitivity remained almost the same, but specificity dropped from 76% to 52%.

The final possible impact from differences in spectrum is where either both sensitivity and specificity either fall or increase, i.e. the distributions of test results move closer together or further apart (Figure 3 h and i). Two examples comparing test results for the detection of coronary artery disease in men and women suggest this pattern. Morise and Diamond[33] and to a lesser extent Weiner and colleagues[34] found that both sensitivity and specificity were higher in men compared to women (Table 2). In other words, the test is more discriminatory (distributions further apart) in men and, less discriminatory (distributions closer together) in women. It is likely that the extent of disease (number of diseased arteries) and clinical

Figure 3 Hypothetical description of the potential impact of spectrum on accuracy



a. Baseline

Represents the distance between the mean of the test results in diseased and nondiseased

$\mu_2 - \mu_1$

Assumptions

Prevalence 50%: $n_1 = n_2$

Same distribution of results in diseased and nondiseased: $\delta_1 = \delta_2$

b. Distributions shift up Se ↑ Sp ↓

c. Distributions shift down Se ↓ Sp ↑

d. Diseased moves up Se ↑ Sp ~

e. Diseased moves down Se ↓ Sp ~

f. Nondiseased moves up Se ~ Sp ↓

g. Nondiseased moves down Se ~ Sp ↑

h. Distributions closer Se ↓ Sp ↓

i. Distributions move apart Se ↑ Sp ↑

Se – sensitivity; Sp – specificity; n – number of participants, δ – standard deviation (distribution of results), μ – mean test result

Table 2 Primary studies demonstrating potential spectrum effects

Example	Disease	Test subgroup	Total n	Prevalence of disease	Sensitivity	Specificity	LR+	LR-	DOR
Ginsberg, 1998[32]	Pulmonary embolism	d-Dimer	1177	16.7%	84.8%	68.4%	2.68	0.22	12.0
		low probability	703	3.4%	79.2%	76.0%	3.30	0.27	12.0
		moderate probability	382	26.4%	80.2%	51.6%	1.66	0.38	4.3
		high probability	92	78.3%	93.1%	45.0%	1.69	0.15	11.0
Hlatky, 1984[7]	Coronary disease	Exercise electrocardiography	2269	61.7%	70.3%	84.4%	4.51	0.35	12.8
		Typical angina	1083	87.4%	79.6%	80.9%	4.17	0.25	16.5
		Atypical angina	825	45.9%	52.8%	82.7%	3.05	0.57	5.3
		non-anginal	361	20.8%	41.3%	88.8%	3.69	0.66	5.6
Lachs, 1992[17]	Urinary tract infection	Dipstick	366	19.7%	83.3%	71.4%	2.92	0.23	12.5
		high probability	107	49.5%	92.5%	42.0%	1.59	0.18	8.9
		low probability	259	7.3%	57.8%	77.5%	2.57	0.54	4.7
Morise, 1995[33]	Coronary artery disease	Exercise electrocardiography	788	55.5%	53.5%	77.2%	2.35	0.60	3.9
		men	504	63.1%	56.0%	81.2%	2.97	0.54	5.5
		women	284	41.9%	47.1%	72.7%	1.73	0.73	2.4
Mulherin, 2002[20]	Chlamydia	Enzyme immunoassay	6672	8.8%	0.734	0.994	122.33	0.27	457.1
(based on Miller 2000[35])		Age < 24	4524	11.1%	0.759	0.995	151.80	0.24	626.7
		Age ≥25	2737	3.2%	0.583	0.992	72.87	0.42	173.4
Banks, 2004[36]	Breast cancer	Mammography (postmenopausal only)	92208	1.2%	85.7%	97.3%	31.96	0.15	216.9
		current HRT user	32390	1.2%	80.8%	96.4%	22.19	0.20	111.3
		past HRT user	14610	1.2%	83.6%	97.6%	34.24	0.17	203.5
		never HRT user	45208	1.1%	91.2%	97.9%	44.06	0.09	489.7
Weiner, 1979[34]	Coronary artery disease	Stress testing	2045	58.3%	79.1%	69.1%	2.56	0.30	8.4
		male	1465	69.8%	79.7%	74.0%	3.06	0.27	11.1
		female	580	29.1%	75.7%	63.7%	2.09	0.38	5.5

Total n –sample size; LR+ - positive likelihood ratio; LR- - negative likelihood ratio; DOR – diagnostic odds ratio presentation (e.g. presence or absence of angina, plus atypical presentations in women) influence the likelihood of an abnormal result in diseased patients, however reasons for lower specificity in women seem more difficult to explain.[33] Hlatky and colleagues[9] actually found higher specificity in women, but other studies (referenced in Hlatky 1984[9]) support the findings above.

Banks and colleagues[36] found a similar pattern of results for mammography for the detection of breast cancer in postmenopausal women, according to HRT use. Mammography was found to be more discriminatory in those women who have never used HRT (Figure 3 i) compared to those who are current users (Figure 3 h). Sensitivity and specificity were both highest in those who had never used HRT (91% and 98% respectively) and both were lowest in those who were current HRT users (81% and 96%). Although the difference in specificity was small (1.5%), given the extremely large numbers of women who are screened the impact in real terms on number of women receiving false-positive diagnoses would be quite significant. In this case, HRT use is likely to be a marker for breast density; mammograms are easier to read in women who have never used HRT and therefore more breast cancers are detected (higher sensitivity) and benign breast lumps are easier to distinguish (lower false-positive rate and high specificity).

It is worth re-emphasising that not only might the positioning of the distribution curves for diseased and nondiseased in relation to threshold vary between groups, but the variability in test results between groups might also vary considerably.

1.3 Spectrum effects and systematic reviews

1.3.1 Systematic reviews of test accuracy

Systematic reviews provide a means of synthesising information from a number of studies to "establish where the effects of healthcare are consistent and where they may vary significantly",[37] for example across populations, settings, and differences in treatment. Systematic reviews of therapeutic interventions are now commonplace in many if not most areas of healthcare, and in recent years interest has turned to applying similar techniques to research evaluating diagnostic tests. The UK Health Technology Assessment (HTA) Programme has funded a large number of such reviews, and the Cochrane Collaboration are also introducing reviews of diagnostic test accuracy into the Cochrane Database of Systematic Reviews (CDSR) which is published within the Cochrane Library.

Systematic reviews of any form of intervention follow key stages, including formulation of the question, setting of inclusion criteria, searching the literature, quality assessment and data extraction of included studies, and synthesis of the evidence. Work is ongoing to develop each of these stages specifically for diagnostic test reviews, for example in literature searching[38,39] and quality assessment,[21] and several authors have published general guidelines for the conduct of reviews of test accuracy.[40,41] Meta-analytic techniques for combining diagnostic studies are also being developed and improved in order to better estimate test accuracy.[11,26,40-45] Similarly, the conclusions of all systematic reviews are only as reliable as the included primary studies and there are a variety of sources of heterogeneity that can affect conclusions.

Between study differences - or heterogeneity in results - can result from chance, from errors in calculating accuracy indices or from heterogeneity caused by differences in design, conduct, participants, tests and reference tests. These are outlined in some detail in Chapter 2. The main focus of this thesis is on heterogeneity due to variation in patient spectrum.

1.3.2 Potential impact from spectrum variation

Mulherin and Miller[20] present a similar theoretical example to that of Deeks[11] discussed in section 1.2.1 that can be extended to consider how variation due to spectrum alone could impact on the conclusions of a systematic review.

Table 3 Hypothetical example of how spectrum could affect the conclusion of a systematic review (adapted from Mulherin and Miller[20])

	Number of patients				**Accuracy**				
	Total	**Prev**	**tp/dis**	**tn/non-dis**	**Sens**	**Spec**	**LR+**	**LR-**	**DOR**
True test performance									
Aged <50	1000	0.50	475/500	375/500	0.95	0.75	3.8	0.1	57
Aged ≥50	1000	0.50	375/500	475/500	0.70	0.85	4.7	0.4	13
Studies with varying age spectrum recruited									
Study A: Equal recruitment of both age groups									
Aged <50	500	0.50	475/500	375/500	0.95	0.75	3.8	0.1	57
Aged ≥50	500	0.50	375/500	475/500	0.70	0.85	4.7	0.4	13
Overall	1000	0.50			**0.83**	**0.85**	5.7	0.2	32
Study B: 75% aged < 50									
Aged <50	750	0.50	238/250	375/500	0.95	0.75	3.8	0.1	57
Aged ≥50	250	0.50	175/250	213/250	0.70	0.85	4.7	0.4	13
Overall	1000	0.50	413/500	425/500	**0.89**	**0.80**	4.5	0.1	36
Study C: 75% aged ≥ 50									
Aged <50	250	0.50	356/375	213/250	0.95	0.75	3.8	0.1	57
Aged ≥50	750	0.50	88/125	106/125	0.70	0.85	4.7	0.4	13
Overall	1000	0.50	444/500	400/500	**0.76**	**0.90**	8.0	0.2	36

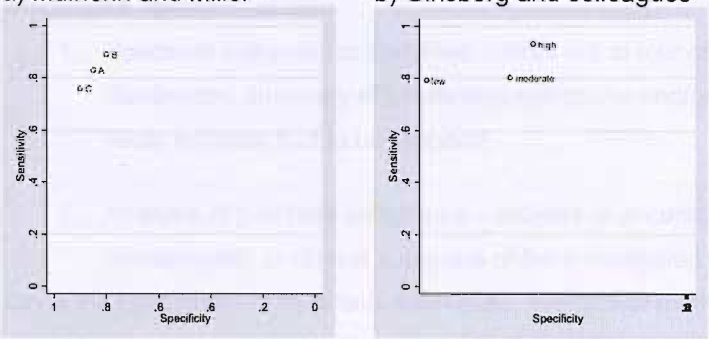
Prev – prevalence; tp – true positives; tn – true negatives; dis – diseased; non-dis – nondiseased; Sens – sensitivity; Spec – specificity; LR+ - positive likelihood ratio; LR- - negative likelihood ratio; DOR – diagnostic odds ratio

Table 3 demonstrates the situation where the true performance of a test varies according to patient age; sensitivity is higher and specificity lower in patients aged less than 50 and vice versa for those aged 50 and over (the distribution of test results would lie predominantly to the right or to the left of the threshold for test positivity). The impact on overall sensitivity and specificity when different spectrums by age group are recruited is then given for hypothetical studies A, B and C (Table 3); disease prevalence is held constant for the sake of simplicity.

The breakdown by age group for each study (A, B and C) shows that test accuracy within the age subgroups is 'unbiased', i.e. reflects the true sensitivity and specificity of the test within those groups, but when the groups are combined, the overall sensitivity and specificity is considerably affected. Equal recruitment of both groups results in almost equivalent values for sensitivity and specificity; preferential recruitment of younger women results in a scenario similar to that depicted in Figure 2b and preferential recruitment of older women, Figure 2c. Plotting these results on a ROC plot (Figure 4a) demonstrates the extent to which results can vary according to the percentage of patients less than 50 who are recruited to that study.

Figure 4 ROC plots demonstrating impact from spectrum on a systematic review

a) Mulherin and Miller[20] b) Ginsberg and colleagues[32]



The same theory can be followed using a real-life example from section 1.2.4. As discussed previously, Ginsberg and colleagues[32] found that sensitivity and specificity of the D-dimer test were affected in opposite directions by the clinical probability of pulmonary embolism being present, sensitivity was highest and specificity lowest in patients with high clinical probability and vice versa for those with low clinical probability. If we assume that each of these subgroups is a separate study of D-dimer, each recruiting patients from different populations with varying probabilities of pulmonary embolism and plot them on a ROC plot (Figure 4b) we can see the variation in sensitivity and specificity that results.

1.3.3 Challenges in investigation of spectrum effects in systematic reviews

There is clearly potential for spectrum variation to have a big effect on the results of both primary studies and systematic reviews. It is rare for diagnostic accuracy studies to be sufficiently large in size or to recruit a sufficiently broad spectrum of participants to allow the influence of spectrum to be examined. It seems that it is more rare for diagnostic accuracy studies to have actually looked for variation due to spectrum. Systematic reviews that include all available studies of a given test for a given disorder, are the best available tool that we have to assess the contribution that a diagnostic test can make to healthcare, assuming that all sources of bias and other causes of heterogeneity are fully investigated. They are also the best tool available to allow us to identify the extent that heterogeneity can affect test accuracy. However there are challenges in such investigations that one needs to be aware of. Two of the major challenges to be faced are

1. limitations in the primary studies, and
2. limitations in the meta-analytic methods available

Limitations in the primary studies

One of the main reasons for any lack of investigation of spectrum effects in systematic reviews is due to the lack of reporting, and potentially recording of, spectrum-related factors in primary studies. It can be particularly challenging to identify let alone record and publish true spectrum-related characteristics; variables such as age and sex are often used instead as proxies. A methodological review by Reid and colleagues[28] assessed 112 primary studies evaluating the accuracy of diagnostic tests against seven methodological criteria, two of which were spectrum-related:

1. Spectrum composition specified – three out of four of age distribution, sex distribution, summary of presenting symptoms and/or disease stage, and eligibility of study subjects had to be reported

2. Analysis of pertinent subgroups – indexes of accuracy cited for any pertinent demographic or clinical subgroup of the investigated population.

Since the publication of Reid and colleagues, five further methodological reviews using similar methodology and assessing diagnostic accuracy studies against the same or similar spectrum-related criteria have been published[b] (Table 4). Across all six reviews, 44% (135/308) of primary studies published between 1970 and 2002 were judged to have adequately specified the spectrum composition of the included study samples and 48% reported either consecutive or random sampling of participants. Only 24% (64/268) had included separate analysis of pertinent patient subgroups. There does appear to have been some improvement over time for all three criteria. Adequate specification of the spectrum composition was 25% for primary studies published up to the early 1990s and 58% for those published from around 1993 to 2002, appropriate participant sampling went up from 40% to 50%, while analysis of pertinent subgroups went up from 8% in the earlier period to 35% in the later one. Nevertheless, these are still not sufficiently high proportions and suggest that the potential impact from patient spectrum on diagnostic test accuracy is inadequately assessed in primary studies.

The publication of the STARD statement[29] should go a long way to addressing this issue. In the 2005 review by Siddiqui and colleagues[48] each of five spectrum-related criteria in 16 primary studies in the ophthalmic literature were assessed. In addition to the two criteria suggested by Reid and colleagues[28] (for which their studies scored 75% and 25%

[b] Other such reviews have also been published, e.g. Sheps and Schechter,[46] Arroll 1988,[47] etc but they did not use the same spectrum-related criteria.

Table 4 Methodologic reviews of studies of diagnostic test accuracy: spectrum-related criteria

Study	Source/Inclusion	Study years	No. papers	Spectrum composition described	Describe participant sampling	Analysis of subgroups	Other spectrum-related criteria
Reid, 1995[28]	Medline, 4 *general medical* journals	1978-1993	112	30 (27%)		9 (8%)	
Heffner, 1998[49]	Medline, 9 general medicine and 6 subspecialty journals for *pulmonary disease* papers	1992-1997	41	25 (61%)	22 (54%)	12 (29%)	indicated study sample, i.e. pts suspected of target disorder: 37 (90%)
Harper, 1999[50]	Papers on *ophthalmic* diagnostic tests selected from recent publications; 9 of which identified from Medline search	1980-1997	20	12 (60%)		11 (55%)	
Rothwell, 2000[51]	Medline plus author references, random sample papers evaluating tests to measure *carotid stenosis*,	1970-90	20	3 (15%)	8 (40%)		
		1993-97	20	15 (75%)	10 (50%)		
Lumbreras-Lacarra, 2004[52]	Medline, 3 *clinical chemistry* journals	1996	18	4 (22%)		8 (44%)	
		2001	27	10 (37%)		7 (26%)	
		2002	34	24 (71%)		13 (38%)	
Siddiqui, 2005[48]	Hand searching of 5 major *ophthalmic* journals Criteria all from STARD guidelines	2002	16	12 (75%)	8 (50%)	4 (25%)	- describe study population - incl/excl criteria, setting, location: 13 (81%) - describe participant recruitment: presenting symptoms, previous test results, or receipt of index/reference test: 13 (81%) - report distribution of severity of disease in those with the target condition and of other diagnoses in those without the target condition: 10 (62%)
	Primary studies published up to c1993			33/132 (25%)	8/20 (40%)	9/112 (8%)	
	Primary studies published from c1993-2002			102/176 (58%)	18/36 (50%)	55/156 (35%)	
	TOTAL			135/308 (44%)	26/56 (46%)	64/268 (24%)	

respectively), they found that 81% of studies adequately described the study population and patient recruitment, and 62% reported the distribution of severity of disease in patients with the target condition and of other diagnoses in those without the target condition. The authors intend for this to provide a baseline against which to evaluate the impact of the STARD statement.

A further problem partially linked to poor reporting of data in primary studies is our ability to investigate sources of heterogeneity at the aggregate level. Where details of patient subgroups within studies are not available, as is likely for diagnostic accuracy studies, one has to rely on aggregated study-level data, such as the percentage of women in each study, or the mean age of study participants, and examine whether that variable explains differences between the studies. Simulation work in the field of RCTs has shown that the statistical power of meta-regression techniques is dramatically and consistently lower than that of individual patient data analysis.[53]

Methods available for meta-analysis

Systematic reviews of diagnostic accuracy studies aim to report both the individual study results together with a summary of the central tendency and variability of the studies. The central tendency of the data is either summarised as a typical *operating point* (average sensitivity and specificity) or as an *SROC curve*, which describes the pattern of values of sensitivity and specificity that could occur across different test thresholds. For a test where the chance of disease increases with the test value, increases in the cut-off value will increase specificity and lower sensitivity according to a curvilinear relationship depicted by the SROC curve (see section 2.2.3 for a fuller explanation of threshold effects).

Obtaining a summary operating point by separately averaging estimates of sensitivity and specificity is frequently used,[11] but is known to potentially be misleading. The approach ignores the negative correlation likely to exist between sensitivity and specificity where there are differences in threshold between the included studies, and produce a summary that falls below the SROC curve. In a systematic review it is likely that the same cutoff has not been applied in all studies, and even when this is not the case, similar threshold type effects can arise through differences in test interpretation between observers, characteristics of the sample and differences in the execution of tests.

One device to overcome the correlation between sensitivity and specificity, is to undertake meta-analysis using a single summary statistic created from them. The diagnostic odds ratio (DOR) is one option, and is computed as the ratio of the odds of a positive test result in a patient with disease compared with a patient without disease (a quantity that has little direct clinical meaning). The DOR allows for a trade-off between sensitivity and specificity as points on a summary ROC curve typically have very similar DOR even when they have different sensitivity and specificity. However, it is possible that the DOR does vary with thresholds, particularly when the diseased and nondiseased groups differ in both the variance of the test measurements as well as the average value of the underlying test result, e.g. patients with disease may on average have higher values of a given marker than patients without disease but they may also have a greater variation in values compared to nondiseased (Figure 2).

The 'SROC approach' allows this variation in DOR across different thresholds and for this reason is usually recommended over straight pooling of DORs.[11,26,42] The method devised by Moses and Littenberg[54] is the most commonly used, however there are problems with it, not least that there is no sound statistical basis for the approach. Alternative so-called 'advanced' methods of meta-analysis, such as the hierarchical SROC regression (HSROC) method[55,56] and the bivariate normal (BVN) method[57,58] have been developed. These do have a sound statistical basis and also allow for threshold effects and for variation in DOR with threshold. These methods are not in widespread use due to the computational difficulties in undertaking them, however they have become much more accessible for use in diagnostic systematic reviews.[59,60] The two approaches were originally proposed as alternative models however recent work has shown that under certain circumstances they are actually different parameterisations of the same model.[59] This and the ability of the methods to investigate sources of heterogeneity requires further examination.

The next chapter gives more detail on these methods and on other key aspects to consider when carrying out systematic reviews of diagnostic test studies.

2 Establishing diagnostic test accuracy in systematic reviews

This chapter considers how test accuracy is established in primary studies and systematic reviews, the various sources of heterogeneity (or variation) in study results, and selected statistical methods for pooling primary studies in systematic reviews.

2.1 *Establishing test accuracy*

2.1.1 In primary studies

Diagnostic test accuracy is established by comparing the results of a test for the detection of a given disorder with the results of some reference standard in a group of patients suspected of having that disorder. These results are classified into a 2x2, or contingency, table and a variety of indices of test accuracy reflecting the new test's (also referred to as the index or experimental test) discriminatory ability, or ability to correctly identify patients with and without the disorder, are then estimated (see Appendix 1). The reference standard (sometimes referred to as the 'gold' standard) should be the best available method for making a definitive diagnosis of the presence or absence of the disease or condition in question and ideally should indicate with 100% certainty the presence or absence of disease, although in practice absolute certainty is rarely achieved.[61] In most cases the reference standard is more invasive, more unpleasant and/or more costly than the test under investigation, hence the search for as accurate an alternative as possible. The most commonly used accuracy indices are sensitivity (proportion of diseased participants who have positive test results) and specificity (proportion of nondiseased participants who have negative test results), but predictive values and likelihood ratios are also presented in reports of primary studies (see Appendix 1 for definitions).

Primary studies of test accuracy are observational and cross-sectional in design, that is they aim to compare the result of the index test with that of the reference standard in the same participant at the same time.[10] They can be either prospective or, commonly, retrospective in design. Diagnostic accuracy studies bear some similarity to the 'cohort' and 'case-control' studies commonly used in epidemiology for the evaluation of aetiology, however diagnostic studies are distinct from studies of aetiology as there is usually little or no time difference between the application of the index test (exposure) and application of the reference test (outcome) so that loss to follow-up, a main drawback of aetiology studies, is not such an a issue.

Rutjes and colleagues[10] have outlined four main variations on the diagnostic accuracy design. The first, termed the "classic" design assembles patients suspected of a disease (ideally in a prospective manner) in whom the new (index) test and then the reference standard are performed, and results are compared. Studies of this design are generally referred to in the literature as case series or cohort studies.

The "reversed flow design" reverses the order in which the reference standard and index tests are applied: cases and controls are sampled from the same population (patients suspected of having the condition), the reference standard is applied first and the index test applied after disease status is known. This bears some similarity to the nested or etiologic case-control design in that cases and controls are selected from the same source population, typically defined by the clinical presentation.[10] This design would enable researchers to recruit a certain number or proportion of diseased participants or to apply the index test to only a random sample of nondiseased (reference test negative) participants.

The remaining two designs can also be thought of as variations on the case-control design but cases and controls this time are sampled from two distinct populations. Diseased participants (cases) are sampled from a clinical (often hospital) population, while controls are sampled either from the 'healthy' general population or from a group of participants diagnosed with a specific alternative diagnosis or diagnoses that is known to produce symptoms and signs similar to those of participants with the target condition.[10] These designs are described as "two-gate designs" either using "healthy controls" or with "alternative diagnosis controls".

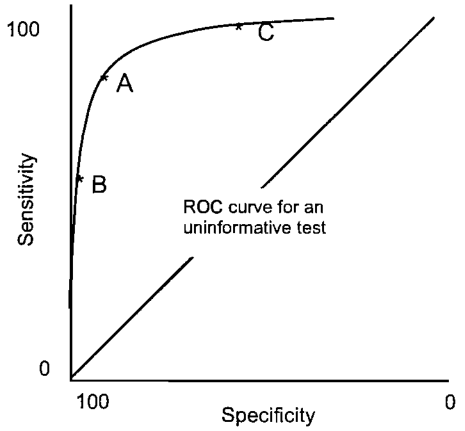
For the latter two designs to generate accurate estimations of sensitivity, representative sampling of cases must be ensured. If participants with advanced disease are over-represented it is likely that sensitivity estimates will be inflated. It is unlikely that specificity estimates from studies using healthy controls will be representative of the test's performance in routine practice as most of them will be unlikely to have alternative diagnoses that might generate false-positive results. In a classic or reverse-flow design, all alternative diagnoses (that are more or less likely to cause false-positive results) will be represented; where specific alternative diagnosis controls are used, specificity could be over- or under-estimated depending on the alternative diagnosis concerned.[10] In the former scenarios the effect on accuracy from spectrum can be considered as an effect modifier, in the latter, if representative sampling is not ensured, it should be considered a 'bias', as introduced in Chapter 1.

A further design termed the "two-gate design with representative sampling" still has two sets of inclusion criteria, one for cases and one for controls, but both are sampled in such a way that both groups are representative of those obtained in the classic diagnostic accuracy study. Such designs would be difficult to achieve and none have been identified in the literature.[10]

Graphical depiction of test results in diseased and nondiseased

The distribution curves for test resuls in diseased and nondiseased persons introduced in Figure 2a can be transformed to ROC space. Primary study results at different thresholds are displayed and an ROC curve drawn through the points to demonstrate the trade-off between sensitivity and specificity (Figure 5). The closer the curve to the top left-hand corner of the plot, the more discriminative is the test. The closer the curve to the centre diagonal (ROC curve for an uninformative test), the less accurate the test. Some authors present the results of their studies in terms of the 'area under the curve' (AUC); perfect tests have areas under the curve of close to 1, whereas poor tests have AUC close to 0.5.[62] The AUC is a global of measure of test accuracy and as such only gives an overall picture of the accuracy of a test; it does not give any indication as to the expected operating point of the test.

Figure 5 ROC curve



Sensitivity and specificity at different thresholds for test positivity:

At cutoff
A: Sensitivity 80%, Specificity 80%
B: Sensitivity 50%, Specificity 95%
C: Sensitivity 95%, Specificity 55%

ROC plots can also be used in systematic reviews to display the sensitivity and specificity pairs from individual studies. Rather than demonstrating the effect of changing threshold on accuracy (as for primary studies above), ROC plots in this context demonstrate the amount of variability in sensitivity and specificity that there is between primary studies (see Figure 6).

2.1.2 In systematic reviews

Systematic reviews of RCTs, particularly where meta-analysis (the use of statistical methods to summarise the results of independent studies) can be used, can provide more precise estimates of the effects of healthcare interventions than those derived from the individual studies included in a review[37] and allow decisions about healthcare to be made that are based on the totality of the available evidence.

The use of statistical methods to combine test accuracy studies is particularly challenging, not least because test accuracy is conventionally represented by a pair of statistics (most often sensitivity and specificity, see Appendix 1) and not by a single measure of effect such as the odds ratio or relative risk. The paired nature of sensitivity and specificity – one increasing and the other decreasing with changing threshold – means that separate pooling of sensitivities and specificities or even positive and negative likelihood ratios is not usually the best approach as it does not allow for this variation with threshold (i.e. doesn't account for the correlation between them). A variety of methods of meta-analysis that allow for variation due to threshold are available and will be discussed below.

2.2 Sources of variation in diagnostic test accuracy other than spectrum

Before conducting a statistical synthesis, sources of variation in study results should be considered. There is almost always considerable variation, i.e. heterogeneity, between the results of diagnostic studies, possibly to a greater extent than is seen for therapeutic interventions, though this comparison has not been quantified in empirical studies. This may be at least partially due to the fact that the importance of rigorous design has been less well appreciated than for therapeutic interventions, consequently diagnostic studies have often been retrospective and not conducted according to standard protocols.

Furthermore, in randomised trials, the statistical outcomes that are considered are usually relative comparisons (such as relative risks and odds ratios) or absolute comparisons (such as risk differences and differences between means) of event rates between treated and control groups made within each trial. While often there is substantial variation in the event rates in the treated groups and in the placebo group between the trials (as displayed in a L'Abbé plot), there may be little variability in relative or absolute comparisons between these event rates. In contrast, for analyses of diagnostic test accuracy the focus is on the event rates in the diseased (test sensitivity) and in the nondiseased (test specificity), and not on relative or absolute comparisons between diseased and nondiseased groups within studies. Thus the level of heterogeneity observed in test accuracy reviews may be higher than that observed in randomised trials due to the statistical focus not being on comparisons within studies but on absolute estimates of event rates.

Between study differences in results can result from:

- chance, from
- errors in calculating accuracy indices or from
- heterogeneity[63] including,
    - methodological heterogeneity or biases in the conduct of studies that can be significantly reduced by rigorous design
    - clinical heterogeneity that arises from true differences in accuracy between different test populations
    - differences in the test under study, and
    - variation in the threshold for positivity or test cut-off.[11,26,40-45]

Empirical evidence for the impact of many of these quality features on test accuracy is still limited. Two studies[64,65] have found several features that significantly over- or under-estimated test accuracy, including the use of case-control design with healthy controls and severe cases of disease, use of different reference tests, selective inclusion of patients and retrospective data collection.[65] Whiting and colleagues have reviewed the literature to provide a summary of the available evidence that supports various sources of bias or variation.[66]

Table 5 outlines various quality features that are often included in quality assessment tools for diagnostic test studies, some of which are discussed in more detail below, and groups them according to whether they are predominantly concerned with internal validity (or study design issues) or with study generalisability. All of these can potentially impact on accuracy estimates obtained from a study and all can be confused with the impact of differences in patient spectrum – the challenge is to differentiate spectrum effects from variation due to other causes.

Table 5 Quality considerations for evaluation of a diagnostic test study[41,67]

Quality feature	Internal validity	General-isability
1. Study population		
Selection bias	√	
Spectrum composition		√
2. Selection and execution of tests		
Verification bias	√	
Use of appropriate reference test	√	
Description of index and reference test execution		√
3. Test interpretation		
Blinding	√	
Inter-observer variability	√	√
4. Data analysis & presentation		
Uninterpretable test results	√	
Interpretation of test results in clinical context		√

2.2.1 Study design and quality considerations

Selection bias

In the evaluation of therapeutic interventions, the term 'selection bias' generally refers to bias resulting from the way that comparison groups are assembled.[68] In that context, randomisation is the only means of allocation that controls for unknown and unmeasured confounders as well as those that are known and measured.[37] It is possible to control or adjust for confounders that are known and measured in observational studies but it is not possible to adjust for those factors that are not known to be confounders or that were not measured. In the epidemiological literature, selection bias in case-control studies reflects selection of either cases or controls that is not independent of the exposure under test. Here it more broadly reflects any bias that may occur due to the selection of subjects for investigation.

The ideal study sample for the evaluation of a diagnostic test is one obtained from a consecutive (or randomly sampled) series of patients recruited from a relevant clinical population and who meet the study inclusion criteria, i.e. a prospectively designed diagnostic accuracy study.[10,11] Selection bias can occur in several ways, for example where:

i) only those who are referred for the reference test or, in a retrospective study those who actually underwent the reference test, are included
ii) inclusion is influenced by the result of the experimental test, i.e. test positives more likely to be included
iii) patients are selected on the basis of another test result that is related to the result of the test under study

In each of these cases, only those patients whom clinicians most expect to have the target disorder will be included in the study such that the proportion of patients who have a positive result will be higher than if all eligible patients had been included. Variations on the "case-control" design tend to be at higher risk from selection bias: cases tend to be selected on the basis of a positive reference test result and the result of the test under evaluation ascertained after true disease status is known; the prevalence of the target disorder tends to be higher than in practice; and cases and controls are often selected from opposite ends of the disease spectrum, e.g. severe cases and healthy controls.[69]

Referral bias occurs where there is a systematic selection of patients for referral to the experimental test who have characteristics differing from those of the entire population. Usually, only patients most likely to have the disorder undergo the experimental test and therefore become eligible for study inclusion. In the presence of selection bias or referral bias patients available for inclusion in a diagnostic test study may be an unrepresentative sample of the population to whom the test will be applied in practice, i.e. both biases may result in variations in spectrum or case-mix.

It is often difficult to establish the extent to which retrospective diagnostic accuracy studies have been subject to selection bias if, for example, the data necessary to identify all people who would have been eligible for a test was not routinely collected, or where test results have not been recorded in any systematic manner.[12]

Lijmer and colleagues found a significant over-estimation of accuracy (increase in DOR 3.0, 95%CI: 2.0, 4.5) to result from the use of a two-gate design using healthy controls as opposed to a classic single-gate design for the evaluation of identical tests.[64] Using the same methodology, Rutjes and colleagues[65] found the bias to be much less in case-control studies that selected controls from patients with diseases more closely resembling the target disorder. Whiting and colleagues identified four studies which on balance tended to show increased accuracy in the presence of distorted selection of participants.[66] It is likely that sensitivity would be over-estimated and specificity under-estimated.

The 'best' diagnostic accuracy studies are prospective in design with consecutive recruitment of patients; this allows evaluation on the full spectrum presenting in that setting, the collection of appropriate baseline information and implementation of rigorous protocols for testing.

Verification bias

Verification bias occurs where the decision to undertake or apply the reference test is influenced by the result of the experimental or index test[11,67,70] (also called ascertainment bias or work-up bias). There are two potential elements to verification bias:

1. Partial verification occurs where only a subgroup of patients who received the index test undergo the reference test (e.g. where the reference test is unpleasant or invasive, such as biopsy or angiography). This incomplete verification may be equal in test positive and test negative cases (i.e. cases missing at random), or it may be differential where those most likely to have the disease tend to undergo the reference test,
2. Differential verification occurs where different tests are used according to the results of the experimental test (e.g. index test positive patients may undergo a more invasive and probably more accurate reference test than those who tested negative on the index test).

For example, in a study of radionuclide ventriculography for detecting coronary artery disease, 31% of index test positive cases underwent verification compared to only 14% of index test negatives.[71] The better the test under evaluation, or at least the stronger the investigator's faith in the test, the greater will be the tendency to preferentially verify index test positives and the greater will be the bias introduced.[72]

Whiting and colleagues[66] found two studies that demonstrated increased accuracy in the presence of differential verification. The 20 studies that investigated the effects of partial verification bias had mixed results, mainly suggesting an increase in sensitivity and decrease in specificity; only two of the 20 studies found no effect from partial verification.

Use of an appropriate reference test

Standard techniques for assessing diagnostic tests assume that a definitive reference test is available, that is, that the reference test used is as close to 100% accurate as can be. However, it can be either that the available test is far from perfect, or that such a test simply does not exist. For example, the diagnosis of metastatic liver cancer can never be definitively determined even at autopsy. The key issue really is not to find a test that confirms a text book definition of disease but to find a test that has practical consequences for patient management, hence the use of the term 'target disorder' as opposed to 'disease'.

In some contexts where a single definitive reference test is unavailable, a reference strategy may be used, where the reference diagnosis is made on the basis of clinical information in combination with a battery of other tests.[11] Incorporation bias occurs where the experimental test is used as part of the reference strategy, i.e. the experimental test and reference tests are not independent, leading to over-estimation of both sensitivity and specificity.[12]

Even the most definitive reference test may have considerable inaccuracies, for example, microbiologic studies of sputum for the detection of tuberculosis can fail to detect mycobacteria that may be picked up by nucleic acid amplification tests, and will incorrectly classify patients with TB as false-positive results.[72] Walter and Irwig[61] refer to a 'substantial body of literature' demonstrating that reference tests may frequently be imperfect. Serious inaccuracies in the reference test will lead to over- or underestimation of the true accuracy of a new test. If the index and reference test are conditionally independent then the new test's characteristics will be underestimated (non-differential misclassification); if the two tests are perfectly correlated, or if the new test makes the same errors as the reference test, the accuracy of the new test will be over-estimated,[11] potentially appearing perfectly accurate regardless of its association with true disease status.[67]

Whiting and colleagues identified 8 studies looking at the effects of an inappropriate reference standard.[66] All 8 studies found some association with sensitivity, specificity or accuracy, but the effects were not consistent across studies.

Blinding or masking

The interpretation of many diagnostic tests involves some degree of subjective interpretation. In clinical practice, test interpretation can be influenced by both the knowledge of the results of other tests and by the specific clinical characteristics of the person being tested. Diagnostic review bias occurs where knowledge of the reference test result influences interpretation of the experimental test, whilst test review bias refers to the opposite situation. Clinical review bias is said to occur where knowledge of patients' clinical characteristics or other test results influences test interpretation (experimental or reference test). For example, to adequately evaluate the accuracy of ultrasound for the detection of rotator cuff tear, observers should not have access to the results of other imaging tests such as x-ray or MRI. This should be distinguished from observer variability which will occur in interpretation of almost any test.

The recommended solution to these biases is to perform a 'blinded' study, where both tests are interpreted without knowledge of the clinical characteristics or the test results[70] to ensure that it is only the diagnostic contribution of the test itself that is being evaluated. Of course this is not the same as routine clinical practice where prior information is used to evaluate the results of subsequent tests. Blinding is particularly important where a new test is intended to replace an existing test, for example the use of MRI instead of ultrasound for the assessment of shoulder pain. Where clinical factors play a significant role in assisting test interpretation, such as in the shoulder pain example above, or where a new test is intended to supplement an existing test, it may be more appropriate to identify the additional diagnostic value added by the test, rather than essentially evaluating the test in isolation.

Whiting and colleagues identified 13 studies looking at the effects of some form of review bias.[66] Only two of the 13 did not find evidence of increases or decreases in either sensitivity, specificity or accuracy. The most common finding was of increased sensitivity in the presence of review bias (8 studies).

2.2.2 Variation in test(s)

The manner in which the index and reference tests have been carried out should be described, not only as good reporting practice and so that the study could be replicated, but to allow a judgement to be made regarding the applicability of the study's results.[41] Just as variations in the timing, duration and dosage or intensity of a therapeutic intervention can affect effectiveness, diagnostic test accuracy may be affected by variations in timing, in technical aspects of any equipment or materials used, inter and intra observer and laboratory variation. Similar variations in the reference standards used must also be considered.

Whiting and colleagues found few studies that investigated the effects of biases and sources of variation associated with the test protocol, making it difficult to draw conclusions on any effect on test performance.[66] They propose that the magnitude of any effect is probably linked to the test and condition under investigation, being more significant for tests that require some expertise to perform and for acute conditions that may change rapidly compared to more chronic diseases.

2.2.3 Threshold effects

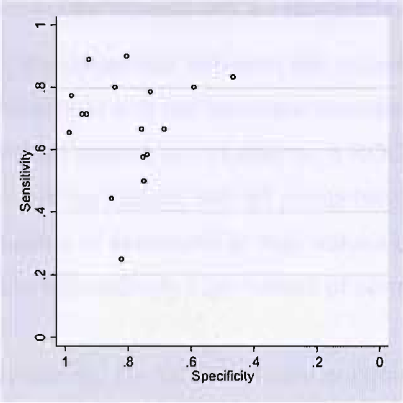
A source of heterogeneity that is unique to meta-analyses of diagnostic tests is variation in the cut-off or threshold chosen to indicate test positivity. Statistics used to report the results of diagnostic tests (e.g. sensitivity and specificity) by nature present a test result as binary, i.e. a test is either positive or negative, disease either present or absent. However, in practice the majority of tests effectively produce continuous data such that an arbitrary cut-off point (diagnostic threshold) is applied to define positive and negative test outcomes. In some cases, such as laboratory tests, this could be explicit numerical cut-offs. Imaging tests, e.g. mammograms, can be interpreted on a categorical scale ranging from definitely normal to definitely abnormal with various categories of suspicion in between. These thresholds can also be affected by variation between laboratories or between observers[11] – one observer's 'mildly abnormal' may be another's 'definitely abnormal'. The diagnostic classification of patients therefore depends on whether the measurement of a given trait is above or below some defined cut-off or threshold value, and the threshold chosen may vary between studies of the same test. The higher the cut-off value chosen, the higher the specificity and lower the sensitivity estimates. The issue of threshold effects is further discussed below.

Whiting and colleagues found very little evidence for any effects from nonarbitrary choice of threshold value.[66]

2.3 Selected methods of meta-analysis and how they may reveal/hide spectrum effects

The first stage of meta-analysis is to plot the results of individual studies graphically in order to assess the degree of variability between study results. As mentioned in section 2.1.1, ROC plots are a useful tool for displaying sensitivity and specificity pairs from individual studies in a systematic review (Figure 6). The pattern of results can also provide an indication as to whether or not there is variation between studies due to threshold, i.e. a threshold effect. Threshold effects are usually interpreted as present if the plotted points mimic the shape of a ROC curve; if the points appear to vary around some central point, there is assumed to be minimal variation due to threshold. However, it is possible for a similar pattern of results to be introduced by variation in the spectrum of diseased and nondiseased patients between study populations.

Figure 6 Sample ROC plot for a systematic review

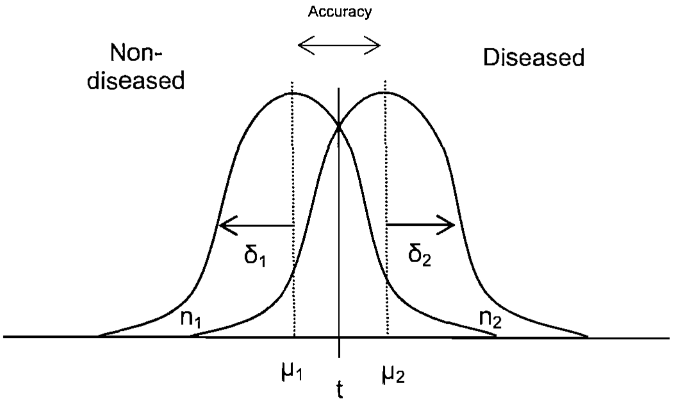


Straight pooling of diagnostic accuracy indices such as sensitivity and specificity does not allow for the presence of any threshold effect and therefore cannot distinguish heterogeneity due to threshold from heterogeneity due to other sources of variation. The exception to this is the diagnostic odds ratio (DOR). The DOR describes the ratio of the odds of a positive test result in a patient with disease compared with a patient without disease. It is easier to understand as a statistical concept than a clinical one and is useful for meta-analysis of test accuracy as it encompasses all four cells of the 2x2 table rather than the two each for sensitivity and specificity (Appendix 1).

Although the DOR allows for a trade-off between sensitivity and specificity, pooling of individual DORs should only be performed if it can be assumed that the relationship between sensitivity and specificity is constant, i.e. that the DOR is constant across different thresholds.

As shown in Figure 7, the discriminatory ability (accuracy[c]) of a test can be defined as a function of the mean test results in nondiseased and diseased groups (μ_1 and μ_2) and the standard deviation from the mean for each group (δ_1 and δ_2). Where the standard deviations from the mean are equal ($\delta_1 = \delta_2$), accuracy is the difference between the means divided by the standard deviation.[73]

Figure 7 Scenario required for symmetric SROC curve



For simplicity, prevalence 50%:
$n_1 = n_2$

Same distribution of results in diseased and nondiseased:
$\delta_1 = \delta_2$

Discriminatory ability:
$(\mu_2/\ \delta_2) - (\mu_1/\ \delta_1)$

as $\delta_1 = \delta_2$,

Discriminatory ability:
$(\mu_2\text{-}\mu_1/\ \delta)$

n – number of participants, δ – standard deviation (distribution of results), μ – mean test result

If the difference between the means remains constant across studies despite differences in threshold and the standard deviations between groups are equal, the DOR will be constant.[d] When studies are plotted on a ROC plot they will be described by a symmetric shaped SROC curve consistent with all points having the same diagnostic odds ratio. This means that the values of sensitivity at high values of specificity will be the same as the values of specificity at correspondingly high values of sensitivity.

However, the DOR will vary at different thresholds when the diseased and nondiseased groups differ in both the average value of the underlying test result and also in the variance of the values, e.g. patients with disease may on average have higher values of a given marker than patients without disease but they may also have a greater variation in values,[74] i.e. $\delta_1 \neq \delta_2$. Where this occurs, the DOR at higher thresholds will be higher than the DOR at lower thresholds. The resulting SROC curve (the derivation of which is described below) will not be symmetric about the sensitivity=specificity line. These concepts are discussed further in section 2.3.2.

[c] The term 'accuracy' can also used to describe a specific index of accuracy, i.e. the proportion of patients in a study who test positive or the proportion of true results (both true positives and true negatives) in the population. In this context it is used as a general term to describe the discriminative ability of the test

[d] Prevalence of disease can vary between studies

2.3.1 Characteristics of an optimal meta-analytic method

Before describing the various methods of synthesising diagnostic test accuracy studies that do allow for threshold effects and for variation in DOR with threshold, it is useful to consider the characteristics of an optimal meta-analytic method:

1. the model should be bivariate in its parameterisation and should allow interpretation in terms of sensitivity and specificity.
2. the model should use appropriate weighting. The number of diseased patients in a study can differ considerably from the number of nondiseased patients, resulting in varying levels of uncertainty in sensitivity and specificity. The different levels of uncertainty or precision associated with the sampling variability in TPR and FPR should therefore be addressed.
3. the model should allow for the threshold relationship or correlation between sensitivity and specificity.
4. the model should use a random effects approach. Considerable heterogeneity between studies is almost always to be expected in a systematic review of a diagnostic test or tests.

2.3.2 Moses and Littenberg SROC method

The Moses and Littenberg SROC method summarises the performance of a test across studies by fitting a summary (or 'average') ROC curve through the observed points.[54,75] Central to the method is the concept that the trade-off between TPR and FPR is most often due to threshold variation, although it can also be due to the other sources of variation in accuracy such as variation in tests and testing methods, methodological differences and variation in patient spectrum.

Model formulation

The model uses the log of the DOR (denoted D) and the log of a proxy measure of threshold (denoted S). D and S are estimated for each study in a meta-analysis in the following way:

$$S = \ln\left(\frac{TPR}{(1-TPR)} \times \frac{FPR}{(1-FPR)}\right) = \text{logit}(TPR) + \text{logit}(FPR)$$

$$D = \ln(DOR) = \ln\left(\frac{TPR}{(1-TPR)} \times \frac{(1-FPR)}{FPR}\right) = \ln\left(\frac{\text{LR + ve}}{\text{LR - ve}}\right) = \text{logit}(TPR) - \text{logit}(FPR)$$

The logit indicates the log of the odds, as used in logistic regression. D, estimated by subtracting the logit of the FPR from the TPR, is the log of the DOR and is a direct measure of how well the test discriminates between diseased and nondiseased. S, estimated by adding the two logits together, is related to how often the test is positive,[75] and increases as threshold decreases. Note that D and S are defined as the difference and sum of the same two measures - TPR and FPR - each of which are estimates and therefore have an unknown degree of error and furthermore covariance may exist between them. Although the uncertainty and covariance can be corrected for it is not usually considered by meta-analysts.[59]

The next stage is to plot D against S for each study and compute the best fitting straight line using the linear regression equation:

$$D = a + bS$$

where *a* denotes the intercept and *b* the slope of the regression line (Figure 8a). This regression line is then transformed into ROC space and an SROC curve generated (Figure 8b). The SROC curve does not connect a set of points as the ROC curve for a primary study does, but rather reflects the central tendency of the data from the primary studies.

Symmetric versus asymmetric SROC curves

As previously mentioned, an SROC curve may be symmetric or asymmetric depending on the relationship of DOR with threshold.

This can be best illustrated by relating D and S back to the distribution of test results in diseased and nondiseased participants (Figure 7), Macaskill shows that TPR and FPR can both be defined as functions of threshold (t), the mean test results in nondiseased and diseased participants (μ_1 and μ_2) and their standard deviations (δ_1 and δ_2),[73] therefore D and S can also be defined as functions of the same parameters.

Where the distribution of results is the same in diseased and nondiseased participants ($\delta_1 = \delta_2$) the formulae can be simplified to show that although S is linearly related to threshold (t), D does not depend on threshold (t).[73] In this case, S can therefore be assumed to be zero. This results in a horizontal regression line and an SROC curve that is symmetric about the sensitivity=specificity line.

Where the distribution of results between diseased and nondiseased participants is not equal ($\delta_1 \neq \delta_2$) both D and S depend on threshold. The resulting regression line has a positive or negative slope and the SROC curve is asymmetric. The degree of asymmetry in the curve will depend largely on the extent of the difference between δ_1 and δ_2. The derivation of an asymmetric SROC curve is as follows:

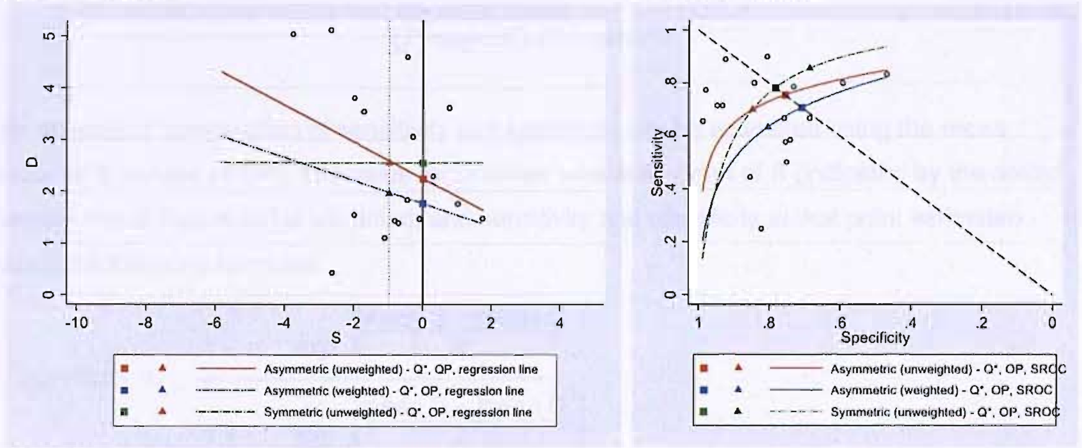
$$sensitivity = \frac{1}{1 + \dfrac{1}{e^{a/(1-b)} \times \left(\dfrac{1 - specificity}{specificity}\right)^{(1+b)/(1-b)}}}.$$

Model interpretation

Returning to the linear regression model, the coefficient for D is the log DOR and indicates the point at which the regression line intercepts the y-axis, i.e. where S is zero or sensitivity equals specificity (Figure 8a). The exponential of *a* therefore gives the DOR associated with the SROC curve at the point where sensitivity=specificity (or the Q* point) (Figure 8b). The higher the value of *a*, the higher the DOR and the closer the SROC curve will be to the top left hand corner of the ROC plot. The intercept value can also be interpreted as a measure of the distance between mean test results in diseased and nondiseased (μ_2-μ_1, Figure 7); the further apart the two distributions, the better the test and the higher the value of *a*.

The coefficient for S indicates how the DOR changes with threshold. If *b*=0, the DOR does not change with threshold, the regression line will be horizontal and the resulting SROC curve will be symmetric (i.e. DOR is constant all along the SROC curve). When the DOR does vary with S (i.e. b≠0), the coefficient for the slope (b) has a considerable effect on the shape of the SROC curve.[59] The higher the value for *b*, the steeper the slope of the regression line and the more asymmetric the SROC curve (the more the DOR varies with threshold). If b has a positive value, DOR increases with increasing test positivity, and vice versa if b has a negative value. Macaskill shows that S can be interpreted as a weighted average distance of true threshold, t, from the mean test results in diseased and nondiseased (μ_1 and μ_2).[73]

Figure 8 Sample Moses plots using data from Scheidler and colleagues[76]



a) 'D vs S' plots

b) SROC curves

Where Q* - point where sensitivity=specificity, OP is the operating point estimated using the mean value for 'S' across studies.

Weighting

The Moses model is usually fitted using weighted or unweighted least squares linear regression.[54] Weighting is commonly carried out using the inverse variance of D. This assumes that variation between studies is due solely to sampling error (like fixed effects). Whilst this carries appeal in that it combines studies according to the precision of their estimates of the odds ratio, it is problematic when sensitivity or specificity (and hence odds ratios) are high, as the formula for the approximate variance of a log diagnostic odds ratio becomes biased (the variance becoming over-estimated) when any of the counts of true positives, true negatives, false positives or false negatives is close to zero.[77] Weighting by inverse variance therefore gives less weight to studies with high sensitivities and specificities (Figure 8b), all other things being equal.

Using an unweighted (or equal weight) regression model gives results more akin to random effect assumptions (i.e. where variation is not just due to sampling error but to real differences in accuracy between studies), because both within and between-study variance are taken into account.[60] The effect is to give relatively higher weight to smaller studies, as would occur in a random effects model when heterogeneity is present.

Estimation of sensitivity and specificity

Difficulties in applying the DOR and associated SROC curve in clinical practice mean that the most likely operating point on the SROC curve is often estimated. A commonly used index is Q*.[78] Q* is the point on the regression line where S=0 or the value on the SROC curve where sensitivity is equal to specificity.

Q* is not useful if the studies in the analysis do not include estimates of sensitivity and specificity near to the Q* point as in the example in Figure 8b. It is estimated using the intercept value *a* estimated from the regression equation (i.e. the log diagnostic odds ratio when the threshold parameter is zero) and inserting it into the equation:

$$Q^*=sqrt(e^a)/(1+sqrt(e^a))$$

An alternative combination of sensitivity and specificity can be estimated using the mean value of S instead of S=0. The value for D where where S=mean of S (indicated by the dotted vertical line in Figure 8a) is identified, and sensitivity and specificity at that point estimated using the following formulae:

$$\text{Specificity} = \frac{1 - \quad \exp\left(\frac{\text{mean S} - \text{mean D}}{2}\right)}{1 + \quad \exp\left(\frac{\text{mean S} - \text{mean D}}{2}\right)}$$

$$\text{Sensitivity} = \frac{\exp\left(\text{mean D}\left(\frac{(1-\text{specificity})}{\text{specificity}}\right)\right)}{1 + \quad \exp\left(\text{mean D}\left(\frac{(1-\text{specificity})}{\text{specificity}}\right)\right)}$$

Because we are using mean values across the dataset, the point identified lies closer to the centre of the data than Q*. This point is indicative of the *average* sensitivity and specificity, however it does not account for the variability in values between studies. It should also be remembered that these points represent only one small part of an asymmetric SROC curve; DOR might vary considerably along the curve.

Investigating heterogeneity in the Moses and Littenberg SROC method

Spectrum effects, or biases, are a source of heterogeneity in a systematic review and are therefore investigated in the same way as other sources of heterogeneity, by extending the model to allow for covariates.[26,42] A covariate, X, can be added to the regression equation for each potential effect modifier:

$$D = a + bS + c_1X_1.$$

The exponential of each of these terms estimates multiplicative increases in diagnostic odds ratios (relative odds ratios) for each factor. An underlying assumption is that the shape of the summary ROC curves is not affected by covariates; i.e. the SROC curves are parallel.

A further extension to the model allows for different shapes for the SROC curves indicated by the covariates. To do this interaction terms between covariates and thresholds are included in the model:

$$D = a + bS + c_1X_1 + d_1S.X_1$$

If the covariate indicates, say differences between two tests, this model is equivalent to fitting separate summary ROC curves for each test. A problem with this model is that it becomes difficult to judge the importance of differences between the curves, as they may differ both in average diagnostic accuracy and shape, and will cross over. Furthermore it would not be possible to identify a source of heterogeneity that had opposing effects on sensitivity and specificity as the overall DOR would not change.

2.3.3 Advanced methods

Rutter and Gatsonis hierarchical SROC method

Rutter and Gatsonis' hierarchical SROC (HSROC) approach models summary ROC curves by estimating the average DOR, the average threshold and the shape (degree of asymmetry) of the curve.[55,56] The HSROC model can be considered as an extension of the Moses model,[55,56,56] allowing for uncertainty at different levels. For this application two levels are considered: variation first within studies, and second, between studies.

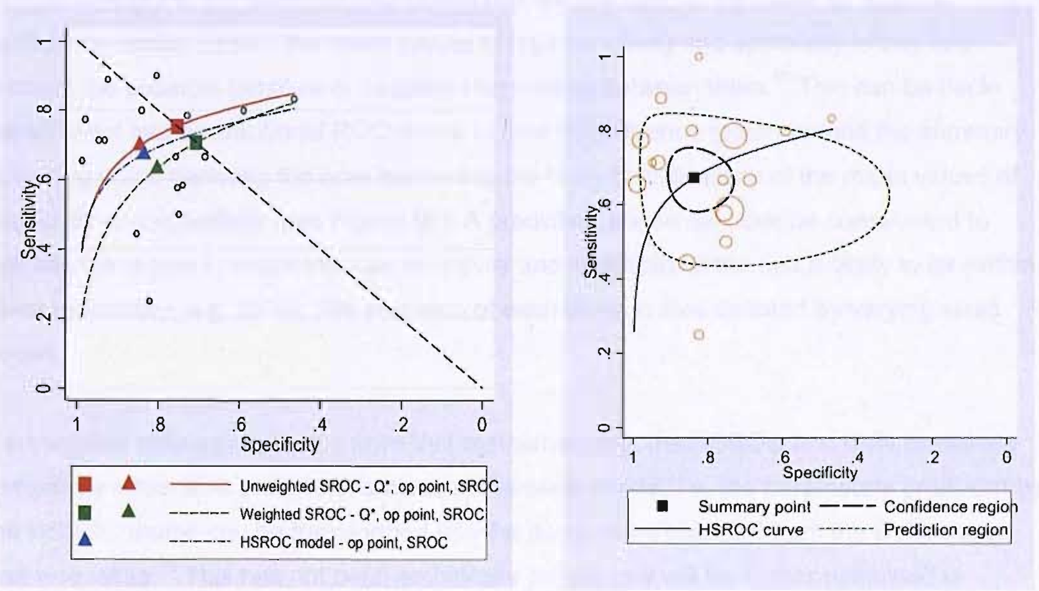
The model is formulated in terms of the probability (π_{ij}) that a patient in study i with disease j has a positive test result, where j=0 for a patient without disease and j=1 for a patient with disease.[79] Appendix 2 provides the full specification of the model. The model yields parameter estimates for

- θ_i (the implicit threshold parameter which models the trade-off between sensitivity and specificity in each study). When $\theta = 0$, the average operating point is at Q*, i.e. where sensitivity=specificity. The value of θ therefore gives an indication of distance from Q*.

- α_i (the log DOR which measures the difference between TP and FP fractions in each study), and
- β which allows for asymmetry in the underlying SROC curve by allowing the log DOR to vary with implicit threshold (i.e. it allows the TP and FP fractions to increase at different rates as θ_i increases). When $\beta = 0$, the DOR for each study does not depend on the cutpoint parameter θ_i and α_i is the log of the DOR. When $\beta \neq 0$ the DOR varies with threshold (θ_i) even if the accuracy parameter (α_i) is held fixed.

The second level of the model fits θ_i and α_i as random effects, so that their average value and variation across studies are estimated. The random effects model takes account of the clustering of TPR and FPR pairs within studies, thereby accounting for the correlation between them. The shape parameter β can only be estimated as a fixed effect (estimated by looking the pattern across studies) because the association between test threshold and accuracy must be derived using data from the studies considered jointly.[60] The precision with which the parameters are estimated is incorporated into the model by weighting in favour of those with more precise estimates.

Figure 9 Advanced method plots



a) HSROC and Moses SROC curves

b) HSROC curve and bivariate plot

Where Q* - point where sensitivity=specificity, OP is the operating point estimated using the mean value for 'S' across studies.

The hierarchical summary ROC curve (Figure 9a and b) is constructed by computing values of sensitivity across the range of specificities using the α (log DOR) and β (shape parameter) estimates from the regression model. The θ (threshold) parameter gives an indication of position on the curve rather than the shape or location of the curve.

Bivariate normal model

The BVN model uses the same hierarchical approach as the Rutter and Gatsonis method,[56] but preserves the sensitivity/specificity parameterisation of the studies, rather than converting test values to estimates of diagnostic odds ratios.[80]

The model uses a random effects approach assuming that the true logit sensitivities for the individual studies are normally distributed around some common mean value $\mu_{A,i}$ with a between study variability of σ^2_A. The same random effect assumption is made for true logit specificities, with mean value $\mu_{B,i}$ and between study variability of σ^2_B. The potential correlation σ_{AB} between sensitivity and specificity (acknowledging the pairing of data within each study and the possibility of threshold effects) is addressed by explicitly including this correlation into the analysis.[57-59] The precision with which sensitivity and specificity have been estimated is also incorporated into the model by weighting in favour of those with more precise estimates. The full model specification is provided in Appendix 3.

The model yields parameter estimates for:

- mean sensitivity, mean specificity and their 95% confidence intervals
- estimates of between study variability in sensitivity and specificity and
- estimates of the covariance between sensitivity and specificity.

The parameters of the bivariate distribution can also be used to calculate an elliptical confidence region around the mean values of logit sensitivity and specificity taking into account the possible (positive or negative) correlation between them.[59] This can be back-transformed into conventional ROC space to give a confidence region around the summary operating point, denoting the area containing the likely combinations of the mean values of sensitivity and specificity (see Figure 9b). A prediction ellipse can also be constructed to indicate the region in which the true sensitivity and specificity of the test is likely to lie (within a given probability, e.g. 95%). The precision of each study is also denoted by varying sized circles.

Harbord and colleagues have shown that mathematically, the HSROC and BVN model are essentially alternative parameterisations of the same model, i.e. the parameters produced by the HSROC model can be transformed into the parameters obtained from the BVN model, and vice versa.[79] This has not been empirically proven and will be further examined in Chapter 4.

Investigating heterogeneity using the advanced methods

Sources of heterogeneity, incluing spectrum effects or biases, are again examined by extending the models to allow for covariates. Under the HSROC parameterisation covariates can be added to the accuracy, threshold and shape components of the model, and are fitted as fixed effects. The significance of covariates can be evaluated by testing the model terms for the covariates, and differences may be noted in whether covariates alter (a) diagnostic odds ratios, (b) the threshold and (c) the shape of the ROC curve. The remaining unexplained variance in both the accuracy and threshold parameters is also given by the random effect terms.

Under the bivariate normal parameterisation covariates are added to both the logit sensitivity and logit specificity components of the model. The effect of each covariate on sensitivity and specificity is thus estimated separately so that any variable that increases one but decreases the other, for example, could be detected. The remaining unexplained variance in both sensitivity and specificity is also given as is the covariance between sensitivity and specificity.

After the introduction of a covariate, the similarity of the two models' output can only be easily maintained if no interaction of the covariate with shape is allowed for the HSROC model. Recall that for an individual study, the distribution of test results in diseased and nondiseased participants determines the shape of the curve; where the two distributions are not equal, the ROC curve will be asymmetric. At review level, the distributions of results in diseased and nondiseased across all studies are considered, to determine whether any differences in distributions are consistent across studies and therefore lead to asymmetry in the SROC curve. Where a covariate is introduced to the HSROC model one can either assume that the two (or more) curves have the same shape (parallel) or that they might have different shapes (crossing curves). If the curves are allowed to have different shapes, one is saying that the distributions of test results in diseased and nondiseased may differ between the subsets, although the degree and statistical significance of any such difference may vary. Parallel curve models ignore any differences in distributions by covariate and model shape for both groups using the whole set of studies.

The BVN model cannot directly consider 'shape' in the same way as the HSROC model. When no covariates are added, the unexplained variances in sensitivity and specificity from the bivariate model are used to estimate the HSROC shape parameter. When a covariate is added, the effect is specified in terms of the effect of that covariate on mean sensitivity and mean specificity and but not on the variances of the two. As the variances are not affected, no change in the shape parameter can be estimated. The HSROC parameters can be converted to bivariate model parameters with or without a shape interaction with a covariate.

The differences in the accuracy, threshold and shape parameters indicate whether the subgroups of studies by covariate have different SROC curves (difference in accuracy), are on the same curve but at different points on the curve (difference in threshold), or on different curves with different shapes (difference in shape), or some combination of these. For example if there is no evidence of differences in accuracy but strong evidence of differences in threshold, the two groups of studies are likely to be on the same curve but at different points on that curve. If the shape term is significant but the accuracy term not, they are likely to be on differently shaped curves but at the same point (e.g. at or near the point where the curves cross).

It is important to remember that DORs and RDORs are estimated at a particular point on the SROC curves. The natural model output estimates both DOR and RDOR at the Q* point, however as previously mentioned, Q* may not be representative of the datasets in the review. The model output can be used to estimate DOR at any point on the curve, for example at the average operating point. The choice of point at which to estimate RDOR can be more complex. For example, if the average operating points of the two groups are some distance apart and the two curves have considerable differences in shape, the distance between the curves could be quite different at each operating point, leading to big differences in RDOR. This issue will be explored in Chapter 4.

2.4 Extent to which the three methods possess the characteristics of an optimal meta-analytic method

The extent to which the three methods possess the characteristics of an optimal meta-analytic method as listed in section 2.3.1 is discussed below. A summary of the characteristics of the three approaches is provided in Table 6.

Table 6 Comparison of statistical methods

Method	Weight	Threshold correlation	Random effects	Parameterisation
M&L (eq)	none (equal)	Yes	No	DOR + S
M&L (w)	1/var(lnDOR)	Yes	No	DOR + S
HSROC	binomial error for sens + spec	Yes	Yes	DOR + threshold
BVN	binomial error for sens + spec	Yes	Yes	sens + spec

M&L (eq) – Moses and Littenberg model with equal weights, or unweighted; M&L (w) - Moses and Littenberg model weighted by inverse variance of the log of the diagnostic odds ratio (lnDOR); sens – sensitivity; spec - specificity

1. The model should be bivariate in its parameterisation and should allow interpretation in terms of sensitivity and specificity.

Although the Moses approach is bivariate in that it is based on two parameters (D and S), its output can only be interpreted within a one-dimensional framework. The original formulation of the model produces a summary ROC curve and allows that curve to have different shapes, but it does not indicate where on that curve we are likely to be. This is because the parameterisation between sensitivity and specificity is lost when DOR is estimated. The method is akin to pooling a single statistic but is an improvement on straight pooling of DOR as it allows threshold variation, or different shaped curves. The Moses model output can be used to estimate summary point and interval estimates for sensitivity and specificity but a value for either sensitivity or specificity must first be specified. Such values are just an arbitrary choice of possible values and may not be representative of values in the primary studies.

The two advanced models are both bivariate in their parameterisations which allows the model output to be interpreted in terms of both test accuracy (DOR) and threshold and sensitivity and specificity.

2. The model should use appropriate weighting to allow for sampling variability

There are two main sources of uncertainty in the Moses model. The first is the variance in D, which the model attempts to account for, and the second is variation in 'S', which the model cannot allow for (linear regression assumes no error in the explanatory variable) and must therefore incorrectly assume to be absent.

The variance in D is allowed for by weighting by inverse variance of D (the log of the DOR). There are two problems with this approach. Firstly, because the sensitivity/specificity parameterisation is lost when weighting by inverse variance of D, the different levels of uncertainty associated with the sampling variability in TPR and in FPR cannot be incorporated (i.e. the number of diseased and nondiseased patients can considerably differ within a study therefore leading to differences in precision between sensitivity and specificity). This can potentially lead to inappropriate significance levels in DOR and its association with threshold. Secondly and more fundamentally, Deeks and colleagues have shown that there are problems with bias in the variance of D,[81] especially where there are zero cells and/or very high values of sensitivity or specificity.

The advanced methods however, appropriately account for different precision of sensitivity and specificity within each study by preserving the sensitivity/specificity parameterisation of the studies. The uncertainty in modelling diseased (sensitivity) and nondiseased (specificity) is considered separately so that the uncertainty in each proportion is accounted for correctly. Studies with more precise estimates of sensitivity and/or specificity therefore get more weight for the estimate of that parameter.

3. The model should allow for the threshold relationship or correlation between sensitivity and specificity

The threshold relationship can be considered at two levels. The first is to allow for threshold-type effects, which all three methods do. The second is direct estimation of the correlation between sensitivity and specificity, which only the advanced methods do. The Moses approach allows DOR to vary with S, the proxy measure of threshold. Because D and S are computed before undertaking the modelling required for the SROC curve, individual information on sensitivity and specificity (and the degree of correlation between them) are lost.[60] Separate pooling of sensitivity and specificity does not allow for this correlation either, but nor does it allow for any variation with threshold.

The BVN model directly estimates the strength of the correlation between sensitivity and specificity by assuming a bivariate normal distribution (combined distribution of two correlated normally distributed variables) between their logit transforms. As logit sensitivity and specificity can be estimated by linear combinations of the HSROC parameters accuracy and threshold, the HSROC model indirectly assumes the same bivariate normal distribution.[79]

4. The model should use a random effects approach, i.e. should consider both between-study and within-study variability

The advanced model(s) properly estimate random effects and properly test for the significance of any effects to account for variability beyond chance (heterogeneity). The BVN model uses a random effects approach in the estimation of summary estimates of sensitivity and specificity and their corresponding 95% confidence intervals. Logit sensitivities and specificities from individual studies are each assumed to be approximately normally distributed around some mean value with a certain variability around this mean estimated. [59] This takes into account the heterogeneity beyond chance between studies. The HSROC model fits threshold and accuracy parameters as random effects, so that their average value and variation across studies are estimated. The random effects model takes account of the TPR and FPR pairs within studies, thereby taking account of the correlation between them. With both approaches, the unexplained between-study variability can be either modelled with covariates and/or be considered random due to unknown sources of variability.

The Moses approach produces a fixed effect estimate of DOR and no estimate for average threshold or variability, but it places no restriction on threshold. It does not account for the variability in sensitivity and specificity as the modelling is undertaken only using the log DOR and S, the proxy measure of threshold. Any between-study variability is not directly modelled and can only be explained by covariates – any remaining unexplained variability is not estimated.

2.4.1 Summary

In summary, the advanced methods have several theoretical advantages over the Moses method, making their results more statistically reliable and accurate. They provide additional information on threshold and shape and the significance of any changes in sensitivity and specificity, avoiding any perceived need for separate meta-analyses using both pooling and SROC methods, which may give inconsistent results.[60] They also estimate the size of the variance in all of the parameters. Furthermore, the drawing of 95% confidence ellipses around the average operating point should enhance our understanding of the heterogeneity between studies and the correlation within studies.

2.5 Outline of thesis and research questions to be addressed

Chapter 1 has established the extent to which there is potential for spectrum variation to impact on the results of both primary studies and systematic reviews. As it is rare for diagnostic accuracy studies to be sufficiently large in size or to recruit a sufficiently broad spectrum of participants to allow the influence of spectrum to be examined, systematic reviews that include all available studies of a given test for a given disorder provide the best means available to assess the impact from heterogeneity, notwithstanding the limitations of the primary studies in terms of design and reporting. Until recently, our ability to investigate sources of heterogeneity have been limited, in some cases by the ability of the methods available, but also by their accessibility. Recent work has made the advanced methods much less computationally demanding and therefore more accessible, and it is timely that their ability to investigate sources of heterogeneity should be examined and compared with more commonly used methods. These methods are explained in detail in Chapter 2.

Chapter 3 reports a methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy to date. This provides a picture of current practice in systematic reviews and meta-analyses in terms of how often spectrum effects have been considered in systematic reviews, whether and how they have been investigated, and what impact if any they have had on test accuracy.

In Chapter 4, a case study of the identification of spectrum effects comparing three meta-analytic methods is reported. A systematic review of two polymerase chain reaction (PCR) tests for the detection of active pulmonary tuberculosis is used to demonstrate the ability of the Moses SROC method,[54] the HSROC method[55,56,60] and the BVN model[57-59] to investigate sources of heterogeneity. Conducting a systematic review from scratch allows one to become intimately familiar with the data in the analysis, such that it is clear which data of interest were actually provided by study authors in the original study publications and also allows fully systematic methods to be employed throughout the review process. This particular dataset was chosen because the two tests are both commercially produced and are fairly standardised in terms of their application, thereby reducing one potential alternative source of heterogeneity. The studies are also generally very large in size, were well-designed and fairly recently published, this reduces to some extent the degree of methodological heterogeneity introduced into the review.

Chapter 5 reports the results of a re-analysis of previously published systematic reviews using both of the Moses methods and the HSROC model. The BVN model was not used as it gives results almost identical to the HSROC model assuming parallel curves and it cannot easily model an interaction of covariate with curve shape as the HSROC model does. Systematic reviews presenting contingency table data plus data on at least one spectrum-related factor per study were analysed. In this way, it was possible to investigate whether the findings from Chapter 4 were replicated across a large sample of datasets and also allowed a more thorough examination of effects from spectrum-related characteristics across a range of tests and conditions.

Chapter 6 presents the discussion and conclusions.

3 A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy

This chapter reports a methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy to date. The first part of the chapter provides a picture of recent practice in systematic reviews and meta-analyses in terms of how often spectrum effects have been considered, whether and how they have been investigated, and what impact if any they have had on test accuracy. The second part of the chapter looks at the same questions for reviews using the advanced methods of meta-analysis that have been published over the last four years.

3.1 Methods

3.1.1 Eligibility criteria

To be included, reviews must have evaluated a diagnostic or screening test by including studies that compared a test to a reference test with the aim of establishing test accuracy. Studies were assessed for inclusion by one reviewer.

3.1.2 Literature search

The Centre for Reviews and Dissemination's Database of Abstracts of Reviews of Effects (DARE) was used to identify existing systematic reviews of diagnostic studies. This is a database of quality assessed systematic reviews identified by hand searching key major medical journals, regular searching of bibliographic databases and by scanning grey literature since 1994[e].

Diagnostic reviews indexed on DARE up to April 2001 had already been screened to identify diagnostic reviews for a previous project[21] and were automatically included. Reviews indexed between April 2001 and December 2002 were also screened for inclusion. Only the reviews for which structured abstracts had been written were considered eligible. Due to the considerable time lag in loading reviews onto DARE at the time of the search, additional high quality systematic reviews not yet indexed on DARE but meeting the inclusion criteria were included. These were identified from sources such as the INAHTA and MEDION[f] databases. Nineteen of the 32 reviews identified from these two databases have since been added to the DARE database.

[e] further details about DARE can be found at http://agatha.york.ac.uk/darehp.htm

[f] INAHTA is the International Network of Agencies of Health Technology Assessment and MEDION is a database of diagnostic systematic reviews updated by a group of Dutch and Belgian researchers.

None of the reviews identified from the searches described above were found to have used the advanced methods of meta-analysis. As a considerable time interval had passed since these searches and as reviews using the advanced methods were known to have been published in more recent years, additional citation searches of the Science Citation Index and Social Science Citation Index were conducted in November 2007 to identify any systematic reviews that had used the advanced methods of meta-analysis. This was carried out to allow the examination of how these reviews dealt with issues of heterogeneity and spectrum in comparison to the previously identified set of reviews. Reviews identified from these searches were analysed subsequently to and separately from the main dataset.

3.1.3 Data extraction

A data extraction form for recording relevant information from each systematic review was designed and piloted. Data were extracted on a variety of items including:

- the experimental test, reference tests and condition tested for;
- the review methodology including the literature search and approach to quality assessment;
- review synthesis methods and approach to statistically identifying heterogeneity in study results
- methods of exploration of variability in study results and variables investigated.

The full systematic reviews were pre-screened independently by two reviewers. Those meeting the inclusion criterion were data extracted by one reviewer and the completed data extraction forms checked against the full paper by a second reviewer. Any disagreements were resolved by consensus or by referral to a third reviewer if necessary.

3.1.4 Data synthesis

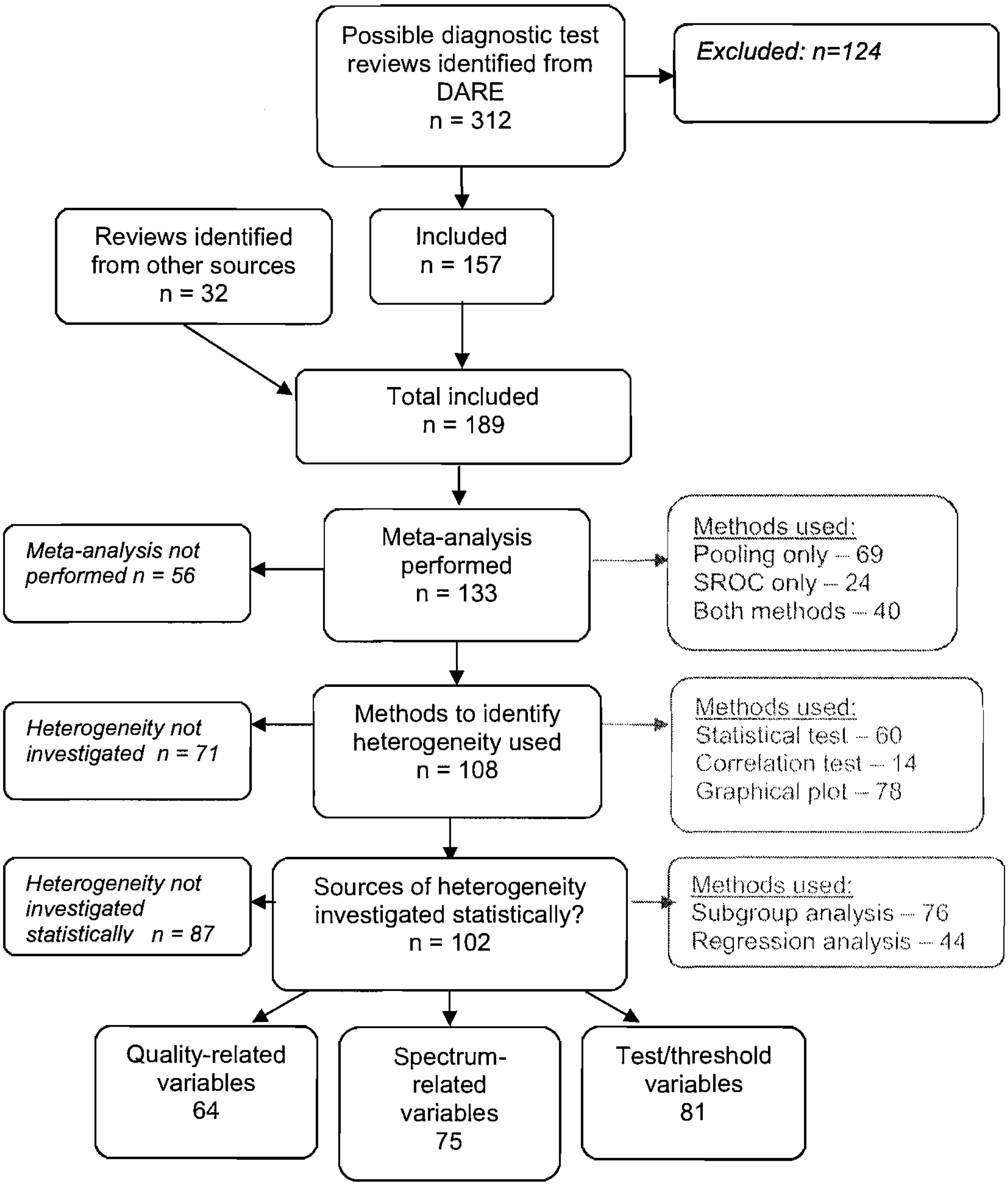
A narrative synthesis is presented. The reviews are considered primarily in terms of the statistical methods used, and the results section is structured to reflect the steps involved in the synthesis of diagnostic test accuracy studies, i.e.

- identification of heterogeneity
- meta-analysis
- investigation of sources of heterogeneity.

Particular focus is given to the extent to which spectrum effects or bias are considered, both as part of quality assessment and spectrum-related characteristics investigated in subgroup or regression analyses. The frequency of investigation of spectrum-related variables as sources of heterogeneity in relation to test or quality-related variables is quantified as is the frequency of reporting of statistically significant effects.

The reviews are considered in two parts: first of all looking at the main sample of reviews identified from the initial searches; and secondly looking at the reviews that used the advanced methods of meta-analysis.

Figure 10 Flowchart of review inclusion process



Possible diagnostic test reviews identified from DARE n = 312
Excluded: n=124
Reviews identified from other sources n = 32
Included n = 157
Total included n = 189
Meta-analysis not performed n = 56
Meta-analysis performed n = 133
Methods used:
Pooling only – 69
SROC only – 24
Both methods – 40
Heterogeneity not investigated n = 71
Methods to identify heterogeneity used n = 108
Methods used:
Statistical test – 60
Correlation test – 14
Graphical plot – 78
Heterogeneity not investigated statistically n = 87
Sources of heterogeneity investigated statistically? n = 102
Methods used:
Subgroup analysis – 76
Regression analysis – 44
Quality-related variables 64
Spectrum-related variables 75
Test/threshold variables 81

3.2 Results – Reviews using established methods of meta-analysis

3.2.1 Summary of reviews identified

Of 312 systematic reviews identified from the DARE searches, 189 met the inclusion criteria and were included in the review. Figure 10 provides a flowchart of the review selection process. Summary details of the 189 included reviews, according to whether they used a narrative (n=56, 30%) or a statistical method of synthesis (n=133, 70%), are provided in Table 7 to Table 11; fuller details of the reviews are available elsewhere.[74]

3.2.2 Description of review methods

The reviews cover a wide range of target disorders and test types, from the low technology of clinical examination for the detection of diseases such as left sided heart failure,[82] deep vein thrombosis[83] or carpal tunnel syndrome[84] at one end to highly equipment intensive tests such as nucleic acid amplification tests for detecting infection,[85-87] or positron emission tomography for the detection of cancer or Alzheimer's disease.[88]

Just over half (52%) of all reviews included searched only one electronic database (Medline) to identify primary studies (Table 7). This was less often the case for narrative reviews (38%) compared to those using statistical syntheses (59%). 59% of reviews used language restrictions in their searches; in 84% of these this was to restrict studies to English language only. Only 14% (27/188) of reviews applied no language restrictions. These proportions were similar regardless of whether the reviews carried out narrative or statistical syntheses.

Half of reviews applied inclusion criteria to restrict studies to those of a higher standard on at least one quality criterion. Most commonly this was to ensure that studies had compared the index test to an appropriate reference standard (86% of meta-analyses and 48% of narrative reviews applying quality-related criteria). The next most commonly used criteria were to ensure blinding had been used (19%), to include only prospective studies (16%) and to ensure verification bias had been avoided (15%). Restriction to higher quality studies was more common in meta-analyses than in narrative reviews and meta-analyses were more likely to apply more than one quality-related criterion.

Quality assessment of included primary studies was reported to have been carried out in 69% of reviews (Table 7), with most (88/131) using a quality assessment tool apparently developed by the authors themselves (only 43 reported using a previously published tool). An analysis of the way in which these reviews considered spectrum-related factors as part of their quality assessment is provided below; a further detailed analysis of all of the items included in a sample of these quality assessment tools is provided by Whiting and colleagues.[21] A further three reviews carried out a classification of evidence such as that outlined by the US Preventive Services Task Force.[89]

Table 7 Summary of reviews found

		TOTAL n (%)	Statistical synthesis n (%)	Narrative synthesis n (%)
Total no. of reviews		189	133 (70%)	56 (30%)
Review methods				
Medline only *electronic* source		99 (52%)	78 (59%)	21 (38%)
No. using language restriction	Restricted	111 (59%)	78 (59%)	33 (59%)
	English only	94 (84%)	63 (80%)	31 (94%)
	No restriction	27 (14%)	20 (15%)	7 (13%)
	Not stated	51 (27%)	35 (26%)	16 (29%)
No. using quality restrictions	Restricted	94 (50%)	69 (52%)	25 (45%)
	Appropriate reference test	71 (76%)	59 (86%)	12 (48%)
	Blinding used	18 (19%)	16 (23%)	2 (8%)
	Prospective only	15 (16%)	9 (13%)	6 (24%)
	Avoids verification bias	14 (15%)	12 (17%)	2 (8%)
	Adequate sample descrip	9 (10%)	5 (7%)	4 (16%)
	Consecutive enrolment	8 (9%)	8 (12%)	0
	Adequate test descrip	2 (2%)	2 (3%)	0
	Complete follow-up	2 (2%)	2 (3%)	0
	No restriction	95 (50%)	64 (48%)	31 (54%)
No. using quality assessment	Not conducted	58 (31%)	40 (30%)	18 (32%)
	Conducted	131 (69%)	93 (70%)	38 (68%)
	Authors' own	88 (67%)	68 (73%)	20 (53%)
	Existing tool	43 (33%)	25 (27%)	18 (47%)
Median (IQR) no. of studies		18 (IQR 20)	22 (IQR 20)	11 (IQR 13)
	No. studies not reported	7 (4%) reviews	3 (2%)	4 (7%)
Median (IQR) no. of patients		3,161 (IQR 6,815)	4,007 (IQR 7,553)	1,726 (IQR 3619)
	No. patients not reported	68 (36%) reviews	34 (26%)	24 (43%)

IQR – inter-quartile range

The median number of studies included in the reviews was 18. Meta-analyses have a higher number with a median of 22 studies compared to 11 for narrative reviews. The number of patients included in the studies was not clearly reported in 36% of all reviews; less so for narrative reviews (not reported in 43%).

Consideration of spectrum-related items in quality assessment

Of the 131 reviews carrying out quality assessment, 51% (n=67) considered patient spectrum in some way. Appendix 7 provides a full description of the items per review. A summary of the items related to spectrum are described in Table 8.

Just over a third of reviews (37%; 48/131) required a judgement regarding the appropriateness of the spectrum of patients included in the study; most asked whether the spectrum had been appropriate or whether the patients included were representative of those to whom the test would be applied in practice. 16 of the 48 reviews in this group (33%) also specifically mentioned the nondiseased as well as the diseased group. Seven reviews asked whether both diseased and nondiseased patients were included and one wanted 'a continuous spectrum of patients that included normal patients',[90] but seven required that control groups should include patients with disorders 'commonly confused' with the target disorder or with 'signs suggestive' of the target disorder.

Table 8. Breakdown of quality assessment items related to patient spectrum

Quality assessment (QA) items related to:	No. (%) of reviews conducting QA (n=131)
'Appropriateness' of patient spectrum (with specific mention of the nondiseased group)	48 (37%) 16 (33%)
Spectrum or sample *described*	30 (23%)
Study setting described	23 (18%)
Participant sampling described	14 (11%)
Analysis of pertinent subgroups	4 (3%)
Spectrum-related items not included in quality assessment	64 (49%)

Thirty reviews (23%; 30/131) asked whether a description of patients characteristics had been provided by the study authors and 23 (18%) were interested in the setting in which the study had been undertaken. Participant sampling was considered in 11% (14/131) of reviews and analysis of pertinent patient subgroups in only 3% (n=4).

3.2.3 Description of statistical methods used

Summary details of the statistical methods used in the reviews are presented in Table 9.

3.2.4 Identification of heterogeneity

Heterogeneity is best identified by visual comparison of study results on a graphical plot. Statistical tests to identify heterogeneity and threshold effects are available, however these lack power. A reasonable recommendation is to assume that heterogeneity and threshold effects are present, and that it therefore does not make sense to test for their presence. Instead, random effects models that allow for these features should be employed.

Graphical plots to identify heterogeneity

Over half (75/133, 56%) of meta-analyses used graphical plots to demonstrate the spread in study results. In 79% (59/75) of cases study results were plotted in ROC space, 13 reviews plotted sensitivity and/or specificity on forest plots and three reviews used 'D vs. S' plots.

Only two of the 56 reviews using a narrative synthesis presented study results graphically, all using ROC plots.

Statistical tests to identify heterogeneity

Statistical tests to identify heterogeneity were used in 60 of 189 (32%) of reviews (Table 9). Of the 133 reviews using statistical syntheses, 55 (41%) used statistical tests to identify heterogeneity, 34 using the Chi square test and 7 Fisher's exact test. A further five meta-analyses made a narrative statement regarding the presence of heterogeneity. Fewer meta-analyses (16%; 21/133) used correlation coefficients to test for a threshold effect, most (14) choosing the Spearman correlation coefficient.

Table 9 Summary of statistical methods used

		TOTAL n (%)	**Statistical synthesis n (%)**	**Narrative synthesis n (%)**
Total no. of reviews		189	133 (70%)	56 (30%)
Statistical methods used				
Test for heterogeneity reported		60 (32%)	56 (42%)	4 (7%)
	Chi	36 (60%)	34 (61%)	2 (50%)
	Fisher	7 (12%)	6 (11%)	1 (25%)
	Breslow-Day	5 (8%)	4 (7%)	1 (25%)
	Q statistic (ORs)	3 (5%)	3 (5%)	0
	Kardaun-Kardaun	1 (2%)	1 (2%)	0
	observed v predicted values	6 (10%)	5 (8%)	1 (25%)
	miscellaneous tests[a]	5 (8%)	5 (7%)	0
	test used but not reported	8 (13%)	8 (14%)	0
Test result	Statistically significant	47 (78%)	44 (79%)	3 (75%)
	Not significant	10 (17%)	10 (18%)	0
	Not reported	4 (7%)	3 (5%)	1 (25%)
Correlation test for threshold effects		21 (11%)	21 (16%)	0
	Spearman correlation		14 (67%)	
	Pearson correlation		3 (14%)	
	Kardaun-Kardaun		1 (5%)	
	test used but not reported		2 (10%)	
Correlation test result	Significant correlation		14 (67%)	
	No correlation		6 (29%)	
	Not reported		1 (5%)	
Study results plotted graphically		77 (41%)	75 (56%)	2 (4%)
	ROC plot	57 (74%)	59 (79%)	2 (100%)
	forest Se and/or Sp	12 (16%)	12 (16%)	0
	forest DOR or log DOR	1 (1%)	1 (1%)	0
	D vs S plot	3 (4%)	3 (4%)	0
	miscellaneous plots[b]	12%)	9 (12%)	0
Type of synthesis used	*Narrative*		0	56 (100%)
	Pooling methods		109 (82%)	
	Sensitivity/Specificity		97 (84%)	na
	LRs		26 (24%)	na
	PVs		11 (10%)	na
	DOR		10 (9%)	
	Effectiveness score		8 (7%)	
	Accuracy		5 (5%)	
	AUC		3 (3%)	
	Miscellaneous[c]		4 (4%)	
	SROC:		64 (48%)	
	Weighting not specified		27 (42%)	na
	Unweighted		13 (20%)	na
	Inverse variance weighted		11 (17%)	
	Sample size weighted		6 (9%)	na
	Variance weighted		1 (2%)	na
	'Weighted'		7 (11%)	na
	Robust resistant regression		2 (3%)	na
	Estimated from DOR or ES		3 (5%)	
	Data presentation:			
	DOR		4 (6%)	
	AUC		10 (16%)	
	SROC parameters		7 (11%)	
	Q*		18[e] (28%)	
	Se or Sp at fixed Sp or Se		20 (31%)	
	SROC curve only presented		10 (16%)	
	Comparison of ≥ 2 curves		4 (6%)	
	Other methods[d]		2 (3%)	
Paired data considered separately (meta-analyses only)	Yes		12 (9%)	
	No		42 (32%)	
	No paired data (or can' tell)		79 (59%)	

		TOTAL n (%)	Statistical synthesis n (%)	Narrative synthesis n (%)
Method of investigating heterogeneity	Not done	17 (9%)	10 (8%)	7 (13%)
	Narrative	68 (36%)	19 (14%)	49 (87%)
	Subgroup	74 (39%)	74 (56%)	na
	Regression	45 (24%)	45 (34%)	na
	Method not described	2 (1%)	2 (2%)	na

ROC – receiver operator characteristic; Se – sensitivity; Sp – specificity; DOR – diagnostic odds ratio; LR – likelihood ratio; PV – predictive value; AUC – area under the curve; ES – effect size;

[a] including effectiveness score (2 studies); comparison of fixed vs. random effects results (1 study); 'covariate adjustment' (1 study); and goodness of fit test (1 study).

[b] including: funnel plots using ES (1 study) or log DOR (1 study); scatterplots of AUC (1 study), LR (1 study) or Se (1 study) per study; Se (1 study) or NPV (1 study) plotted against prevalence; Se and Sp as function of prevalence (1 study); and Se/Sp plotted against sample size (1 study).

[c] including: fraction positive (1 study); correlation coefficient (1 study), Youden index (1 study), odds of false-negative on index vs reference test (1 study)

[d] Including: ratio of ORs (1 study); estimation of LR, method not reported (1 study)

[e] in two reviews LR was estimated from Q*

Five of the reviews using a narrative synthesis used statistical tests to identify heterogeneity (Table 9), four of which reported that statistically significant heterogeneity was found. A further four reviews specifically stated that the studies were too heterogeneous to be pooled, though no formal evidence for this was provided.

Identification of heterogeneity according to type of synthesis used

Of the 133 (70%) reviews in which meta-analysis was performed, 52% (n=69) carried out statistical pooling alone, 18% (24) conducted only SROC analyses, and 30% (40) used both methods of statistical synthesis (Table 10). None of the included reviews used the more advanced methods of meta-analysis outlined in Chapter 2 above, i.e. the BVN model and the HSROC model. Although 57% (76/133) of meta-analyses presented study results graphically, these were primarily reviews that had used SROC regression models: only 19/69 (28%) using statistical pooling alone presented results graphically.

Table 10 Statistical tests and graphical approaches used according to method of synthesis

	Narrative syntheses	Statistical syntheses	Statistical syntheses by method of synthesis used		
Type of synthesis	**56 (30%)**	**133 (70%)**	**Pooling only** 69 (52%)	**Pooling and SROC** 40 (30%)	**SROC only** 24 (18%)
Graphical presentation of results	2 (4%)	76 (57%)	19 (28%)	35 (87%)	22 (92%)

Many meta-analyses using SROC methods stated that these methods allow for the presence of a threshold effect (37/64), so presumably did not see the need to specifically test for threshold effects.

3.2.5 Type of syntheses used

Narrative syntheses of data

A narrative synthesis was used in 56 (30%) of reviews. In eight reviews the authors indicated that this was due to the presence of between-study heterogeneity, but the remainder did not state whether they had considered using statistical approaches to study synthesis.

Meta-analyses of sensitivities and specificities, predictive values and likelihood ratios

Of the 109 reviews that pooled accuracy indices, 87% pooled sensitivity and/or specificity, 23% pooled likelihood ratios and 10% pooled predictive values. A further 5% of reviews pooled test 'accuracy', which is the percentage of diagnoses that were correct (i.e. number true positive plus number true negative as a proportion of all test results).

Pooled single summaries of test performance

Single summaries of test performance, estimated by pooling results from individual studies or by logistic regression methods (akin to fixed effects pooling) were carried out in only a handful of studies: 9% (10/109) pooled diagnostic odds ratios; 7% (8/109) pooled the 'effectiveness score' (akin to the DOR), and 3% (5/109) pooled area under the curve data from individual studies.

Single summaries of test performance using SROC regression models

For those reviews presenting SROC curves, all except four used regression models such as that described by Moses and colleagues[54] to create the curves. Three of the exceptions estimated SROC curves from the pooled DORs or effectiveness scores and the other did not describe the method used. For the remainder, the main differences between the models used are the weights chosen for the regression model. In 42% of cases (27/64) the use of, or choice of, weight was not provided by the review authors (Table 9). In 13 reviews (20%) the models were unweighted; in 17% inverse variance weights were used; and in 9% sample size weights were used. In a further 11% (6/64) models were simply described as 'weighted'.

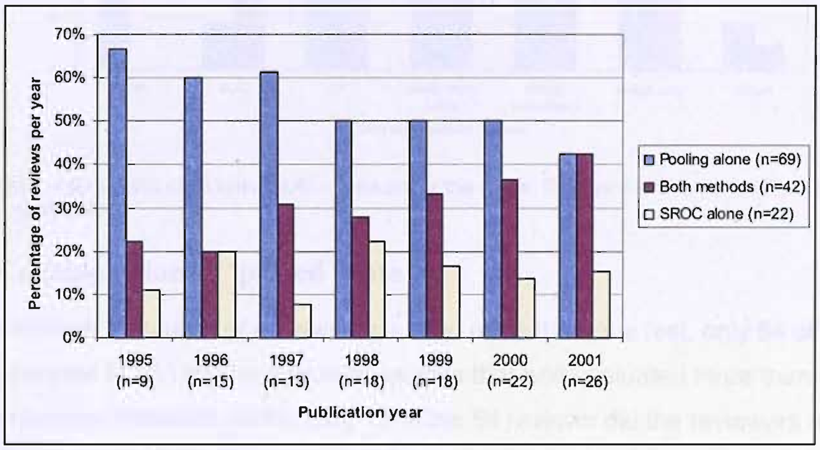
As discussed in section 2.3.2, SROC curves can be interpreted in several ways. The methods most commonly used were those that converted certain points of the SROC curve to sensitivity and specificity pairs (Table 9): the Q* (maximum joint sensitivity and specificity) was presented in 28% (18/64) of reviews, sensitivity and specificity pairs were 'read' from the SROC curves in 31% (20/64) of reviews, e.g. sensitivity at mean specificity or 95% specificity, or sensitivity and specificity at mean threshold. Ten reviews (16%; 10/64) chose to provide area under the curve (AUC) data and only four (6%; 4/64) interpreted the SROC curve as a DOR. The underlying SROC model parameters were provided by 11% (7/64) of reviews, 16%

(10/64) presented the SROC curve only with no summary statistics and 6% (4/64) compared two or more curves for different tests.

Type of statistical synthesis according to publication year

Figure 11 shows the proportion of reviews using each method of synthesis according to publication year for reviews published between 1995 and 2001 (insufficient numbers of reviews were available for other years). The proportion of reviews using statistical pooling alone has slightly declined over that time period (from 67% in 1995 to 42% in 2001, with a corresponding increase in the use of SROC methods (from 33% of all reviews in 1995 to 58% in 2001). However, two thirds of those using SROC methods have also carried out statistical pooling rather than presenting only SROC models (42/64). The tendency to carry out both methods in the same review has on the whole increased over time.

Figure 11 Type of meta-analytic method used by publication year

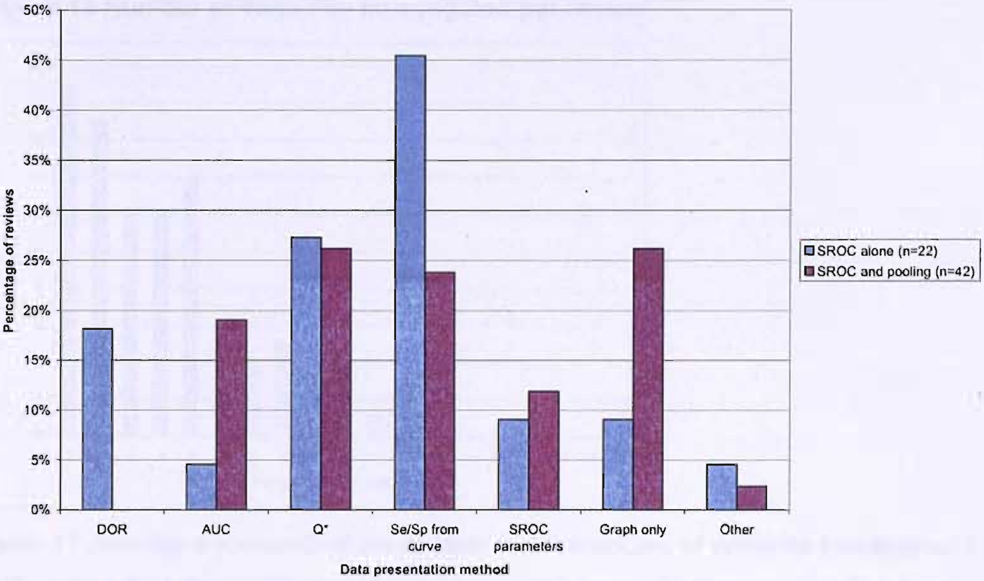


Data presentation according to type of syntheses used

Given the difficulties in the clinical application of SROC curves it was hypothesised that where SROC analysis alone was used, reviews would be more likely to present the SROC results as some combination of sensitivity and specificity rather than using alternative means of data presentation.

Figure 12 shows a breakdown of methods of presenting SROC analyses according to whether or not statistical pooling was also performed. When only SROC analysis was carried out, reviews were more likely to report the results as pairs of sensitivity and specificity data (45% compared to 24% of reviews that also conducted pooling), providing some support for this hypothesis. It is not clear whether these sensitivity and specificity pairs were in fact 'read' from the SROC curve or were actually estimated by some form of averaging. It seems possible that the point estimates reported were computed by pooling sensitivities and specificities and may not have been points on the ROC curve. When both pooling and SROC analysis were reported to have been carried out (i.e. where the pooled estimates were clearly presented), reviewers were more likely to present area under the curve data, less likely to present DORs and more likely to simply present the curves themselves with no further interpretation.

Figure 12 Means of presenting results of SROC analyses (n=64)



DOR – diagnostic odds ratio; AUC – area under the curve; Q* - point where sensitivity=specificity; Se – sensitivity; Sp – specificity

Consideration of 'paired' data

Although a number of reviews evaluated more than one test, only 54 of the 133 meta-analyses (41%) included primary studies that had evaluated more than one test against a reference standard, and in only 12 of the 54 reviews did the reviewers attempt to deal with the fact that they had 'paired' data, e.g. by analysing the data from those reviews separately.

3.2.6 Investigation of sources of heterogeneity

Methods of investigating heterogeneity

Of the 56 narrative reviews, 49 (87%) carried out a narrative review of factors that might cause variation in the results of the primary studies and seven did not really appear to deal with the question of heterogeneity at all.

Of the 133 meta-analyses, 29 (24%) provided either a narrative discussion of factors affecting heterogeneity (19) or did not consider heterogeneity at all (10). The remaining 102 attempted to statistically investigate possible sources of variation: 74 (56%) using subgroup analysis and 45 (34%) using some form of regression analysis. Regression analyses were usually undertaken by extending the SROC regression model, though 10 reviews reported using logistic regression models and one used meta-regression. A further two did not report the method they had used. For those reviews using subgroup analyses, although several reported P-values for the differences between groups very few reported the test used to detect any statistically significant difference: seven reviews reported using a t-test or Mann-Whitney U test to compare subgroups, two used the chi-square test and three the Wilcoxon test (paired or unpaired).

Figure 13 Number of variables investigated per review

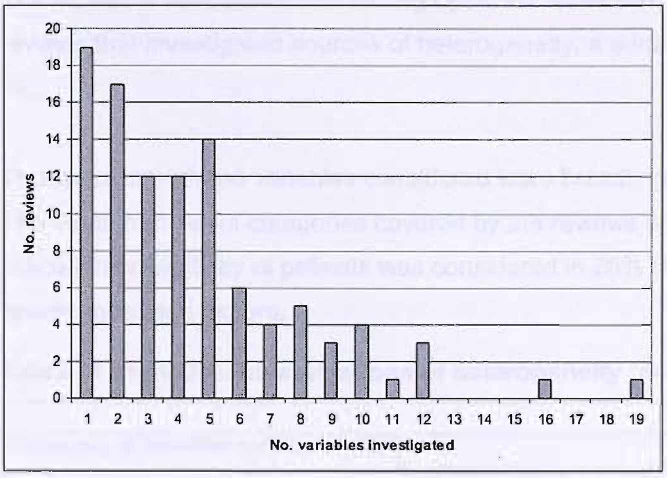


Table 11 provides a summary of the number and breakdown of variables investigated by the 102 reviews that statistically investigated possible causes of heterogeneity. The median number of variables investigated in these reviews was four, ranging from one in 20% of reviews to over six in 27% of reviews (Table 11). In general, a large number of variables were investigated in these analyses in comparison to the number of studies included in the review. The ratio of median number of variables to median number of studies was one to six. Only 38% of reviews complied with the typical recommendation to have at least 10 studies for every characteristic investigated.

Test and quality-related variables investigated

At least one quality-related variable was investigated in 63% (64/102) of reviews (Table 11). Within this subset of 64 reviews, the most commonly considered variables were use of blinding (41% of reviews; 26/64), sample size (33%; 21/64), the reference test used (28%; 18/64) and the avoidance of verification bias (25%; 16/64). The inclusion of an appropriate spectrum of patients and impact of study design chosen were among the variables considered in a small minority of reviews, 9% and 5% respectively. Around a third of reviews (36%) tried to look at the overall effect of study quality on accuracy, for example by classifying studies as low, medium or high quality or by using the quality score to subdivide studies.

Test- or threshold-related variables were examined by 79% (81/102) of the reviews (Table 11). Most (69%; 56/81) considered items related to variations in the test used, for example by looking at the effect of variations in the field strength used in MRI, or in the level of expertise of the person interpreting the test. 38% (31/81) of reviews considered threshold by subdividing studies according to threshold used. Publication year, which could be a proxy for changes in a test over time or changes in the patient population tested, was considered important by 36% (29/81) of reviews.

Spectrum-related variables investigated

The impact of clinical or socio-demographic variables were investigated in 68% (69/102) of reviews that investigated sources of heterogeneity, a summary of which is provided in Table 12.

The spectrum-related variables considered were broadly grouped into eight main categories. The mean number of categories covered by the reviews was 1.8 (range 1-6). The clinical indication or eligibility of patients was considered in 28% (19/69) of reviews that looked at spectrum-related factors.

Table 11 Statistical investigations of heterogeneity (n=102)

		N (%) of reviews
Median no. of variables considered (IQR)		4 (IQR 4)
	% considering only 1 variable	20 (20%)
	% considering 2 to 5 variables	55 (54%)
	% considering > 6 variables	28 (27%)
Ratio of median no. variables investigated to median no. studies included		1 : 6
	Reviews with ratio < 1:10	63 (62%)
Categories	**Variables investigated:**	
Quality-related variables	Not investigated	38 (37%)
	Investigated	64 (63%)
	Blinding	26 (41%)
	Sample size	21 (33%)
	Ref test used	18 (28%)
	Verification bias	16 (25%)
	Consecutive enrol	12 (19%)
	Prosp/retrospective	9 (14%)
	Spectrum	6 (9%)
	Disease prog bias	4 (6%)
	Sample description	3 (5%)
	Cohort/case-control design	3 (5%)
	Other items	9 (14%)
	Quality 'rating'/score	23 (36%)
Test- or threshold-related variables	Not investigated	21 (21%)
	Investigated	81 (79%)
	Test	56 (69%)
	Threshold	31 (38%)
	Publication year	29 (36%)
Clinical or socio-demographic variables	Not investigated	33 (32%)
	Investigated	69 (68%)

IQR – interquartile range

For example, in a review of ultrasonography for detecting peripheral arterial disease, de Vries[91] and colleagues examined whether clinical indications for testing included peripheral arterial disease only or whether other diagnoses were considered as well. Oosterhuis and colleagues[92] in a review of mean corpuscular volume for vitamin B12 deficiency also grouped studies according to clinical indication, i.e. whether patients were in a screening setting, had the test ordered to exclude B12 deficiency as part of treatment, and whether patients were considered to have pernicious anaemia. Other reviews looked at particular elements of the clinical indication. For example De Bruyn and colleagues[93] and De Bernardinis and colleagues[94] examined whether the presumed underlying causes of the target disorder (cirrhosis[93] and acute pancreatitis[94] respectively) affected test accuracy.

In a review of exercise tests for detecting coronary artery disease in women, Kwok and colleagues[95] examined whether exclusion of patients with history of myocardial infraction, baseline ECG abnormalities, or taking digoxin explained heterogeneity. Berry and colleagues[96] examined whether the inclusion of asymptomatic patients affected the accuracy of spiral and electron beam CT for the detection of hepatic lesions, pulmonary embolus or CAD, and Leitich and colleagues[97] examined the impact of including multifetal gestations on the accuracy of cervicovaginal fetal fibronectin for predicting preterm delivery.

Fifteen (22%) of the 69 reviews examined studies according to the symptom status (usually asymptomatic or not) or risk status of participants, for example references [87,97-102]. A further 15% (n=10) considered disease severity or stage, such as in references [76,103-107]. Over a quarter of reviews examined specific demographic characteristics such as age (26%; 18/69) for example references [84,108-112] or sex (13%; 9/69), for example reviews [84,95,113,114].

Table 12 Summary of spectrum-related heterogeneity investigations

			Number (%) of reviews (n=69)
Category of spectrum-related variable	Clinical indication/eligibility		19 (28%)
	Symptoms/risk status		15 (22%)
	Disease severity/stage		10 (15%)
	Age		18 (26%)
	Sex		9 (13%)
	Prevalence		21 (29%)
	Setting/source of pts		20 (29%)
	Sampling/study design		7 (10%)
Mean no. of spectrum-related categories investigated			1.8 (range 1-6)
Results clearly presented?	Yes		33 (48%)
	Partially		21 (30%)
	No		14 (20%)
Significant effect identified from	Spectrum	Yes	41 (59%)
		No	28 (41%)
	Test	Yes	20 (29%)
		No	31 (45%)
		not investigated	18 (26%)
	Quality	Yes	17 (25%)
		No	30 (43%)
		not investigated	22 (32%)

Prevalence was investigated as a source of heterogeneity in 29% (21/69) of reviews and 29% (20/69) also considered the setting or source of participants as a variable. In some reviews, for example Peters and colleagues,[115] Berger and colleagues,[106] or Hoffman and colleagues,[116] these compared accuracy in patients from a general or screening population to a referred population, while in others variation in the geographical setting was considered, such as the reviews by Kinkel and colleagues,[117] Loy and colleagues,[118] or Visser and colleagues.[119] In seven reviews, the method of sampling (consecutive or random versus other) or study design (case-control vs case series) were also considered. Although these were included as quality-related features they have also been included here as they can impact on the spectrum of included subjects.

3.2.7 Result of heterogeneity investigations (for reviews that examined spectrum-related factors)

This section considers only those reviews that examined spectrum-related factors (n=69).

The results of the heterogeneity investigations were clearly presented in just under half of this subgroup of reviews (48%; 33/69) (Table 12). In a further 30% (21/69), results were partially presented. In many cases, only results for those variables having a statistically significant impact were presented in detail; the other variables were described as having no significant impact, for example references [85,99,100,113,120]. In other reviews, only the P-values for the differences between subgroup were given and the full regression results or pooled accuracy in subgroups were not provided, for example in references [107,117,121]. In the remaining 14 reviews (20%) the results of the heterogeneity investigations were not presented in detail but were discussed narratively, usually by listing which variables did and did not have a significant impact on results.

Of the sample of reviews that considered spectrum-related variables in their heterogeneity investigations, 59% (41/69) found these variables to have a significant impact on test accuracy (Table 12), this is in comparison to 29% (n=20) that found an effect from test-related variables and 25% (n=17) that detected an effect from quality-related variables. Six reviews included 'avoidance of spectrum bias' as an item in their quality assessment of studies and examined the effect of meeting this criterion on accuracy; five reviews reported a non-significant effect[90,96,113,116,122] and one[123] did not report the result.

Fourteen reviews reported both their results in detail and looked at spectrum-, test- and quality-related covariates.[87,94,98,109,114,116,123-129] Of these, eight (57%; 8/14) found a statistically significant impact from spectrum-related factors, eight (57%) from test-related and six (43%) from quality-related covariates. These reviews used a variety of methods to investigate heterogeneity including looking at pooled sensitivity, specificity, log DORs and effect sizes in subgroups and adding covariates to SROC regression models.

3.3 Results - Reviews using advanced methods of meta-analysis

As no reviews using advanced methods of meta-analysis were identified from the original searches for this chapter, subsequent searches were undertaken to identify more recently published reviews known to have used these methods. These were examined according to whether spectrum effects were considered, how they were investigated, and what impact if any they have had on model parameters.

3.3.1 Summary of reviews identified

The citation searches identified 27 potential systematic reviews using the advanced methods of meta-analysis. 10 of these were excluded as they did not use the advanced methods and the copies of four could not be obtained in time.

Thirteen reviews using advanced methods of meta-analysis were identified. A summary of the reviews is provided in Table 13 and Table 14 and details of review methods, analysis methods and results are given in Appendix 6 to Appendix 8.

3.3.2 Review methods

The reviews predominantly examined the accuracy of imaging tests such as ultrasound, CT, MRI and PET-scanning (10 reviews). Other tests evaluated were cytology or biochemical markers.[80,130,131] The most commonly investigated topic was the diagnosis or staging of various forms of cancer (5 reviews).[80,132-135]

Table 13 Summary of reviews using advanced methods

		TOTAL n (%)
Total no. of reviews		13
Review methods		
Medline only *electronic* source		1 (8%)
No. using language restriction	Restricted	5 (38%)
	English only	1 (8%)
	No restriction	8 (62%)
	Not stated	0 (0%)
No. using quality restrictions	Restricted	0 (0%)
No. using quality assessment	Not conducted	0 (0%)
	Conducted	13 (100%)
	Authors' own	7 (62%)
	Existing tool	6 (38%)
	QUADAS[136]	5 (38%)
Synthesis method used		
	BVN	10 (77%)
	HSROC	3 (23%)
Heterogeneity investigation	Not conducted due to insufficient studies	4
	Univariate analyses	9
	Multivariable model developed	6

QUADAS – Quality of Diagnostic Accuracy Studies tool; BVN – bivariate normal model; HSROC; hierarchical SROC model

Only one review (8%) relied on a single electronic database (Medline) to identify primary studies (Table 13) compared to 59% (78/133) of the meta-analyses in the main 'DARE sample' . Language restrictions were used in 5 (38%) reviews compared to 59% of the main sample. Only 1 (8%) restricted studies to English language only compared to 47% (63/133) in the main sample.

Table 14 Summary of heterogeneity investigations

Study	Param eters	Type of characteristics investigated			No. invest-igations[a]	No. significant results[c]	Detail of spectrum-related factors investigated
		Spectrum	Design	Test			
Bipat, 2003[132]	Se,Sp	0	6	0^b	48	None	None
Bipat, 2004[133]	Se,Sp	0	7	0^b	77	None	None
Bipat, 2005[137]	Se,Sp	1 Adequate description of patients	9	0^b	60	12 (20%)	Outcome - Diagnosis Helical CT – sufficient description of patient popl (sens P<0.05 and spec P<0.01) MRI – sufficient description of patient popl (sens P<0.01) US – sufficient description of patient popl (sens P<0.01 and spec P<0.01)
Bipat, 2005a[134]	Se	1 Adequate description of patients	8	0^b	36	Per patient 4 (8%) Per lesion 6	Per lesion MRI (1.0T) spectrum of patients was representative of patients in practice (regression coefficient P<.001)
Glas, 2003[80]	Se,Sp	5 Adequate description of patients Type of controls BCG therapy % haematuria Tumour differentiation	6	1	66	8 (12%)	Sensitivity and specificity not correlated with spectrum-related variables (data not shown). Correlations with cohort versus case-control designs were observed, however.
Koelemay, 2004[138]	Se,Sp	0	5	1	12	1 (8%)	
Shaheen, 2007[130]	lnDOR	4 median age %men inclusion of HIV/HCV co-infected patients prevalence of significant fibrosis/cirrhosis	4	1	18	3 (17%)	APRI accuracy for detecting significant fibrosis not affected by patient-related factors: Age of study population (P=0.1), sex (P=0.96), prevalence of significant fibrosis (P=0.46), inclusion of HIV/HCV co-infected patients (P=0.60) For detection of cirrhosis, APRI accuracy was greater in studies containing higher proportion of men (P=0.001), younger participants (P=0.04), and HIV/HCV co-infected patients (P=0.03). The other

Study	Parameters	Type of characteristics investigated			No. investigations[a]	No. significant results[c]	Detail of spectrum-related factors investigated
							covariates were not significant (data not given).
Whiting, 2006[139]	α, θ	0	1	0	1	1 (100%)	
Williams, 2007[140]	α, θ, β	7 Severity of renal artery stenosis Hypertension and other features Hypertension with or without chronic renal failure Hypertension moderate or unspecified Hypertension and peripheral vascular disease Transplant recipient Peripheral vascular disease	8	3	72	2 (3%)	Population characteristics had no significant effect on test performance (data not shown)

Se – sensitivity; Sp – specificity; lnDOR – natural log of the DOR; α - accuracy parameter; θ - threshold parameter, β - shape parameter

[a] number of covariates x number of tests x number of outcomes

[b] separate subgroup assessment according to test characteristics

[c] listed if spectrum-related

None of the included reviews applied inclusion criteria to restrict studies to those of a higher quality standard, although one[139] did restrict most of their analyses to cohort studies only. Quality assessment was conducted in all of the thirteen reviews: five used the QUADAS tool or a modification of it, one used a list published by the Cochrane Methods Group on Diagnostic tests and seven did not state a source for the tool. 70% (93/133) of meta-analyses in the main sample used some form of quality assessment.

The number of studies included ranged from 8[131] to 90,[133] the latter including 299 datasets.

3.3.3 Statistical methods

Ten reviews employed the BVN model for their analyses and three the HSROC model (Table 14). Seven of the reviews using the BVN method included authors based in the Academic Medical Centre at the University of Amsterdam,[80,132-135,137,138] reflecting the fact that the adaptation of the BVN method to make it more 'user friendly' was also carried out there.[59]

Twelve of the thirteen reviews presented forest plots of sensitivity and/or specificity[132,134,137,141] or of DOR[130] or ROC plots of individual studies[80,130,131,133,137-140,142] to demonstrate the presence of heterogeneity. Only the review by van Westreenen and colleagues[135] did not provide any graphical presentation of data. 56% of the DARE sample presented data graphically.

Investigation of sources of heterogeneity was carried out in 9 reviews. The remaining four recognised the presence of heterogeneity in their reviews but did not consider that they had sufficient studies to investigate the causes.[131,135,141,142] The characteristics investigated were generally affected by the quality of reporting in the primary studies. Most reviewers were unable to carry out all of their planned investigations.

3.3.4 Result of heterogeneity investigations

All nine reviews conducted univariate analyses to determine the variables that individually had a significant effect on results, usually to $P<0.10$. Six reviews went on to develop multivariable models including characteristics that individually had a significant effect on test accuracy.[80,132-134,137,138]

Of the seven studies employing the BVN model that also investigated sources of heterogeneity, five examined the effect of the variables on sensitivity and specificity, one considered effects on sensitivity only, and another[130] did not appear to investigate heterogeneity within BVN model framework (Table 14). Shaheen and colleagues[130] used a random effects meta-regression model to examine the effect of the covariates on the natural log of the DOR. Of the two reviews using the HSROC framework that also investigated sources of heterogeneity, one assumed no interaction of covariate with curve shape and therefore examined the effect of on accuracy and threshold only[139] and the second examined effects on the accuracy, threshold and shape parameters.[140]

Most reviews focused on the investigation of study design and analysis related characteristics (Table 14). Five reviews included spectrum-related variables. In two cases this was confined to a determination of the adequacy of the description of patients. Statistically significant results were obtained for less than 20% of investigations. Of the five reviews investigating spectrum-related characteristics, three[130 134,137] found significant effects (Table 14). In two of these the factors investigated were quality related, in one review whether the studies gave a sufficient description of the patient population[134] and in the other, whether the spectrum of patients was representative of patients in practice.[137] For the review by Shaheen and colleagues,[130] the heterogeneity investigations do not appear to have been undertaken within the BVN model framework, but a separate random effects regression model to examine the effect of covariates on lnDOR developed.

3.4 Discussion

Due to the timing of the literature seaches undertaken for this chapter, the main focus of it is on reviews published up to 2002, however as there are currently only a small number of published reviews using advanced methods the findings from the main sample of reviews are likely to apply to most reviews published to date.

Spectrum-related issues

Overall, spectrum-related factors appear to be under-considered in systematic reviews of diagnostic tests. Of the 189 reviews included, only 35% (n=67) considered spectrum as part of quality assessment and 36% (n=69) investigated the potential impact from spectrum-related factors statistically. These percentages increase when one considers only those reviews that actually carried out quality assessment (51% of which included spectrum-related criteria) those that carried out meta-analysis (52% of which investigated spectrum-related covariates), and those that reported reported carrying out heterogeneity investigations (68% of which investigated spectrum-related covariates) nevertheless these percentages are still low especially when one considers that heterogeneity of study findings is common.

Of the reviews that included spectrum-related criteria in their quality assessments, the majority (72%; 48/67) required a judgement on the 'appropriateness' or otherwise of the included patients only 16 of which (33%) specifically mentioned the appropriateness of the nondiseased patients. Only four reviews included an item on whether pertinent subgroups had been investigated in the primary studies. This is arguably a more important aspect of the generalisability of a study as inclusion of an appropriate or representative sample may mask the fact that a test's discriminatory capacity varies between subgroups.

The sample of reviews included do not allow any strong conclusions regarding the potential impact from spectrum on test accuracy to be made. Of the reviews that reported having considered spectrum (n=69), only 48% (n=33) clearly reported the results of the heterogeneity investigations and 59% (n=41) stated that they had identified a statistically significant impact from the spectrum-related variable in question. When the small number of reviews that considered spectrum-, test- and quality-related variables were examined (n=14), the proportions finding statistically significant effects from each of these categories were similar (57%, 57% and 43% respectively). Although the total number of reviews was small, this does suggest that spectrum is an equally important source of heterogeneity that should at least be considered if not always investigated. It is worth noting however that there will be an unknown number of reviews that did investigate spectrum-related or other characteristics but that did not find any significant effects and therefore did not report having carried them out. There is a difference between how often spectrum actually matters and how often it has been reported to matter.

Review methods used

The preferential use of the pooling approach is not least because of the challenge in reporting SROC methods, as the results are not easily interpreted by clinicians. Clinicians tend to prefer to have point estimates of the sensitivity and specificity of a test, whereas ROC curves describe a series of estimates.[143] The majority (82%) of reviews carrying out statistical syntheses opted to pool aspects of test performance independently, i.e. separate pooling of accuracy indices, with little consideration paid to the possibility of a threshold effect, whilst 48% of meta-analyses undertook SROC regression analyses either alone, or in combination with the pooling approach. Investigation of sources of heterogeneity was undertaken most commonly either by pooling data according to subgroups (56%) or by extending the SROC regression model with the addition of covariates (34%).

The addition of a covariate to a regression model produces a regression coefficient for that covariate that is akin to the relative diagnostic odds ratio, i.e. the extent to which the DOR would be increased or decreased in the presence of that covariate. For example, Fleischman and colleagues[109] found that amongst studies of exercise echocardiography for the detection of coronary artery disease both increasing age and later publication year led to significant decreases in DOR (univariate analysis results: RDOR -0.22 per year; 95%CI: -0.31, -0.12 and -0.41 per year; 95%CI: -0.58, -0.24). This sort of information is not particularly meaningful to many clinicians, therefore other reviewers have attempted to overcome this problem by presenting their results as some combination of sensitivity and specificity at given points on the SROC curve.

Whitsel and colleagues[114] in a review of the QTc (or heart rate corrected QT interval) for the detection of autonomic failure in people with diabetes, estimated accuracy in each subgroup using separate SROC models (similar to assuming the presence of an interaction of covariate with curve shape) and used this data to estimate the relevant sensitivities for each subgroup at the overall pooled specificity of 86%. Others have reported Q*, or point of maximal joint sensitivity and specificity. The clinical relevance of these points is not always clear and in some cases the points chosen appear rather arbitrary. Furthermore a series of potential operating points were never quoted; this could be a real deficiency for asymmetric SROC curves where the DOR varies along the curve. The other real problems for reporting of heterogeneity investigations is that some covariates will affect the diseased group more than the nondiseased such that any differential impact on sensitivity and specificity will be masked by presenting the regression coefficients or RDOR.

The problems are further highlighted in a review of the Papanicolau test for cervical precancer. Fahey and colleagues[125] both added covariates to an SROC model and pooled sensitivity, specificity and DOR in the same subgroups for four spectrum-related, test-related and quality-related variables, and reported their results in detail. In general, the pooled DOR analyses indicated differences in the same direction but of a slightly different magnitude than suggested from the regression analyses (Appendix 11). The exception was earlier publication year which resulted in an increase in the pooled DOR by about a half, whereas the regression analysis found no impact from publication year. These differences were accentuated when inverse variance weighting was used in the regression model. At the same time pooled sensitivity dropped from 68% in the earlier studies to 58% while specificity increased from 64% to 70%. Although these results don't account for differences in threshold as mentioned earlier, it is possible that a change in the way the test was applied or in the population tested over time affected sensitivity and specificity in opposite directions such that the overall DOR was not affected. This example demonstrates how potentially conflicting results can be produced from the same data according to method of meta-analysis.

Use of advanced methods

Problems with the most commonly used methods can potentially be overcome using advanced statistical methods. No reviews in the main sample attempted to pool studies using these methods, despite them having been available since 1995. This is likely to be because of the time needed for new and more complex methodologies to diffuse into routine practice, but may also may reflect difficulties in applying methods in unconventional software such as WinBUGS. The development of the advanced methods to make them more easily accessible[59,60] has led to the publication of at least 13 reviews using the advanced methods since 2003.

The overall standard of these reviews is very high and a big improvement on that found in the main sample of reviews. This is less likely to be a reflection of increasing knowledge of best practice review methods than the fact that the review authors are mainly based in centres of academic excellence, some at the forefront of development of these methods. The HSROC method has been less used than the BVN model but that too is a reflection of the affiliations of the respective authors.

Consideration of heterogeneity in the reviews was of a high standard. Spectrum-related characteristics were investigated in five of the nine reviews that investigated sources of heterogeneity, but in two cases only the presence of an adequate description of patients was examined. This is very likely due to lack of recording or reporting in the primary studies. Of the five reviews that included details on spectrum-related characteristics, three found statistically significant effects from these variables, supporting the finding from the main dataset that when spectrum variables are reported to have been considered they are often found to have a significant effect.

General comments on investigation of heterogeneity in diagnostic test systematic reviews

Graphical plots to demonstrate the presence of heterogeneity are rarely reported in reviews using narrative syntheses of diagnostic test accuracy (reported in 4%) and furthermore, are not always reported in reviews using meta-analytic techniques (reported in 57%). Graphical presentation of results was mainly carried out by those conducting SROC analysis, e.g. individual study results as well as the SROC curve were presented in ROC space. Of those authors opting only to pool data, less than a quarter (19/69) used any form of graphical presentation of results, only nine of which presented data on a ROC plot thereby demonstrating any potential correlation between sensitivity and specificity.

Given the high degree of heterogeneity amongst diagnostic test studies, graphical presentation of individual study results are a useful aid to conveying complex information, even in reviews choosing to use a narrative synthesis – a perfectly defensible option where studies are highly variable. Plotting pairs of sensitivity and specificity in ROC space is an easy way to display heterogeneity of both indices as well as allowing potential threshold effects to be detected. It is also true however that visual examination of study results to identify heterogeneity also has limited power to detect bias if the number of studies is small. At the very least, reviewers should explicitly acknowledge and assess the potential for heterogeneity to be present. It is encouraging that all 13 reviews using the advanced methods included some form of graphical presentation of data.

The wide variation in methods chosen to combine the results of primary studies again perhaps reflects uncertainty in the most appropriate methods to use and also greater familiarity with more traditional indices of test accuracy (e.g. sensitivity and specificity). It would be extremely difficult to make a judgement as to whether or not the approach taken by the individual reviewers was appropriate or not without looking at the primary studies, but some issues are of particular concern.

Narrative reviews may have been carried out due to assumed but unreported heterogeneity, or due to insufficient numbers of studies (the median number of studies in narrative reviews was only half of that in meta-analyses), but few reported having considered the option of using statistical synthesis. Although the median number of studies may have been lower, in principle many did include a sufficient number of studies to consider meta-analysis. Given that considerable heterogeneity is a given in diagnostic test meta-analysis, some discussion of it is warranted regardless of the synthesis method chosen. Furthermore, reviewers should recognise that a justification for the approach chosen, whether narrative or statistical, should be provided in systematic reviews.

For those carrying out statistical syntheses, most opted to pool aspects of test performance independently, as discussed above, with little consideration paid to the possibility of a threshold effect. Similarly to the presence of heterogeneity, many would agree that a correlation between sensitivity and specificity is to be expected in diagnostic reviews. As might be expected, around half of those using SROC approaches (37/64) stated that they did so because this technique allows for any threshold effect. It is likely that the results for many reviews that only carried out pooling of sensitivity and specificity or likelihood ratios would differ if methods that allow for heterogeneity and threshold variation were employed.

Given the likely presence of heterogeneity, it can certainly be argued that potential sources of heterogeneity should always be investigated in systematic reviews of diagnostic test accuracy studies. This should be limited by the number of studies in the review and will also be limited by the level of reporting in the primary studies. It is unclear, even for reviews of intervention effectiveness, how many covariates can reliably be investigated, and how this might depend on the number of studies, the extent of the heterogeneity and the relative weights awarded to the different studies.[144] For the investigation of characteristics affecting primary study results a ratio of one variable for every 20 participants is often recommended; for systematic reviews, a ratio of one variable for every 10 studies is more usual. Three-quarters of meta-analyses included in the main sample here attempted to investigate sources of heterogeneity. On average one characteristic was investigated for every six studies included in these reviews; indicating likely over-investigation of study characteristics.

The most appropriate choice of variables to be investigated will depend on the specific context of the review and the included studies however spectrum-, test- and quality-related variables should at the very least be considered for investigation in any review. Quality-related variables were considered in less than two-thirds of reviews in this sample. Blinding, sample size and overall quality classification were the most commonly considered criteria. Half of all reviews only included studies that met certain quality-related criteria and so may have decided that further investigation of the effect of quality on accuracy was not warranted. However, poor reporting on the part of the authors of the primary studies and until recently[139] the fact no standard quality assessment tool has been available will partly explain this under-investigation. A tool for the quality assessment of diagnostic studies developed using standard scale development techniques is now available.[21] The authors hope that as well as providing a standardised tool for systematic reviewers, the project may also play a role in bringing about greater awareness regarding the important quality issues involved in diagnostic accuracy studies and help to raise the standards of such trials.

Poor reporting is a particular problem with diagnostic accuracy studies such that it is often difficult to ascertain what procedures to avoid bias were actually followed by study authors. The STARD initiative[30] (Standards for Reporting of Diagnostic Accuracy) aims to promote the completeness and quality of reporting of diagnostic accuracy studies similarly to the CONSORT statement for reports of RCTs.[31] Greater awareness of methodological principles for diagnostic accuracy studies will also help inform the design and analysis of primary studies.

Nearly all reviews focus on undertaking meta-analyses comparing the results of a new test with a reference standard. Very few reviews analysed only studies which compared results of several tests in the same patients with a reference standard and only 12/54 (22%) reviews that included at least some 'paired' data on two or more tests considered those studies separately. One can argue that heterogeneity will be less likely to be so problematic in meta-analyses of within study comparisons between tests, as many of the factors (such as the patient group) will be identical for both tests. Statistical methodology for investigating heterogeneity and threshold effects in studies of paired test comparisons requires further development, but may in time lead to more robust evidence about the relative performance of alternative diagnostic tests.

Other issues highlighted by this methodological review includes the significant potential for publication bias in these reviews – 84% restricted studies to those published in English only and 52% searched only one electronic database (Medline). Publication bias is known to be a real problem in reviews of therapeutic interventions.[67,145] Although its extent has not yet been quantified for test accuracy reviews it seems likely that it will be as much, if not more of an issue for tests. The retrospective nature of many diagnostic test studies would imply that authors may only publish if they have found particularly good results with a test.

It has not been possible to study any variation in the standard of review methods within different areas of medicine or types of test. This is hard to categorise across reviews and numbers within sub-categories would be small.

A strength of this review was use of the DARE database. Systematic reviews have to meet a certain standard of methodological quality before being included on the database. Due to the considerable time lag in loading reviews onto DARE at the time of the original search, some reviews (32/189) identified from other sources were also included to try to make the sample as current as possible. Nineteen of the 32 reviews (59%) identified in this way have since been added to the DARE database however it is possible that some of the remaining 13 (7% of the total sample) may not have met DARE's quality standards. However in the main, the reviews included in this chapter are of higher quality than many that are published, so that systematic review standards may be worse in practice than has been shown here. Given that the majority of reviews in the main sample were published prior to 2002 and that comprehensive guidelines on carrying out systematic reviews of diagnostic tests were not published before 2001,[11,146,147] it is likely that review methods have improved significantly since that time. The superior quality of the reviews using the advanced methods of meta-analysis may partly reflect this, but as they were mainly carried out in centres of excellence, one would expect the general standard to be amongst the highest.

3.5 Conclusions

It is clear that a proportion of published meta-analyses use inappropriate methods of analysis. The likely presence of a correlation between sensitivity and specificity and of between study heterogeneity is ignored in both the analysis and presentation of their results, and in many cases average values of sensitivity and specificity (or occasionally likelihood ratios) are presented. There is a danger that these reviews may be disseminating a misleading message that implies consistency of test performance when in fact the data that they have collected clearly display inconsistency. Such inadequate analyses could in the worst instance lead to inappropriate diagnostic investigations, interpretations and the use of inappropriate interventions. Where people have reported investigation of spectrum effects the majority have found statistically significant associations.

Where heterogeneity has been considered, the variability in approaches taken is a reflection of the level of difficulty and complexity of carrying out such reviews. The methodology is still developing and there is considerable uncertainty in the most appropriate techniques to use. High profile guidelines on undertaking diagnostic test reviews[11,40] should go some way to improving standards, as will the inclusion of diagnostic test accuracy reviews in the Cochrane Library. Nevertheless, carrying out many of the statistical analyses required for these reviews requires a high degree of familiarity with statistics and statistical software packages. There is as yet no truly user-friendly software package that can be used by non-statisticians in the way that packages such as RevMan is used for meta-analyses of therapeutic interventions. It is highly recommended that diagnostic test accuracy meta-analyses should not be carried out without the involvement of a statistician familiar with the field.

Difficulties with investigating heterogeneity at review level also points to the need for sufficiently large, prospective, well-designed multicentre studies that evaluate a number of diagnostic tests (or variations on a test), in order to establish test accuracy and also allow the investigation of the influence of patient characteristics on accuracy.

4 A case study comparing three meta-analytic methods

This chapter applies the three meta-analytic methods outlined in Chapter 2 to a dataset from a previous systematic review of diagnostic tests for the detection of tuberculosis[78] to illustrate similarities and differences in results, and finally to explore the effect of adding covariates to each model. The rationale for choosing the TB dataset and the main systematic review methods are provided in Appendix 9 and Appendix 10, respectively. Methods employed to compare across meta-analytic models are described below.

4.1 Methods used to compare meta-analytic models

Given the statistical rigour of the advanced methods, throughout this chapter they are treated as the benchmark against which the Moses methods can be assessed, as Harbord and colleagues did in their wider empirical evaluation of methods.[148] It is worth noting that the superiority of the advanced methods has not yet been empirically proven, however against the criteria outlined in 2.3.1, the advanced models appear best. The comparison of methods can be split into the comparison of the primary analyses of the complete dataset, and the comparison of the heterogeneity investigations across models. The models are examined in three ways, comparing:

a. the results of the unweighted and weighted Moses models
b. the HSROC and bivariate normal model results
c. the Moses model results and the advanced methods.

For the heterogeneity investigations, these comparisons are stratified by whether the models are assumed to have parallel SROC curves or 'crossing' SROC curves, i.e. where the covariate interacts with curve shape so that the SROC curves for the subgroups can have different shapes.

For all models, accuracy (DOR) was estimated at Q* (the point at which sensitivity=specificity) and at a point nearer to the centre of the data. The latter is estimated using the mean threshold of the studies in the dataset and is referred to as the "DOR at the average threshold". See section 2.3.2 "Estimation of sensitivity and specificity" for a discussion of the potential lack of representativeness of Q*. For the investigation of heterogeneity with 'crossing curves', the difference in accuracy between groups (RDOR) was estimated at Q* and at the average threshold of the studies in each group, i.e at the average threshold of the reference group and at the average threshold of the comparator group. Where parallel SROC curves are modeled the RDOR is constant all the way along the curves.

Three covariates were selected to compare the investigation of heterogeneity across the three models. The covariates examined were deliberately chosen to reflect increasing levels of complexity in results, illustrating the effect of the covariates on DOR, threshold, and shape. The covariates examined were as follows

- effect on accuracy (RDOR): *index test blinding*. For the TB dataset, index test blinding occurred where the PCR test in question was performed and interpreted without knowledge of the reference test results (actual diagnosis). Even tests such as these whose interpretation require a certain numerical threshold to be reached before being considered positive may involve some degree of subjective interpretation.
- effect on threshold: *test type*. Two tests were included in this dataset, MTD and Amplicor. Although operating on the same principles, these were developed by different manufacturers and may have different accuracy properties in the same way that different different drugs within a class can have.
- effect on curve shape: *reference test used*. There is no definitive reference standard for the detection of pulmonary tuberculosis. A commonly used reference standard is culture alone, however it is known that microbiologic studies of sputum for the detection of tuberculosis can fail to detect mycobacteria that may be picked up by PCR tests and will incorrectly classify patients with TB as false-positive results.[72] A compromise solution is to use a reference strategy, where the reference diagnosis is made on the basis of clinical information in combination with culture and other tests such as chest x-ray, however this may 'over-diagnose', and identify patients as having TB who in fact do not have the disease.

  Because the definition of 'diseased' is relatively tight when defined by culture results alone in comparison to where a combined reference strategy is used, and because PCR works on the same principle as culture (amplifying the presence of mycobacterial DNA as opposed to 'growing' it), there will be less variance in the distribution of PCR results when the presence of disease is defined by culture alone as opposed to a combined reference strategy.

The results of both the primary analysis and the heterogeneity investigations are presented primarily in tabular format, with the studies and SROC curves plotted in ROC space.

4.1.1 Comparison of the Moses methods

A key factor potentially leading to differences between the results of an unweighted and weighted Moses analysis is the effect of the weighting system used. The Moses model is commonly weighted by the inverse of the variance (or standard error) of the log of the DOR, i.e. the SE(lnDOR) which, as Deeks and colleagues have shown,[81] can be subject to bias. To demonstrate how bias can be introduced into the SE(lnDOR), it is broken down into three components for studies with the highest DORs:[81]

i) a sample size dependent term (SSdep) which reflects unequal numbers in diseased and nondiseased groups

ii) a proportion test positive term (PTPdep), the effect of which is minimised when the numbers of true negatives and false positives are equal (specificity 50%). As the balance of TNs and FPs changes due to increasing or decreasing threshold, the PTPdep term increases multiplicatively.

iii) a DOR dependent term (DORdep), which is 0 when DOR=1 (or where sensitivity=specificity) and which rises with DOR; the actual magnitude of the term decreases or increases with smaller or larger numbers of diseased, respectively

The formula for estimating the SE(lnDOR) is

$$SE(\ln DOR) = \mathrm{sqrt}(SSdep*PTPdep) + DORdep$$

The full formula and details on each of the components of SE(lnDOR) are given in Appendix 15. Only the sample size dependent term will operate appropriately (is unbiased) under the particular characteristics of diagnostic meta-analyses as follows:[81]

- DORs can be extremely high in value, often with very small or zero cells in the 2x2 contingency table. The SE(lnDOR) is an asymptotic estimate and therefore may be invalid where proportions are close to 1, as occurs with zero cells,
- individual studies often vary in the threshold for test positivity, and finally
- diagnostic studies often have unequal sizes of diseased and nondiseased groups.

To help demonstrate the presence of bias in the SE(lnDOR), scatter plots of study weight against effect (essentially funnel plots) are presented. Funnel plots are often used to look at publication bias, or small study effects.[149,150] The bias in this case, however, is not a bias in the data obtained, such as whether all of the studies have been identified or whether some are of lower quality, but is in the statistical method itself. Essentially, the calculation of the SE(lnDOR) does not appropriately reflect the precision or 'value' of large studies that show large effects.[81]

Three types of funnel plot are presented to help investigate whether the relationship between study weight and effect may explain the differences in results between the two Moses models. The first plots lnDOR against its standard error; if there is bias in the SE(lnDOR), studies showing large effects will have high standard errors. The second plots lnDOR against sample size to look for a sample size related effect. This will show whether studies with large effects and large standard errors also have low sample sizes. The third plots lnDOR against the inverse square root of the effective sample size (where the effective sample size (ESS) is the sample size needed in equal-sized groups to achieve the available power where there are groups of unequal sizes). The latter has been shown to provide a more robust indication of any sample size related effect in a diagnostic systematic review.[81]

The effect of the bias in the SE(lnDOR) on the results of the Moses unweighted and weighted analysis is then illustrated by plotting the two regression lines and the individual studies on a 'D versus S' plot and then by examining the effect on the analysis of removing the studies with the most biased SE(lnDORs).

4.1.2 Comparison of the Moses model against the HSROC model

To help understand whether the Moses and HSROC models are influenced by individual studies in the same way or whether they treat the studies differently, deletion residual analysis was employed. Essentially this removes each individual study in turn to identify the effect that this has on each of the model parameters. The results were then examined to identify any patterns or categories of studies having the biggest effects on the analyses, and whether these patterns were the same across models.

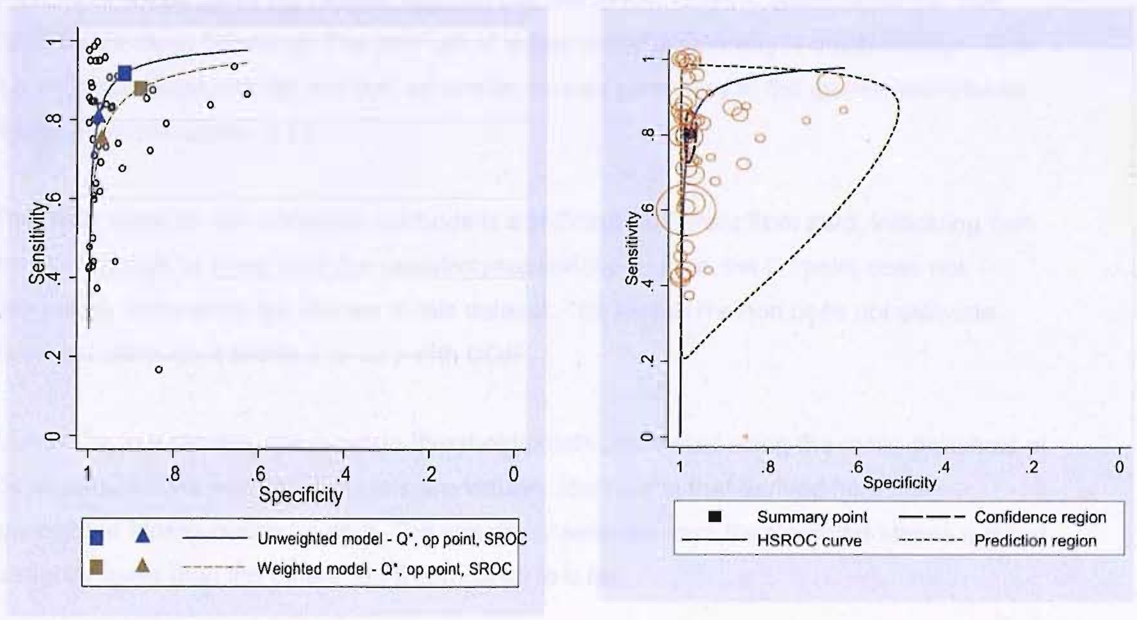
4.2 Primary analysis of the TB dataset

The results of the primary analysis of the 51 studies using the four models are given in Table 15 and displayed graphically in Figure 14. Details of all of the datasets included in the review are provided in Appendix 14.

Figure 14 ROC plots for the three meta-analytic methods

a. Moses methods

b. Advanced methods



Where Q* - point where sensitivity=specificity, op point - the operating point estimated using the mean value for 'S' across studies.

Figure 14 displays the ROC plots for the Moses and advanced methods. The apparent differences in the location of some of the studies between the Moses method plot and the plot for the advanced methods is due to the zero cell correction (the addition of 0.5 to every cell of a 2x2 table that contains a zero) that is needed in order to carry out the Moses analysis. The zero cell correction is not required for the advanced methods. The size of the circles in the plot for the advanced methods indicate the precision of the study; the smaller the circle, the more precise the study estimate. The figures show considerable heterogeneity between studies in both sensitivity and specificity and a considerable range in precision.

Table 15 shows that regardless of the parameterisation used, the two advanced models give near identical results across all parameters. There are considerable differences in DOR between the two Moses methods, with the unweighted model producing a DOR closest to that of the advanced models. These differences are explored in section 4.2.1 below.

Table 15: Main model parameters[a]

Method	**Mean accuracy (DOR) at Q***	**Mean accuracy (DOR) at OP**	**Mean threshold**	**Shape**	**Average sensitivity**[b]	**Average specificity**[b]
Moses method - unweighted	121 (52, 284)	181	x	-0.17, P=0.21	*0.81*	*0.98*
Moses method - weighted	53 (32, 88)	97	x	-0.26, P=0.01	*0.75*	*0.97*
HSROC	139 (76, 253)	198 (89, 307)	-0.79 (-1.30, -0.28)	0.35, P=0.06	*0.80 (0.75, 0.86)*	*0.98 (0.97, 0.99)*
Bivariate normal	*139 (76, 254)*	*198 (89, 306)*	*-0.80 (-1.30, -0.29)*	*0.35, P=0.06*	0.80 (0.75, 0.86)	0.98 (0.97, 0.99)

DOR – diagnostic odds ratio; Q* - point where sensitivity=specificity; OP – average operating point of studies, estimated using mean threshold

[a] – figures in *italics* denote derived values, i.e. parameters which are not the natural output from the model in question but have been transformed from model parameters

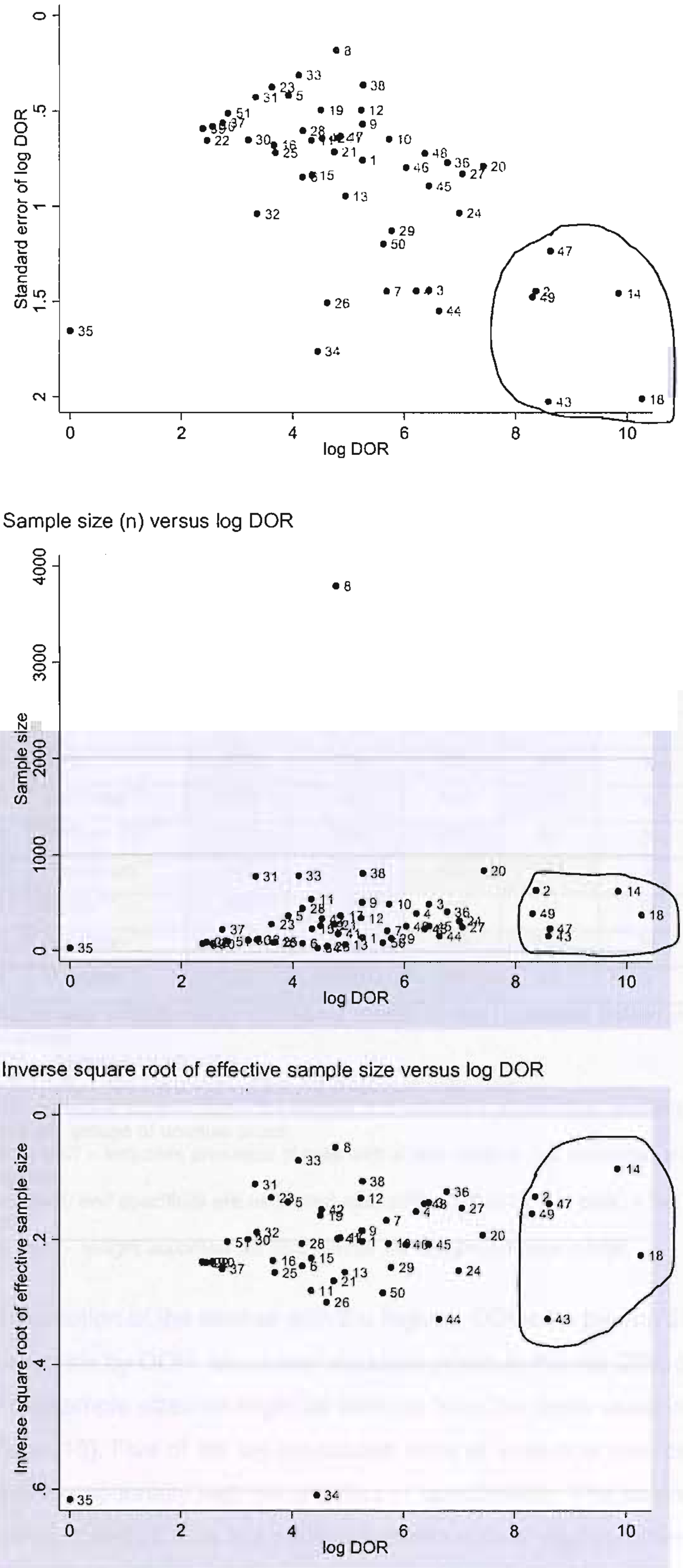
[b] – for the Moses methods, the average sensitivity and specificity are estimated using the mean of 'S' from the primary studies

In terms of shape, all of the models suggest that the SROC curve is asymmetric, i.e. that DOR varies along the curve. The strength of evidence for asymmetry is much stronger from the weighted Moses model and the advanced models compared to the unweighted Moses model (P-values closer to 0).

The theta value for the advanced methods is significantly different from zero, indicating that the study points lie away from the sensitivity=specificity line, i.e. the Q* point does not adequately summarize the studies in this dataset. The Moses method does not estimate threshold although it allows it to vary with DOR.

Across the four models, the average threshold points (estimated using the mean threshold of the studies) for the advanced models are virtually identical to that derived from the unweighted Moses method output. The sensitivity estimate from the weighted Moses method is slightly lower than the others (0.75 compared to 0.80).

Figure 15 Scatter plots of log of diagnostic odds ratio (DOR)



a. Standard error of log DOR [SE(lnDOR)] versus log DOR

b. Sample size (n) versus log DOR

c. Inverse square root of effective sample size versus log DOR

4.2.1 An exploration of the differences between Moses models

The scatter plot of lnDOR against its standard error reveals a relationship between DOR and its SE, with less precise studies having higher DORs (Figure 15a). The studies at the far bottom right of the plot (numbers 14, 2, 47, 18, 43 and 49) – the studies with the 6 highest lnDORs - have the most influence on this pattern. Plotting lnDOR against total sample size (Figure 15b), the six studies singled out above lie directly to the right of the main group of studies, suggesting no association of DOR with sample size. Plotting lnDOR against the inverse square root of the ESS (Figure 15c), as suggested by Deeks and colleagues[81] in fact shows the vast majority of studies located within a large group at the centre-top of the plot. The same six studies lie to the right of the main group, confirming no sample size related effect in this dataset.

Table 16: Studies with the highest quartile of diagnostic odds ratios

Id	Author	DOR in top 25%	SE (lnDOR)	Total n	ESS	Zero cells?	Specificity	Sensitivity	%weight
18	Devallois[151]	**28741**	**2.01**	372	79	fp + fn	**1.00**	**0.98**	**<0.01%**
14	Chedore[152]	**18969**	**1.46**	**618**	**533**	fn	0.98	**1.00**	**<0.01%**
47	Wang[153]	**5538**	**1.23**	230	**198**	x	0.99	**0.99**	**<0.01%**
43	Smith[154]	**5415**	**2.03**	153	37	fp + fn	**1.00**	**0.95**	**<0.01%**
2	Abu-Amero[155]	**4292**	**1.45**	**628**	**233**	fp	**1.00**	0.79	**<0.01%**
49	Yam[156]	**4045**	**1.48**	387	159	fp	**1.00**	0.86	**<0.01%**
20	Eing[157]	**1669**	0.79	**833**	108	x	**1.00**	**0.89**	**0.01%**
27	LaRocco[158]	**1145**	0.83	246	179	x	0.98	**0.95**	**0.01%**
24	Hoffner (b)[159]	**1088**	1.03	309	64	x	**0.99**	**0.88**	**0.01%**
36	Piersimoni[160]	**872**	0.77	**402**	**268**	x	**0.99**	0.85	**0.01%**
44	Smith[154]	**757**	**1.55**	153	37	fn	0.98	**0.95**	**<0.01%**
3	AlZahrani[161]	**630**	**1.44**	**489**	**204**	fp	**1.00**	0.42	**<0.01%**
45	Vuorinen[162]	**627**	0.89	256	93	x	**0.99**	0.85	**0.01%**

Shaded cells indicate values at or above median for that parameter; Bolded cells indicate values in top 25% for that parameter.
DOR – diagnostic odds ratio
SE(lnDOR) – standard error of the log diagnostic odds ratio;
ESS - effective sample size is the sample size needed in equal-sized groups to achieve the available power where there are groups of unequal sizes;
Zero cells? – indicates presence of cells with a zero value in 2x2 contingency table; fp – false positive, fn – false negative
Sensitivity and specificity are estimated after adding 0.5 to all four cells of 2x2 tables which have at least one zero cell.
%weight – weight accorded per study under the weighted Moses model

Examination of the studies with the highest DORs (in the top 25% of the dataset) shows that the top six by DOR also have standard errors in the top 25% of the dataset but do not have small sample sizes as might be inferred from the more usual interpretation of funnel plots (Table 16). Five of the top six studies have at least one zero cell in their 2x2 tables and all have exceptionally high sensitivities or specificities. The inverse relationship between precision and DOR is therefore not due to a small sample effect but is more likely to be explained by the estimates of SE(lnDOR) being overly influenced by extreme diagnostic threshold (high specificities) and/or high test accuracy.

4.2.1.1 Biased SE(lnDOR)

The upper section of Table 17 breaks down the SE(lnDOR) of the six studies with the highest DORs into its three components. The sample size dependent term which reflects unequal numbers in diseased and nondiseased groups, does not make a big contribution to the SE for the majority of these studies; only one (43 Smith) has a large SSdep term compared to the rest of the dataset.

The opposite is true of both the PTPdep term and the DORdep term. The effect of the PTPdep term on the SE is minimised when the numbers of true negatives and false positives are equal (specificity 50%). For the six studies under consideration here, the PTPdep value is above the median for all six studies and in the top 25% for four of them. Given that the smallest possible value for this term is four, all of the studies clearly have very high thresholds with values ranging from 50 up to 1127 (Table 17).

The DOR dependent term is zero when DOR=1 (or where sensitivity=specificity) and rises with DOR; the actual magnitude of the term decreases or increases with smaller or larger numbers of diseased, respectively. The DOR dependent term is above the median for five of the six studies and in the top 25% for four of them. For all six of these studies except Chedore (14), the term is negative indicating that specificity is higher than sensitivity for these studies.

Table 17: Breakdown of SE(lnDOR) for selected studies

id	Author	DOR	SE	Components of SE(lnDOR)			SSdep*
			(lnDOR)	SSdep	PTPdep	DORdep	PTPdep
Studies with six highest DORs							
18	Devallois[151]	**28741**	**2.01**	0.05	**703**	**-31.43**	35.15
14	Chedore[152]	**18969**	**1.46**	0.01	51	1.74	0.51
47	Wang[153]	**5538**	**1.23**	0.02	80	-0.10	1.6
43	Smith[154]	**5415**	**2.03**	**0.11**	**287**	**-26.56**	31.57
2	Abu-Amero[155]	**4292**	**1.45**	0.02	**1127**	**-17.24**	22.54
49	Yam[156]	**4045**	**1.48**	0.03	**685**	**-15.04**	20.55
Studies with six lowest DORs							
51	dos Anjos Filho[163]	17	0.51	0.04	7	-0.01	0.28
37	Piersimoni[160]	15	0.56	0.06	8	-0.02	0.49
40	Sato[164]	13	0.58	0.06	5	0.07	0.27
22	Gomez-Pastrana[165]	12	0.66	0.06	17	-0.20	0.94
39	Sato[164]	11	0.59	0.06	4	0.12	0.24
35	Osumi[166]	1	**1.66**	0.38	7	0	2.74

Shaded cells indicate values at or above median for that parameter; bolded cells indicate values in top 25% for that parameter
DOR – diagnostic odds ratio; SE(lnDOR) – standard error of the log diagnostic odds ratio; SSdep - sample size dependent term; PTPdep - proportion testing positive term; DORdep - DOR dependent term.

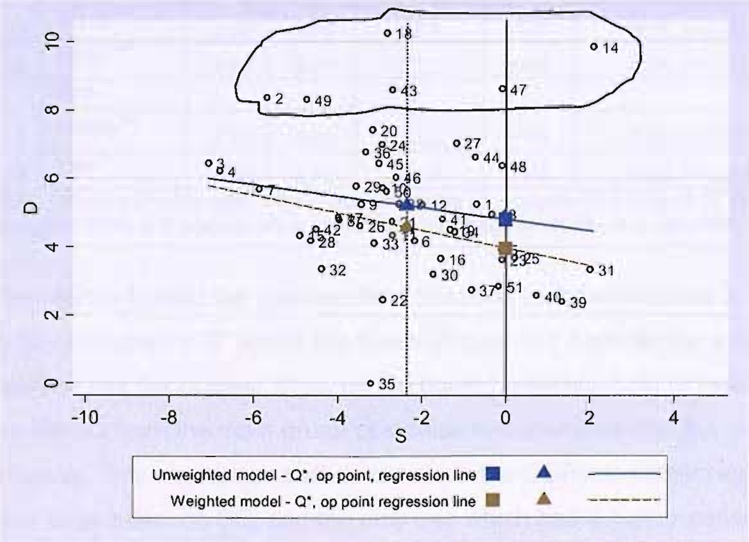
In contrast, for the studies in the bottom half of Table 17 (the six with the lowest DORs), where specificities are all below 0.95 and sensitivities are also generally lower (less than 0.86), the SSdep terms are all above the median, while the PTPdep and DORdep terms are below the median. The PTPdep terms are all very close to the lowest possible value (four). These studies are therefore not affected by zero or small cells or by threshold variation, therefore in most cases it is the SSdep term that drives the SE(lnDOR).

4.2.1.2 Effect of biased SE(lnDOR)

The presence of bias in SE(lnDOR) explains the differences in results between the unweighted and weighted Moses models. As the latter model is weighted by the inverse of SE(lnDOR), studies with high SE(lnDOR) get very little weight and vice versa for studies with low SEs. The final column of Table 16 lists the weights accorded to each study for the weighted Moses analysis; the weights for each of the 51 studies are given in Appendix 14. The unweighted Moses model by its nature gives all studies equal weight (i.e. 1/51 or 2%) whereas for the weighted Moses model, weights range from less than 0.01% to 21% (Appendix 14). From Table 16 one can see that all of the studies with the highest DORs receive 0.01% of the weighting or less, i.e. less than 200 times the 'weight' that they receive in the unweighted or equal weight analysis.

Figure 16 plots the lnDOR, or D, against S for all studies in the dataset. The studies with the six highest DORs and biased SE(lnDORs) are circled, lying above the main dataset. All of these studies receive weighting of less than 0.01% for the weighted model. This explains why the SROC curve for the weighted model in Figure 14a is considerably below that for the unweighted model and the DOR considerably smaller.

Figure 16 D vs S plot for all studies



Q* - point where sensitivity=specificity, op point is the operating point estimated using the mean value for 'S' across studies (mean S = -2.36).

Table 18 shows the effect of individually deleting each of the six studies from the overall pooled analysis using the Moses models. It is immediately noticeable that the deletion of two of the six studies has a relatively much larger effect on the overall pooled analysis compared to the others, and secondly that for all studies the effect is greater for the unweighted model. The latter observation would be expected due to the lower weighting attributed to these studies in the weighted model.

Table 18 Result of pooled analysis using Moses models minus each study with biased SE(lnDOR)

a. Unweighted Moses

id	Author	Individual study values		Pooled analysis minus individual studies				
		D	S	DOR *b'line* *121.1*	% change	S *b'line* *-0.17, P=0.21*	% change	sens, spec
18	Devallois[151]	10.27	-2.84	112.5	-7%	-0.16, P=0.22	-7%	0.80, 0.98
14	Chedore[152]	9.85	2.08	80.2	-34%	-0.29, P=0.03	73%	0.79, 0.98
47	Wang[153]	8.62	-0.09	101.5	-16%	-0.21, P=0.12	24%	0.80, 0.98
43	Smith[154]	8.60	-2.71	114.7	-5%	-0.16, P=0.22	-3%	0.80, 0.98
2	Abu-Amero[155]	8.36	-5.69	126.8	+5%	-0.13, P=0.36	-25%	0.81, 0.98
49	Yam[156]	8.31	-4.75	123.2	+2%	-0.14, P=0.31	-18%	0.80, 0.98

b. Weighted Moses

id	Author	Individual study values		Pooled analysis minus individual studies				
		D	S	DOR *b'line* *52.5*	% change	S *b'line* *-0.26, P=0.01*	% change	sens, spec
18	Devallois[151]	10.27	-2.84	52.2	-1%	-0.26, P=0.01	-1%	0.75, 0.97
14	Chedore[152]	9.85	2.08	48.3	-8%	-0.28, p<0.01	11%	0.74, 0.97
47	Wang[153]	8.62	-0.09	49.8	-5%	-0.27, p<0.01	5%	0.75, 0.97
43	Smith[154]	8.60	-2.71	52.3	+0%	-0.26, P=0.01	0%	0.75, 0.97
2	Abu-Amero[155]	8.36	-5.69	53.2	+1%	-0.25, P=0.01	-4%	0.75, 0.97
49	Yam[156]	8.31	-4.75	52.9	+1%	-0.25, P=0.01	-3%	0.75, 0.97

DOR – diagnostic odds ratio; b'line – baseline value for analyses including all 52 datatsets; % change - percentage change in DOR or S from baseline; sens, spec – average sensitivity and specificity

The reason behind the greater effect from two of the six studies is that they are the only two to have values for 'S' above the mean (Figure 16). In particular study 14 (Chedore), whose deletion has the biggest effect on the pooled analysis, both in terms of DOR and shape, stands out from the main group of studies and therefore has the greatest leverage on the analysis. This was the one study for which the DORdep component of the SE(lnDOR) had a very large effect on SE, and the only one which had a higher sensitivity than specificity. This effect is explored further in the next section.

None of the studies has a large effect on the sensitivity or specificity at the average threshold.

4.2.2 Moses versus HSROC comparison

The identification of bias in the SE(lnDOR) in several studies and the effect of this on the Moses models in section 4.2.1 above, suggests that certain studies may have a greater impact on the overall pooled analysis than others. Because the differences between the unweighted and weighted Moses models had already been investigated in some detail, a deletion residual analysis was undertaken to examine how the unweighted Moses model agrees with the HSROC model.

Twenty-one studies were identified whose removal affected at least one model parameter by 5% or more, either with the HSROC (19 studies) or Moses (20 studies) model analyses. The effect from these studies is summarised in Table 20 and plotted in Figure 17; full details are given in Appendix 13 (unweighted Moses) and Appendix 14 (HSROC). Eighteen of the 21 studies affected the parameters of both models by 5% or more, one (study 4[161]) affected only the HSROC model results and two (id 20[157] and 35[166]) affected only the Moses model results.

Table 19 Categories of study with biggest influences on model results

Category	Total no. of studies in dataset	No. having ≥5% effect on at least one model parameter
sensitivity greater than specificity	5	5
minimal (less than 5%) difference between sensitivity and specificity	11	9
high values (over 93%) for sensitivity	8	8
exceptionally high specificity (99.5% or more)	8	6
studies with zero false negatives	5	4
studies with zero false positives	8	5
studies with lowest sensitivities (40% or less)	2	2

Review of the deletion residual analysis for the two models suggested that these studies fall into seven categories (Table 19 and Table 20). In general, because the majority of studies in this dataset have specificity considerably greater than sensitivity, studies in the first two categories, i.e. with sensitivity greater than specificity or sensitivity close in value to specificity have by far the biggest effect on the analyses. The studies that lie around the edges of the dataset, i.e. those with more extreme values have the biggest effects.

For both models, the removal of study 14 (Chedore), which is positioned in the far top left of the ROC plot (Figure 17), has by far the biggest effect on all parameters, changing threshold, accuracy and shape by -17%, -21% and +46% for the HSROC model and accuracy and shape by -34% and +73% for the Moses model (Table 20). This study had the highest sensitivity (99.7%) in the dataset but also had an extremely high value for specificity (98.0%).

Table 20 Summary of deletion residual analysis: percentage change in model parameters following removal of each individual study

		HSROC: % change in			Moses (eq): % change in		1 sens>spec	2 min sespdiff	3 sens ≥ 0.934	4 spec ≥ 0.995	5 zero cell FN	6 zero cell FP	7 two lowest sensitivities
id	Author	accuracy (DOR)	shape	threshold	accuracy (DOR)	shape							
Studies whose removal had ≥5% effect on at least one model parameter for either HSROC or Moses analyses													
14	Chedore, 1999[11]	-21%	46%	-17%	-34%	73%	Y	-	Y	-	Y	-	-
47	Wang, 1999[153]	-12%	19%	-6%	-16%	24%	-	Y	Y	-	-	-	-
39	Sato, 1998[164]	10%	-19%	14%	+15%	-24%	Y	-	-	-	-	-	-
40	Sato, 1998[164]	9%	-14%	11%	+13%	-19%	Y	-	-	-	-	-	-
18	Devallois, 1996[151]	-9%	0%	0%	-7%	-7%	-	Y	Y	Y	Y	Y	-
31	Middleton, 2003[167]	7%	-15%	13%	+9%	-15%	Y	-	Y	-	-	-	-
51	dos Anjos Filho, 2002[163]	7%	-9%	7%	+9%	-12%	-	Y	-	-	-	-	-
48	Wang, 1999[153]	-6%	10%	-2%	-7%	10%	-	Y	Y	-	-	-	-
27	La Rocco, 1994[168]	-6%	8%	-2%	-7%	7%	-	Y	Y	-	-	-	-
35	Osumi, 1995[166]	x	x	x	6%	13%	-	-	-	-	-	-	Y
43	Smith, 1999[154]	-6%	-1%	0%	-5%	-3%	-	Y	Y	Y	Y	Y	-
44	Smith, 1999[154]	-5%	7%	-2%	-6%	8%	-	Y	Y	-	Y	-	-
23	Hoffner, 1996 (a)[169]	5%	-6%	6%	+6%	-8%	-	Y	-	-	-	-	-
20	Eing, 1998[157]	x	x	x	-2%	-5%	-	-	-	Y	-	-	-
5	Alcala, 2001[170]	4%	-5%	6%	+4%	-6%	-	Y	-		-	-	-
49	Yam, 1998[156]	-3%	-11%	3%	+2%	-18%	-	-	-	Y	-	Y	-

		HSROC: % change in			Moses (eq): % change in								
id	Author	accuracy (DOR)	shape	threshold	accuracy (DOR)	shape	1 sens>spec	2 min sespdiff	3 sens ≥ 0.934	4 spec ≥ 0.995	5 zero cell FN	6 zero cell FP	7 two lowest sensitivities
25	Kambashi, 2001[171]	3%	-4%	5%	+6%	-8%	Y	-	-		-	-	-
2	Abu-Amero, 2002[155]	-2%	-17%	5%	+5%	-25%	-	-	-	Y	-	Y	-
32	Mitarai, 2001 (a)[172]	1%	12%	-10%	1%	12%	-	-	-	-	-	-	Y
4	Al Zahrani, 2000[161]	1%	-8%	-3%	x	x	-	-	-	Y	-	Y	-
3	Al Zahrani, 2000[161]	1%	-6%	-3%	+2%	-6%	-	-	-	Y	-	Y	-
Studies in the same categories whose removal did not have ≥5% effect on any model parameter													
1	Abe, 1993[173]	x	x	x	x	x	-	Y	-	-	-	-	-
7	Bemer-Melchoir, 2000[174]	x	x	x	x	x	-	-	-	Y	-	Y	-
13	Cavusoglu, 2002[175]	x	x	x	x	x	-	Y	-	-	-	-	-
26	Kang, 2002[176]	x	x	x	x	x	-	-	-		-	Y	-
34	Neu, 1999[177]	x	x	x	x	x	-	-	-	-	Y	-	-

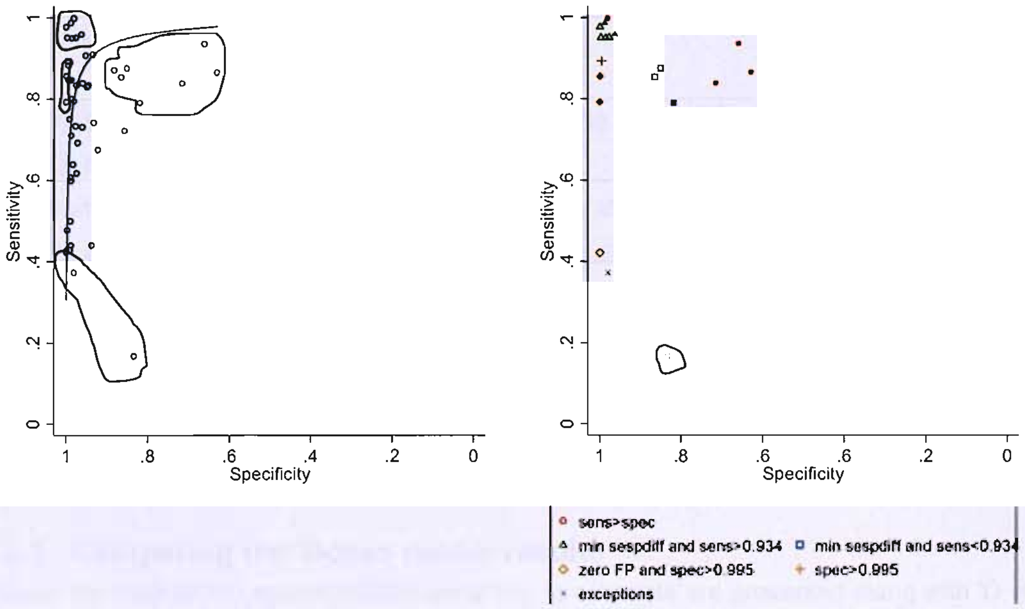
Studies are sorted by the magnitude of the effect on the pooled DOR using the HSROC analysis
x indicates studies whose removal does not impact on at least one parameter by 5% or more
1 sens>spec: sensitivity greater than specificity
2 min sespdiff: minimal difference between sensitivity and specificity
3 sens ≥ 0.934: sensitivity ≥ 0.934
4 spec ≥ 0.995: specificity ≥ 0.995
5 zero cell FN: zero false negative results
6 zero cell FP: zero false positive results
7 studies with lowest sensitivities

The shape (and threshold for the HSROC model) parameter was more commonly affected than DOR, with the removal of only four (three for HSROC) studies affecting DOR by 10% or more. Eleven studies impact on the shape parameter in the Moses model by 10% or more.

For the HSROC model, 9 studies affected shape by 10% or more; five of which also affect the threshold parameter by 10% or more (Table 20).

Figure 17 Plots of studies having effect of 5% or more on at least one model parameter

a. All studies b. Selected studies



Filled in symbols indicate effect of 10% or more

4.2.3 Summary

For this dataset, the advanced models (HSROC and BVN models) have near identical results when all studies are pooled together. However, there are considerable differences in results between the two Moses methods, in terms of both DOR and shape. Potential explanations for the difference between the two Moses methods are:

a. the zero cell correction that is needed in order to carry out the analysis, i.e. the addition of 0.5 to every cell of a 2x2 table that contains a zero. Adding the correction will attenuate the effect, more so in small studies. However the correction is added for both equal and weighted models so this is not a likely explanation here.
b. studies receiving the highest weight under the weighted model having the lowest DOR, possibly due to sample size, pulling the SROC further away from the top-left had corner of the ROC plot, where accuracy is the highest.
c. as the studies in the dataset generally had very high specificities and some also have exceptionally high sensitivities (e.g. studies with small or zero cells), it is more likely that the approximate variance of the log DOR is biased, as shown above. Weighting by inverse variance of lnDOR will therefore give less weight to the studies with the

highest specificities in particular. This can effect both the magnitude of the DOR and the shape of SROC curve.

For this dataset the DOR from the unweighted Moses model was closest to that from the advanced models, although the shape of the curves were different. Whether or not this is a consistent finding will be explored in the next chapter.

Studies having the most effect on the analyses for both the unweighted Moses and HSROC models were those for which sensitivity was higher than specificity or sensitivity was close in value to specificity, especially at higher levels of sensitivity. The shape term was most commonly affected.

4.3 Investigation of heterogeneity in the TB dataset using the 3 methods

The results of the heterogeneity investigations are presented in tabular format in Table 21, Table 22 and Table 23, first with no shape interaction allowed, i.e. parallel curve models (top half of tables), and then allowing for a shape interaction, i.e. crossing curve models (bottom half of tables). The associated SROC curves are plotted in Appendix 15 to Appendix 17. For the investigation of heterogeneity, the comparison between models focuses on comparing SROC curves between groups as opposed to comparing operating points. Where there are multiple thresholds, a comparison of operating points is not useful.

4.3.1 Comparing the Moses model results

Plots of the SE(lnDOR) against lnDOR according to covariate are presented along with 'D vs S' plots for both parallel and crossing curve models. To assist in the comparison of Moses model results, the six studies in the dataset with the highest DORs are circled according to covariate.

4.3.1.1 Moses models comparison: Index test blinding

When studies are examined by the presence or absence of index test blinding, the weighted Moses model shows a much smaller, and nonsignificant difference between groups compared to the unweighted model (Table 21). This difference is maintained with or without the interaction of covariate with shape, and regardless of whether RDOR is estimated at Q* or at the respective average threshold points. Figure 18 shows that all six studies with biased SE(lnDOR) identified from section 4.2.1 above (the group of six studies to the far bottom right of the plot) fall into the reference case group (blinding not described).

Table 21 Difference in model parameters: blinded index test interpretation (comparator) versus blinding not reported (reference)

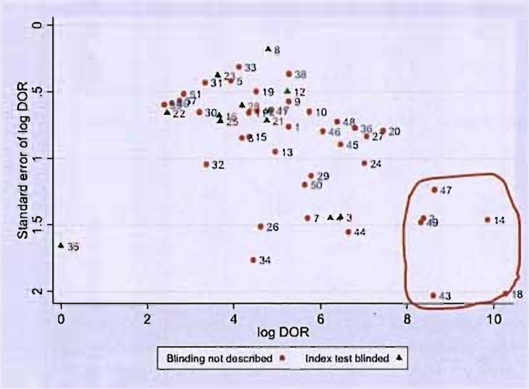
	Effect on accuracy (RDOR)[a]			Effect on other parameters[b]			
Method	**RDOR at Q***	**RDOR at ref threshold**	**RDOR at comp threshold**	**threshold**	**shape**	**sensitivity**	**specificity**
No shape interaction (parallel curves)							
Moses method - equal weight	0.21, *P*=0.02	as at Q*	as at Q*	-	-	*-0.21*	*-0.01*
Moses method - weighted	0.59, *P*=0.14	as at Q*	as at Q*	-	-	*-0.11*	*0.01*
HSROC	0.25, P=0.02	as at Q*	as at Q*	-0.47, P=0.19	-	*-0.17, P=0.02*	*-0.01, P=0.61*
BVN	*0.25, P=0.02*	as at Q*	as at Q*	*-0.46, P=0.19*	-	-0.17, *P*=0.02	-0.01, *P*=0.61
With shape interaction (crossing curves)							
Moses method - equal weight	0.15, P=0.07	0.19, P=0.02	0.21, P=0.02	-	-0.12, P=0.68	*-0.21*	*-0.01*
Moses method - weighted	0.74, P=0.63	0.62, P=0.19	0.57, P=0.13	-	0.09, P=0.66	*-0.11*	*0.01*
HSROC (shape)	0.21, *P*=0.03	0.47, P=0.08	0.26, P=0.13	-0.25, *P*=0.65	*0.22, P*=0.63	*-0.17, P=0.11*	

Figures in *italics* denote derived values, i.e. parameters which are not the natural output from the model in question but have been transformed from model parameters

[a] – RDOR – relative diagnostic odds ratio or difference between the curves, at Q* (the point where sensitivity=specificity), at the average threshold of the reference group (ref threshold) or of the comparator group (comp threshold). Studies reporting blinded index interpretation form the comparator group (numerator) and studies where blinding is not reported the reference case (denominator)

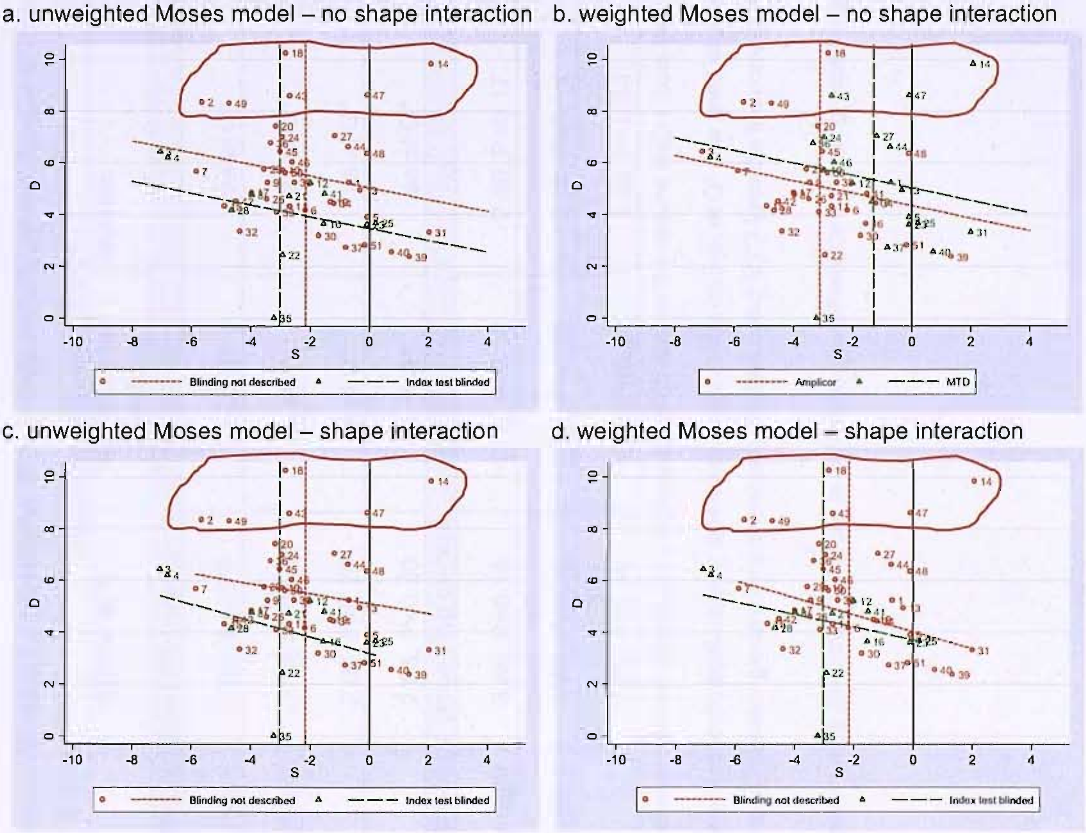
[b] – the effect on parameters other than accuracy is defined as the difference between groups in each parameter and the P-value for the difference

Figure 18 Plot of SE(log DOR) against log DOR – index test blinding[9]



The effect on the pooled analysis can be seen by looking at the same studies on the D vs S plots (circled in Figure 19). Whether the regression lines are parallel or crossing, the lines are much closer together when the weighting is applied because the six studies get very little weight due to their high SEs. The DORs for the two groups are therefore very similar under the weighted model (Figure 19b and d).

Figure 19 D vs S plots – by index test blinding


a. unweighted Moses model – no shape interaction b. weighted Moses model – no shape interaction

c. unweighted Moses model – shape interaction d. weighted Moses model – shape interaction

the various average threshold points are denoted where the regression lines cross the vertical lines at S=0 (Q), S= -2.16 (ref group average threshold), S= -3.03 (comparator group average threshold).

generated.

[9] See section 4.1.1. for a description of this comparison

Table 22: Difference in model parameters: MTD (comparator) versus Amplicor (reference)

	Effect on accuracy (RDOR) [a]			Effect on other parameters [b]			
Method	**RDOR at Q***	**RDOR at ref threshold**	**Method**	**RDOR at Q***	**RDOR at ref threshold**	**Method**	**RDOR at Q***
No shape interaction (parallel curves)							
Moses method - equal weight	1.99, *P*=0.28	as at Q*	as at Q*	-	-	*+0.15*	*-0.02*
Moses method - weighted	2.16, *P*=0.08	as at Q*	as at Q*	-	-	*+0.17*	*-0.03*
HSROC	2.06, *P*=0.20	as at Q*	as at Q*	1.06, *p*<0.01	-	*0.17, p<0.01*	*-0.02, P=0.06*
BVN	2.05, P=0.20	as at Q*	as at Q*	*1.06, p<0.01*	-	0.17, *p*<0.01	-0.02, *P*=0.06
With shape interaction (crossing curves)							
Moses method - equal weight	3.86, P=0.14	1.48, P=0.58	2.59, P=0.17	-	0.31, P=0.31	*+0.15*	*-0.02*
Moses method - weighted	1.63, P=0.38	2.82, P=0.06	2.04, P=0.11	-	-0.18, P=0.43	*+0.17*	*-0.03*
HSROC	2.29, *P*=0.16	1.59, P=0.24	1.64, P=0.14	0.81, *P*=0.12	-0.23, *P*=0.56	*0.17, P=0.01*	*-0.02, P=0.57*

Figures in *italics* denote derived values, i.e. parameters which are not the natural output from the model in question but have been transformed from model parameters

[a] – RDOR – relative diagnostic odds ratio or difference between the curves, at Q* (the point where sensitivity=specificity), at the average threshold of the reference group (ref threshold) or of the comparator group (comp threshold). Studies using MTD form the the comparator group (numerator) and studies of Amplicor the reference case (denominator)

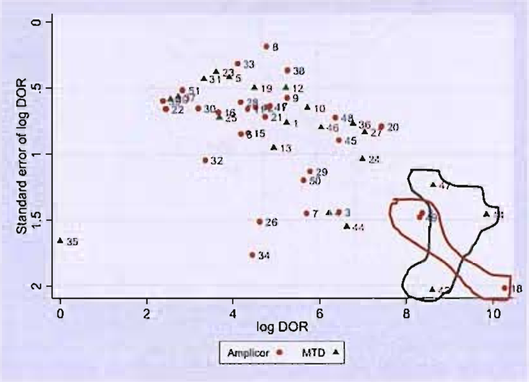
[b] – the effect on parameters other than accuracy is defined as the difference between groups in each parameter and the P-value for the difference

4.3.1.2 Moses models comparison: Test type[h]

When the test covariate is added to the Moses models, the weighted and unweighted results are very similar, as long as no interaction with shape is allowed (RDORs 2.16 and 1.99 respectively, Table 22). Of the six studies with biased SEs, three are of MTD and three of Amplicor (Figure 20). Where the SROC curves are assumed to be parallel, the effect is spread across the two groups, and the difference between groups remains similar whether weighting is applied or not (Figure 21a and b).

Where an interaction of test type with curve shape is allowed however, the weighted and unweighted models no longer give similar results (Table 22). Not only does one model give a larger RDOR than the other, but the RDOR and the model giving the largest RDOR varies according to the point on the curves at which the RDOR is

Figure 20 Plot of SE(log DOR) against log DOR – test type



Where RDOR is estimated at Q*, the unweighted model finds MTD to be much more accurate than Amplicor compared to the difference shown by the weighted model. At the average threshold for the reference group, it is the weighted model that shows MTD to be a considerably more accurate test (p<0.10), while the unweighted model shows a much smaller difference between tests. At the average comparator threshold, the two models find a similar difference between groups.

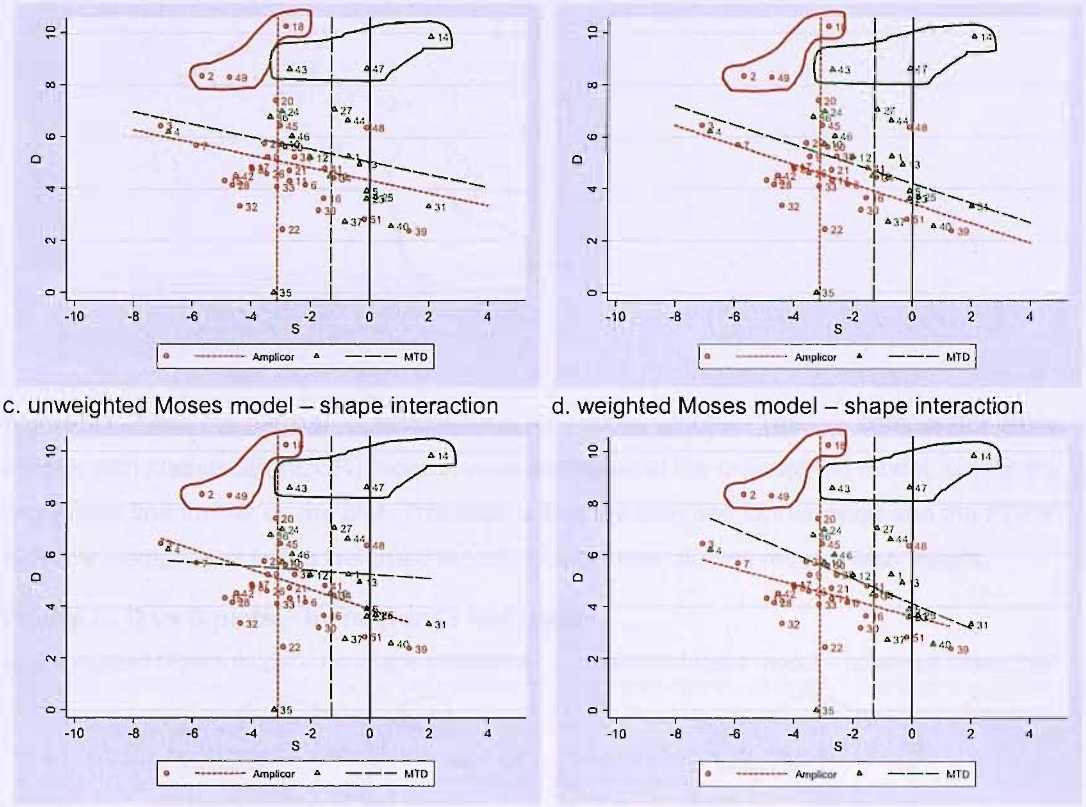
The reason for these differences can be seen from the D vs S plots in Figure 21c and d. The regression line for the reference case (Amplicor) remains similar in position and slope whether weighting is used or not, however the slope of the regression line for the comparator case (MTD studies) has a considerably steeper slope under the weighted Moses model.

These differences in slopes explain why there is a bigger difference in RDOR between the unweighted and weighted models compared to the case where the slopes are assumed to be the same (Figure 21a and b). In particular, study 14 (Chedore) has a high vale for D and a

[h] See section 4.1.1. for a description of this comparison

high value for S. Where S is allowed to vary, this study has a considerable effect on the slope of the regression line and therefore on the point at which the lines intercept zero (Q* point).

Figure 21 D vs S plots – by test type



a. unweighted Moses model – no shape interaction b. weighted Moses model – no shape interaction

c. unweighted Moses model – shape interaction d. weighted Moses model – shape interaction

the various average threshold points are denoted where the regression lines cross the vertical lines at S=0 (Q), S= -3.11 (ref group operating point), S= -1.29 (comparator group operating point).

The changing slopes also explains the differences between models according to where RDOR is estimated. For the unweighted model the regression lines cross at the left hand side of the plot whereas for the weighted model they would cross at the right hand side of the plot if the lines were extrapolated slightly further. At Q* (where S=0), the lines are much further apart under the unweighted model compared to the weighted. At the mean of S for the reference group (S=-3.11), the opposite is the case, whilst at the mean of S for the comparator group the lines are similar distances from each other whether weighted or unweighted. This example demonstrates that even where the difference in shape between groups is not statistically significant, allowing a difference in shape between groups can lead to big differences in RDOR according to where RDOR is estimated.

4.3.1.3 Moses models comparison: Reference test used[i]

At the simplest level (with no interaction of covariate with shape), when the type of reference test used is added to the Moses models, the RDOR for the unweighted model is 2.48 compared to just 1.12 for the weighted model (Table 23). Only one of the six studies with

[i] See section 4.1.1. for a description of this comparison

biased SE(lnDOR) is in the reference case group (combined reference test) for this example (Figure 22).

Figure 22 Plot of SE(log DOR) against log DOR – reference test used

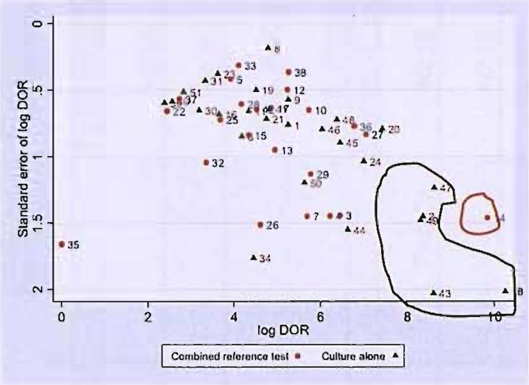
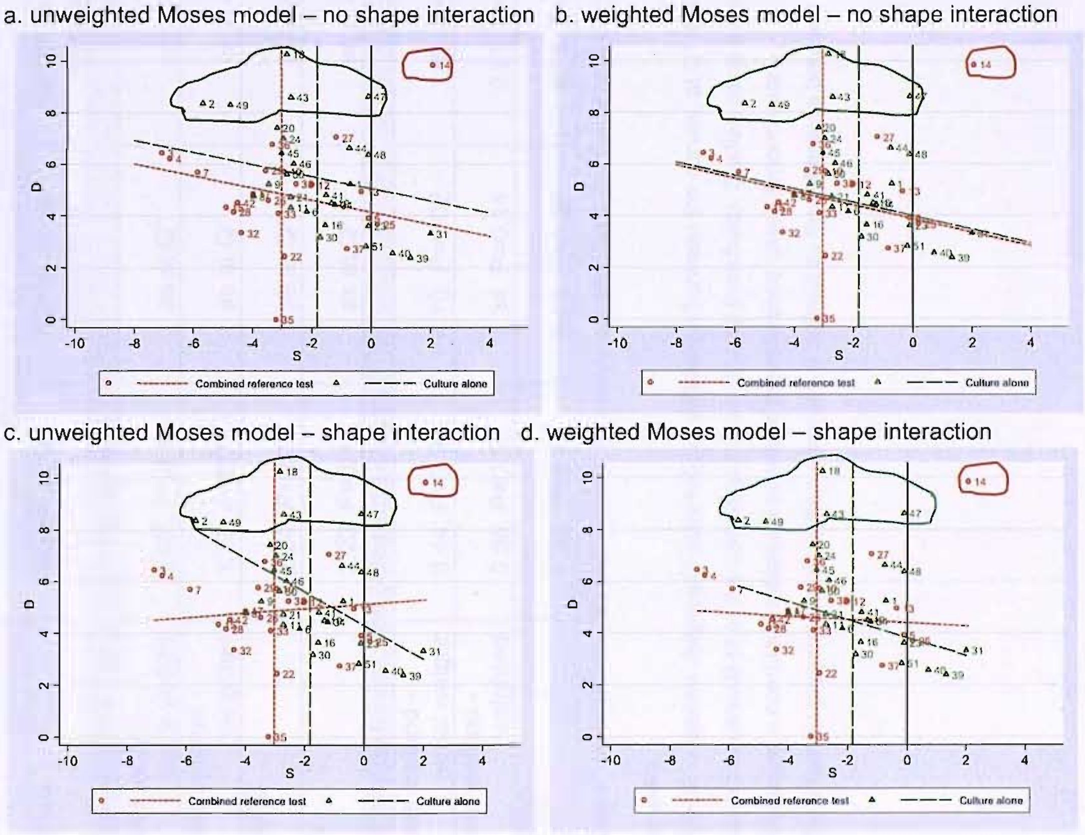


Figure 23 shows the parallel regression lines for the two groups. The five comparator group studies with biased SE(lnDOR) receive more emphasis in the unweighted model, pulling the regression line further up the plot. This means that the lines are further apart and the RDOR higher in comparison to the weighted model, where these studies receive less weight.

Figure 23 D vs S plots – by reference test used


a. unweighted Moses model – no shape interaction b. weighted Moses model – no shape interaction

c. unweighted Moses model – shape interaction d. weighted Moses model – shape interaction

the various average threshold points are denoted where the regression lines cross the vertical lines at S=0 (Q), S= -3.02 (ref group operating point), S= -1.82 (comparator group operating point).

Table 23: Difference in model parameters: culture alone (comparator) versus combined reference test (reference)

	Effect on accuracy (RDOR) ^{a}			**Effect on other parameters**^{b}			
Method	**RDOR at Q***	**RDOR at ref threshold**	**RDOR at comp threshold**	**threshold**	**shape**	**sensitivity**	**specificity**
No shape interaction (parallel curves)							
Moses method - equal weight	2.48^{c}, *P=0.12*	as at Q*	as at Q*	-	-	*+0.15*	*-0.01*
Moses method - weighted	1.12, *P=0.73*	as at Q*	as at Q*	-	-	*+0.10*	*-0.02*
HSROC	2.24, *P=0.14*	as at Q*	as at Q*	0.73, *P*=0.02	-	*0.15, P=0.01*	*-0.01, P=0.40*
BVN	*2.23, P=0.14*	as at Q*	as at Q*	*0.72,P=0.02*	-	0.15, *P*=0.01	-0.01, *P*=0.40
With shape interaction (crossing curves)							
Moses method - equal weight	0.44, P=0.33	4.10, P=0.02	1.69, P=0.35	-	-0.74, P=0.01	*+0.15*	*-0.01*
Moses method - weighted	0.56, P=0.32	1.34, P=0.14	0.95, P=0.88	-	-0.29, P=0.14	*+0.10*	*-0.02*
HSROC	1.22, *P=0.76*	0.76, P=0.32	8.75, P=0.34	1.74, *p*<0.01	0.90, *P*=0.02	*0.14, p<0.01*	*-0.01, P=0.05*

Figures in *italics* denote derived values, i.e. parameters which are not the natural output from the model in question but have been transformed from model parameters

[a] – RDOR – relative diagnostic odds ratio or difference between the curves, at Q* (the point where sensitivity=specificity), at the average threshold of the reference group (ref threshold) or of the comparator group (comp threshold). Studies using culture alone as the reference standard form the comparator group (numerator) and studies using a combined reference standard the reference case (denominator)

[b] – the effect on parameters other than accuracy is defined as the difference between groups in each parameter and the P-value for the difference

Where an interaction of covariate with shape is allowed, however, a more complex picture emerges (Figure 23c and d). If RDOR is estimated at the reference or comparator group average threshold points, the unweighted model suggests that studies using culture alone have higher accuracy than those using a combined reference test. The difference between models is much greater at the average reference threshold with the RDOR from the unweighted model over 3 times that of the weighted model (4.10 compared to 1.34, Table 23).

If RDOR is estimated at the Q* point however, studies using culture alone are shown to be considerably *less* accurate than those using a combined reference test, and the difference between the weighted and unweighted models is much less (RDOR 0.44 for the unweighted and 0.56 for the weighted models respectively, Table 23). The reason for this complex picture is hinted at by the highly statistically significant shape term ($p<0.01$) for the unweighted model.

Figure 23c and d clearly demonstrate the effect of the shape term. Under both models the regression lines for the two groups cross near to the centre of the data, and furthermore the slope of the regression lines changes considerably, with that for the reference group (combined reference test) even changing direction between models. The five studies in the culture alone group with biased SE(lnDOR) receive more emphasis in the unweighted model, pulling the left hand side of the regression line further up the plot. Due to the high SEs, these studies receive less weight when the weighting is applied and the slope is much less steep. The overall effect is that at the average reference threshold, the lines are much further apart under the unweighted model.

Because the regression lines cross near to the centre of the data, the point at which RDOR is estimated has a massive impact on its size and direction. At Q* (S=0), the combined reference test line intercept is above that for culture alone under both models, so that the RDOR favours the combined reference group. At the average threshold for the reference group (S= -3.02), however, the regression line for the culture alone group is above that for the combined reference group, favouring the culture alone group.

4.3.1.4 Summary

The presence of studies with biased SE(lnDOR) can potentially have a huge effect on any heterogeneity investigations under the Moses model framework. Because of the weighting by the inverse of SE(lnDOR), studies with biased SE's get considerably less emphasis under the weighted model.

At its simplest, for example where all studies with biased SEs fall into the same subgroup such as with the presence of index test blinding, the effect manifests predominantly in an underestimation of RDOR for the weighted model compared to unweighted. However, where the biased SE studies fall into different groups the effect is more complex. The biased SE studies were evenly distributed according to test type. The effect on RDOR alone was less as both groups were affected – the overall location of the regression lines on the plot does not change much whether or not weighting is used. However a much larger impact on curve shape (or the slope of the regression line), was found, especially for the group containing the outlying study (number 14). This in turn impacts on RDOR as the distance between the regression lines increases or decreases, depending on the change in slope and the location at which RDOR is estimated (Q* or average threshold point).

The final example illustrates this last effect as the outlying study 14 was the only one in the combined reference test group but this was sufficient to change the slope of the regression line between unweighted and weighted models and contributed considerably to the differences in RDOR according to the point at which it is estimated.

4.3.2 Comparing the advanced model results

The results in the top halves of Table 21 to Table 23 show that for this dataset, where no interaction of covariate with shape is allowed (parallel curve models), the HSROC and bivariate normal model give near identical results to within two decimals places. The main comparison is therefore of the HSROC approach with and without the interaction of covariate with shape, i.e. between the parallel and crossing curve HSROC models. The SROC curves for the HSROC models with and without a shape interaction are plotted in Appendix 15 to Appendix 17.

4.3.2.1 HSROC parallel versus crossing curve models: Index test blinding

The top half of Table 21 shows that by assuming parallel SROC curves, the presence of index test blinding has a large and strong effect on accuracy (RDOR 0.25, P=0.02), suggesting that the two groups are operating on two different SROC curves. There is also some suggestion of differences in threshold between groups (P=0.19). Appendix 15, Figure c shows that the SROC curve for the index test blinded group is considerably below that for the blinding not reported group. Where an interaction of covariate with shape is allowed (bottom half of Table 21), there is no suggestion of differences in shape or threshold between groups (P=0.63 and P=0.65 respectively). The RDOR does vary along the curves but the strong evidence of differences in accuracy generally remains.

The choice of model has a relatively small effect on the conclusions that would be drawn from this dataset.

4.3.2.2 HSROC parallel versus crossing curve models: Test type

Where test is included as a covariate and SROC curves are assumed to be parallel (Table 22), the advanced method models show some evidence of differences in accuracy between groups (RDOR 2.05, P=0.20), but the main difference is in threshold (p<0.01), suggesting the two groups may operate on the same SROC curve but at different thresholds.

With the interaction of test type with curve shape, there is no evidence for differences in shape between groups (P=0.56) but the evidence for a difference in threshold remains although it is less strong (P=0.12) (Table 22). There is now slightly more evidence of a difference in accuracy, depending on where RDOR is estimated. At Q*, the RDOR is similar to that for the parallel curve model (2.29, P=0.16), at the average reference and comparator thresholds, the first of which in particular is considerably closer to the centre of the data than Q*, the RDOR is slightly lower at 1.59 (P=0.25) and 1.64 (P=0.14) respectively.

The inclusion of the interaction of covariate with shape for this example slightly affects the strength of conclusions that would be drawn regarding the difference between tests.

4.3.2.3 HSROC parallel versus crossing curve models: Reference test used

A more complex picture emerges according to reference test used. For the parallel curve models, the advanced method models show reasonable evidence of differences in accuracy between groups (RDOR 2.23, P=0.14), but the strongest evidence is for a difference in threshold (P=0.02).

The bottom half of Table 23 however, shows that where the curves are allowed to have different shapes, there is no real evidence for differences in accuracy regardless of the point at which RDOR is estimated, but strong evidence for differences in shape and threshold (P=0.02 and p<0.01, respectively). Although the differences in accuracy are not statistically significant regardless of where RDOR is estimated, the direction and magnitude of the RDOR varies considerably. At Q*, the RDOR is 1.22, slightly in favour of the studies using culture alone having higher accuracy, at the comparator group average threshold point it is 8.75 and at the average reference threshold it is 0.76 in favour of the combined reference test group having higher accuracy.

Appendix 17, Figure b helps with the interpretation of this data. Where the curves are allowed to have different shapes, they cross very near to the sensitivity=specificity line, or Q* point. This explains why there is so much variation in the RDOR. The model chosen for this example has a large effect on the results of the analysis.

4.3.2.4 Summary

At the simplest level, where a covariate predominantly affects accuracy alone, allowing for an interaction of that covariate with shape can have some effect on the size and statistical significance of differences between groups. For the example here, index test blinding, the overall conclusion drawn regarding the potential effect from the covariate would not be greatly affected although the strength of that conclusion would be somewhat affected by the model chosen (with or without a shape interaction term) and the point at which RDOR was estimated.

For the test type covariate example above, where the strongest evidence was for differences in threshold between groups, choosing a model which allows the SROC curves to have different shapes led to slightly more evidence for differences in accuracy depending on the point at which RDOR was estimated. However the RDORs were not very different from that given when no interaction with shape is allowed and conclusions regarding the importance of test to the analysis would not be greatly affected by the model chosen.

In the most complex example, where type of reference test used is added to the analysis, the model chosen and point at which RDOR is estimated has a big impact on the results. If parallel SROC curves are modeled, the DORs per subgroup are 191 for culture alone, and 85 for studies using a combined reference test. The differences in the parameters suggests that it is threshold rather than accuracy that varies the between groups, i.e. the studies are operating at different points on the same or similar SROC curves.

Where ROC curves have different shapes however, a more confusing picture emerges; with strong evidence for differences in shape and threshold but not accuracy, with large variations in RDOR along the curves. It is very difficult to untangle the different effects for this example, however it is clear that the choice of model has a considerable impact on the conclusion drawn. With parallel curves, one might infer probable differences in accuracy between groups (different SROC curves) and that the two groups operate at different thresholds. With differently shaped curves, one might say that the studies do operate on different SROC curves but the differences in shape and the fact that the studies operate at different points on the curves means that there is no associated difference in accuracy between the groups.

4.3.3 Moses model results versus advanced model results

The Moses and advanced model results can be compared both with and without an interaction of covariate with shape. Again the results of the analyses adding each covariate can be seen in Table 21 to Table 23. Parallel and crossing SROC curves for each model can be compared in Appendix 15 to Appendix 17.

4.3.3.1 Moses versus HSROC models: Parallel SROC curves

For each of the three covariates examined here the unweighted Moses model results most closely resemble the results of the advanced models, in terms of size, direction and strength of evidence from the RDOR (Table 24). For one covariate (test type) the RDOR for the weighted Moses model is similar to that from the HSROC and BVN models with an RDOR of 2.16 compared to 2.06 (Table 22). The associated P-value, however, would suggest a strong difference in accuracy between tests (P=0.08) whereas the advanced models and unweighted Moses model suggest that the effect is not as strong (P=0.20 and P=0.28). The other two examples both show the weighted model under-estimates the RDOR compared to the advanced models (Table 21 and Table 23).

Table 24 Summary of similarity of strength of evidence from HSROC model and Moses model results

	Blinded index test	Test type	Reference test used
PARALLEL SROCs			
RDOR	eq	~	eq
CROSSING SROCs			
RDOR at Q*	eq	eq	neither[a]
RDOR at ref threshold[b]	eq	w	neither[a]
RDOR at comp threshold[b]	w	both	neither[a]
Shape	both	w	eq

[a] neither P-values nor magnitude or direction of RDORs are similar

[b] RDOR, or difference between the curves, at the average reference threshold (ref) or comparator threshold (comp)

eq – unweighted (equal weight) Moses; w – weighted Moses model

4.3.3.2 Moses versus HSROC models: Crossing SROC curves

Where the SROC curves can have different shapes, neither Moses model consistently approximates the results of the HSROC model (Table 24).

For blinded index test interpretation the unweighted model results are very close to the results of the HSROC model except for the RDOR at the average reference threshold point which is over-estimated (0.19 compared to 0.47 for the HSROC model). The weighted Moses model is closest to the HSROC model at this same point, but underestimates the RDOR in comparison to the HSROC model at the other two points. All three models show no strong evidence of differences in shape.

When the test type covariate is added to the models (Table 22), the RDORs for both Moses models are in the same direction as for the HSROC but both either over- or under-estimate their magnitude. On the whole, the strength of evidence for differences in accuracy (P-values for the RDORs) of the unweighted model are most similar to the HSROC. All three models show no evidence of differences in shape.

For the analyses by type of reference test used, neither Moses model comes near to estimating the HSROC model results for differences in accuracy. The RDORs for both models are in the opposite direction to the RDORs from the HSROC model at almost every point (Table 23). The HSROC model suggests no differences in accuracy despite the varying magnitudes of RDOR along the SROC curves. Both Moses models however suggest some evidence of differences in accuracy at the average reference threshold, (P=0.02 for the unweighted model and P=0.14 for the weighted model).

All three models indicate differences in shape between groups, although the evidence is less strong for the weighted Moses model (P=0.14).

4.3.3.3 Summary

For the three covariates examined here, where parallel SROC curves are modeled, the unweighted Moses model results are very close to the results of the advanced models in terms of size, direction and statistical significance of the RDOR and difference in sensitivity and specificity between groups. The weighted Moses model approximates the results of the advanced model for only one covariate (test type). For the other two examples it considerably under-estimates the difference in accuracy (RDOR) and in sensitivity.

Where the SROC curves are allowed to have different shapes, neither model consistently approximates the HSROC model results. For blinded index test interpretation, the equal weighted model is very close to the HSROC model with the exception of the RDOR at the average reference threshold. For the analysis by test type, it is the weighted model results that generally most similar to the HSROC model. For this covariate, the unweighted model over-estimates the effects seen for the HSROC model, again except at the RDOR at the average reference threshold. For reference test used, neither Moses model gives RDORs anywhere near those of the HSROC model and in fact in most cases have results in the opposite direction.

4.4 Discussion

The HSROC and BVN, or advanced, methods have several theoretical advantages over the Moses method, making their results more statistically reliable and accurate. This chapter examined similarities and differences in results between models, and explored the effect of adding covariates. Part of the aim was to identify any suggestion that either Moses method could approximate the results of the more statistically rigorous methods. The covariates examined were specifically chosen to illustrate a range of effects on the different model parameters and the potential differences between models. The effects are not necessarily typical of the effects that would be expected in most systematic reviews.

Reflection on the analyses carried out here shows that for the overall pooled analysis for this dataset:

- there is considerable disagreement between the two Moses models,
- the two advanced models gave almost identical results,
- the unweighted Moses model results were most similar to those of the advanced methods. The weighted Moses model considerably under-estimated the results of the other two models.

With the addition of covariates to the models:

- there was common and sometimes considerable disagreement between the two Moses models regardless of whether parallel or crossing SROC curves were modeled,

- where parallel SROC curves were modeled, the two advanced models give near identical results
- for the advanced models, in some circumstances the interaction of covariate with shape made little difference to the conclusions that would be drawn from the model regarding the importance of a covariate, but in others conflicting results arose
- where parallel SROC curves are modeled, the unweighted Moses model generally has results more similar to the advanced methods than the weighted Moses model
- where curves are allowed to have different shapes, neither Moses model consistently approximated the HSROC model results

The disagreement between the two Moses models, both for the overall analysis and the investigation of heterogeneity was primarily due to bias in the SE(lnDOR), whose inverse was used as the weight for the weighted model. For some studies in this dataset the SEs are biased upwards, so that they have higher SEs than might be expected from their sample sizes. Weighting by the inverse of the SE mean that these studies received a very low emphasis in the weighted Moses analysis, leading to overall under-estimation of effects in comparison to the unweighted analysis.

The circumstances under which biased SE(lnDOR) might be expected are as follows: extreme values of sensitivity and specificity, often with zero FNs or FPs, unequal sample sizes of diseased and nondiseased patients, and variation in the threshold for test positivity leading to variation in the proportion of patients who are test positive. These circumstances are common in diagnostic meta-analysis, therefore bias in the SE is always a risk.

Studies with extreme values in sensitivity and/or specificity, along with studies for which sensitivity estimates were greater than specificity or were similar in magnitude to specificity, also had the biggest individual effect on the unweighted Moses and HSROC models. This was because these studies lie around the edges of the dataset; studies with more extreme values having the greatest effect on an analysis.

The group to which these studies were allocated according to covariate in turn impacted on the difference in model parameters between groups and the complexity of the differences between models. In particular, the relationship of the study with the highest sensitivity and very high specificity (study 14 by Chedore and colleagues,[152] located at the far top left of the ROC plot) to other studies in the same subgroup seemed to particularly influence the importance of the shape term. For example, for the investigations by index test blinding and test type, this study was surrounded by others in the same subgroup, and the main effects of the covariate were on accuracy and/or threshold differences. For the reference test used covariate, the other studies in the same subgroup as Chedore all had high specificity but generally much lower sensitivity so that the Chedore study had the main influence on the shape of the SROC curve, pulling it away from that for the other subgroup. The Chedore study was one of a small group of studies where sensitivity was greater than specificity, and its exceptionally high DOR in relation to the others explains its strong influence on this review.

A further finding from the analyses presented here is related to the misleading nature of the estimation of accuracy and differences in accuracy at Q*. The primary analyses showed a considerable difference in DOR according to where it is estimated and potentially more importantly huge differences in relative DORs when comparisons by covariate are made, to the extent that the direction of effect can change according to where DOR is estimated. This could lead to highly misleading conclusions. For the primary analyses, estimation of DOR near to the centre of the data (e.g. using the mean threshold value as was done here), would seem to give a more representative picture of the data. For the investigation of heterogeneity, however, the choice of point at which to estimate RDOR is more complex, especially if the operating points of the subsets of data are not in close proximity to each other and furthermore, if the SROC curves cross near to the centre of the data.

This and the other findings discussed above require further exploration in other datasets to try to identify how commonly each occurs and ultimately to make some recommendations as to whether for the advanced models, an interaction of covariate with shape aids review interpretation or simply “over-models” the data and whether, in general, either Moses model provides a better approximation to the advanced model results or not. If supported further, the under-estimation of effects from the weighted Moses model in comparison to the unweighted model and the under-estimation of effects from both models compared to the advanced models will have considerable implications for systematic reviews, both past and future.

5 A re-analysis of previously published systematic reviews to identify spectrum effects

This chapter reports an empirical study replicating the methods used in the previous chapter on data obtained from a large sample of previously published systematic reviews of diagnostic tests. The aim is to determine the extent to which the findings in Chapter 4 can be generalised, i.e. to examine the extent to which the meta-analytic models disagree and under what circumstances and to compare the identification of spectrum effects.

5.1 Methods

The methods followed were similar to those presented in Chapter 3 and Chapter 4 with the following differences and additions.

5.1.1 Literature search

The Centre for Reviews and Dissemination's Database of Abstracts of Reviews of Effects (DARE) was again used to identify existing systematic reviews of diagnostic studies. Diagnostic reviews indexed on DARE up to December 2002 were screened to identify diagnostic reviews for inclusion in Chapter 3. This search was updated in July 2005 in order to identify more recently published reviews.

5.1.2 Eligibility criteria

Diagnostic systematic reviews comparing a test to a reference test were included if they presented:

1. sufficient information to allow the construction of a 2x2 contingency table for each primary study. This information was used to calculate relevant accuracy statistics. Studies reporting only summary accuracy statistics without sufficient raw data to allow the construction of a 2x2 table were excluded.
2. information on at least one spectrum-related covariate for each primary study

Studies were assessed for inclusion by one reviewer. Screening was undertaken in two stages, initially the full sample of identified reviews was screened for reviews meeting criterion 1. It was estimated that 30 to 40 reviews would be sufficient for this empirical study and that more recent reviews might be more likely to publish spectrum-related data, therefore the second stage was to screen reviews published between 2000 and 2005 for reviews meeting criterion 2.

5.1.3 Data extraction

A brief data extraction form for recording relevant information from each systematic review was designed and piloted (Appendix 18). Data were extracted on:

- primary study author and year of publication

- the experimental test and target disorder
- 2x2 contingency table data
- data on any potential spectrum-related sources of heterogeneity per study.

The full systematic reviews were data extracted independently by two reviewers. Any disagreements were resolved by consensus or by referral to a third reviewer if necessary.

5.1.4 Data synthesis

Data were synthesised using three meta-analytic models: the Moses model (both unweighted and weighted by inverse variance of lnDOR) and the HSROC model. The BVN model was not applied because

a. the analyses in Chapter 4 showed that it produces results virtually identical to those of the HSROC model
b. the BVN cannot easily allow for an interaction of covariate with shape so that it could only be compared to the other models where parallel SROC curves were assumed.

The meta-analytic methods were undertaken as for the TB dataset in Chapter 4 and described in detail in Appendix 10.

In summary, each model estimates mean accuracy (DOR) with 95% confidence intervals and an estimate of asymmetry in the SROC curve (P-value associated with the shape term). The HSROC model also produces an estimate of mean threshold and its 95% confidence intervals. The models naturally estimate DOR at Q* (the point where sensitivity=specificity), however this point is often nowhere near the centre of the data and therefore not representative. The DORs were therefore also estimated at the average threshold, i.e. near to the average operating point of the dataset.

As before, covariates were added to the models in two ways. First, assuming that the SROC curves for the two groups are parallel; second, allowing the covariate to interact with curve shape (i.e. the SROC curves will cross at some point along their length).

Differences between groups according to covariates can be assessed by:

1. differences in accuracy or the relative diagnostic odds ratio (RDOR). Both parallel and crossing curve models naturally produce the RDOR at Q*. This value is constant along the length of the parallel curve models, but varies along the length of two crossing curves. This can be seen visually by the variation in distance between the curves. For the crossing curve model, the RDOR at Q* does not necessarily adequately represent the data, especially if the curves cross near to the centre of the data. RDOR has therefore been estimated at the average threshold for the reference group and at the average threshold for the comparator group. This gives RDORs near to the average operating points of each subgroup. RDORs have been estimated at all three points to examine

whether the differences between models vary according to the point at which RDOR is estimated and to give a picture of how RDOR varies along the curves.

2. differences in threshold – for the HSROC method only
3. differences in curve shape - for the crossing curve models only

The difference in sensitivity and specificity between groups can also be estimated however, unless one can control for differences in threshold between studies, a comparison of operating points is not useful.

5.1.5 Comparison of meta-analytic methods

As in Chapter 4, three main model comparisons were undertaken:

A. unweighted Moses versus HSROC
B. weighted Moses versus HSROC
C. weighted Moses versus unweighted Moses (primary analyses only)

Comparisons A and B were to determine whether either Moses method produces results akin to the HSROC method (the benchmark). Comparison C was undertaken to identify whether the weighted Moses method consistently underestimates the unweighted method.

Primary analyses of complete datasets

Comparison of DORs

For the overall pooled analyses, the similarity of the DOR estimates were compared by estimating the ratio of DORs (denoted RORs) between models. These were estimated for the DORs at Q* (the point where DOR is often estimated in reviews) and at the average threshold (a point nearer to the centre of the data). RORs were estimated at both points to see if the model results were more less similar at these points.

A summary of the RORs per comparison was provided using box and whisker plots. These give a simple graphical summary, showing the central location of the data (the median), two measures of dispersion (the range and inter-quartile range), the skewness (from the orientation of the median relative to the quartiles) and an indication of any potential outliers. The median ROR tells us what, on average, the bias is between one model and another; if the median ROR is 1, there is on average no bias between the models. The IQR (denoted by the 'box') demonstrates the extent to which individual reviews agree or disagree with the median result. If the disagreements are all very small, the box will be quite tight around the median; if it is possible that reviews have large disagreements then the box and whisker will extend a considerable distance.

Comparison of SROC shape

The extent to which the different models show similar evidence of asymmetry of SROC curves was summarised using a 'P-value plot'. This is a scatter plot of pairwise comparisons of the P-values for the shape terms for each meta-analysis, thereby comparing how two different models measure the strength of evidence. The closer the scatter points are to the central diagonal line, the more similar the P-values are between models. The shape parameter in the Moses model is not directly comparable with the shape parameter in the HSROC model so they cannot be directly compared.

<u>Stratification of the comparison of DORs</u>

The examination of the overall pooled analyses for the ROR at the average threshold were also stratified by the following characteristics:

a. size of DOR, using the pooled estimate from the HSROC model as the basis for the stratification. It was assumed that reviews with very high overall pooled DORs included studies with high DORs and therefore with exceptionally high sensitivities and/or specificities.
b. range in 'S' per review from the Moses model, to reflect variation in the threshold for test positivity. At extremes of 'S', the less equal the numbers of true negatives and false positives.
c. number of zero FP or FN cells in the included studies. The higher the relative number of zero cells per review the more biased the SE(lnDOR) and the bigger the differences between the unweighted and weighted Moses models. Furthermore, where there are lots of zeros, the Moses models will have added 0.5 as a correction in their method, which will lead to downward bias in the estimate of the odds ratio in comparison to the HSROC model.
d. strength of evidence for asymmetry as estimated from the HSROC model (P-value associated with shape term)
e. strength of evidence for the importance of differences in threshold as estimated from the HSROC model (P-value associated with threshold term)

One would expect differences between the unweighted and weighted Moses models according to characteristics a., b. and c. Stratification by characteristics c., d. and e. might help illuminate any circumstances under which the Moses methods can approximate the results of the HSROC model. The stratification of the Moses comparisons by the presence of asymmetry or threshold effects for the HSROC model would not be useful, therefore only the comparison with the HSROC model were stratified by characteristics d. and e.

Comparison of heterogeneity investigations

<u>Comparison of RDORs</u>

As discussed above, for each investigation of a covariate in a review, an RDOR is estimated. To compare between models therefore, a ratio of RDORs was estimated (denoted RROR). Between model comparisons were made between each Moses model and the HSROC model for the:

- RDORs at Q* (for parallel and crossing curve models)
- RDORs at the average reference threshold (crossing curve models)
- RDORs at the average comparator threshold (crossing curve models)

A further within model comparison of RDORs between the parallel and crossing curve versions of each model was also undertaken for the RDORs at Q*.

As the RDOR for a comparison of subgroups can sometimes be less than one and sometimes greater than one, the summary statistics were standardised by coding to ensure that the HSROC model always estimates an RDOR greater than one. For the comparisons where the RDOR for the HSROC model was less than one, the inverse of the RDORs from all three models was taken to ensure standardisation of direction. For the within model comparisons, the summary statistics were coded to ensure that the crossing curve version of each model always estimates an RDOR greater than one.

A summary of the ratio of RDORs (RRORs) estimated at each point per model was again provided using box and whisker plots. P-value-plots were used to display the pairwise comparisons of P-values for each RDOR comparison.

Comparisons of differences in shape

P-value plots were used to display the pairwise comparisons of P-values for the differences in the shape term between models.

Comparisons of differences in threshold

P-value plots were also used to display the pairwise comparisons of P-values for differences in threshold for the HSROC model with parallel versus crossing SROC curves.

5.2 Results

5.2.1 Summary of reviews identified

Of 331 identified reviews, 153 presented sufficient data to complete 2x2 contingency tables per study. Of these, 97 were published between 2000 and 2005. On further examination, 29 presented detailed information on at least one spectrum-related covariate per study. The 29 reviews provided covariate information for 60 spectrum-related investigations of heterogeneity (Table 25). Figure 24 provides a flowchart of the review selection process. Details of the reviews and results of the primary analyses are provided in Appendix 19. Details of the heterogeneity investigations are given in Appendix 20 to Appendix 22.

The median number of studies per review was 17 (IQR 12, 26). The median sample sizes ranged from 20[178] to 7575.[179] the most commonly investigated tests were imaging tests (13 of 29 reviews) followed by clinical assessment or examination (5 reviews). The most commonly

Figure 24 Flowchart of the review selection process

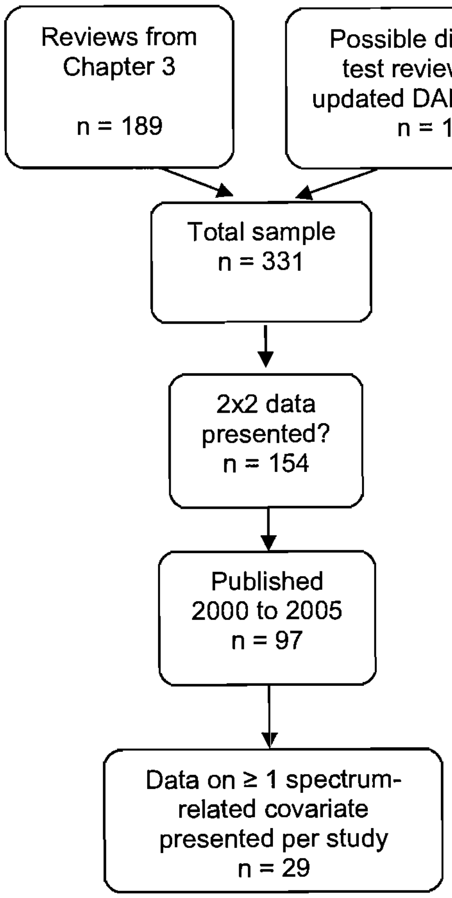

Reviews from Chapter 3
n = 189
Possible diagnostic test reviews from updated DARE search
n = 142
Total sample
n = 331
2x2 data presented?
n = 154
Published 2000 to 2005
n = 97
Data on ≥ 1 spectrum-related covariate presented per study
n = 29

Table 25 Summary details of review topics and covariates

id	Review	N	Median sample size (SD)	Test	Topic	Covariates investigated			
1	Balk 2001[180]	32	101 (355)	presentation myoglobin	acute cardiac ischemia	definition of population			
2	Bricker 2000[179]	11	7,575 (9,324)	ultrasound	pregnancy	setting[b]	gestational age range screened[c]	risk status	
3	Buchanan 2001[181]	21	293 (880)	clinical assessment	dangerous severe personality disorder	sample type	time at risk		
4	Chapell 2002[182]	13	85 (115)	distal motor latency: symptoms/presented patient groups	carpal tunnel syndrome	presence of age bias	presence of bias to easy cases	selection of diseased patients	
5	Delgado 2003[178]	15	20 (12)	F18-FDG PET	detection of primary tumours in patients with metastasis	unknown primary tumour characteristics			
6	Deville 2000[123] [a]	17	182 (928)	straight or cross leg raising test	herniated discs	previous surgery	bilateral excluded[c]	gender	
7	Dijkhuizen 2000[100]	33	120 (174)	endometrial sampling	endometrial carcinoma	menopausal status[bc]	symptomatic status[b, c]		
8	Eden 2001[183]	7	102 (781)	palpation	thyroid cancer screening	source of exposure			
9	Flemons 2003[184]	49	71 (129)	sleep monitors	sleep apnoea	setting	gender	mean apnoea-hypopnea index, i.e no. events per hour of sleep	mean BMI
10	Flobbe 2002[185]	22	213 (478)	mammography	breast cancer	patient identification			
11	Gifford 2000[186]	11	202 (108)	clinical assessment	potentially reversible causes of dementia	age	setting	patient identification	
12	Glas 2003[80]	26	107 (76)	cytology	primary bladder cancer	% Grade 1 tumours	%Grade 2 tumours	%Grade 3 tumours	% urological controls[c]
13	Gould 2001[107]	35	46 (27)	FDG-PET	lung cancer	gender	age		
14	Gould 2003[187]	33	49 (44)	PET	mediastinal staging of non small cell lung cancer	gender	age[b, c]		
15	Gray 2000[188]	14	85 (301)	toludine blue dye in visual screening	oral cancer	patient identification			
16	Ioannidis[189]	10	295 (439)	out-of-hospital ECG	acute myocardial infarction	patient identification	age	gender[b, c]	
17	Kittler 2002[104]	13	172 (890)	dermoscopy	melanoma	inclusion of non-melanocytic lesions			
18	Koelemay 2001[190]	19	96 (71)	MRA	peripheral arterial disease - aortoiliac tract	age	gender	%intermittent claudication	

id	Review	N	Median sample size (SD)	Test	Topic	Covariates investigated			
19	Lysakowski 2001[191][a]	7	66 (32)	transcranial Doppler	vasospasm due to ruptured cerebral aneurysm	heterogeneity of population[b, c]			
20	MSAC 2002[192]	12	77 (176)	cytogenetic tests	fragile X syndrome	gender	patient identification		
21	Nallamothu 2001[193]	14	104 (63)	electron beam computed tomography	coronary artery disease	age	gender		
22	Patwardhan 2004[194]	19	43 (31)	PET	Alzheimer disease dementia	age	type of controls		
23	Romagnuolo 2003[195]	46	63 (53)	MRI cholangiopancreatography	bilary disease - detection of stones	patient identification			
24	Sauerland 2004[196]	13	219 (577)	clinical examination	pelvic fractures	age group			
25	Sotiriadis 2003[197]	12	4,308 (4,642)	intracardiac echogenic foci	Down syndrome	age	risk setting		
26	Varonen 2000[198]	7	156 (74)	ultrasound	acute maxillary sinusitis	setting			
27	Visser 2000[119]	21	404 (739)	Duplex ultrasound	peripheral arterial disease	gender	age	setting - country	
28	Whitsel 2000[114]	17	58 (772)	Bazett's heart rate-corrected QT interval (QTc)	autonomic failure in diabetes	age	gender	% type 1 diabetes	mean duration of diabetes
29	Wiese 2000[129][a]	30	175 (294)	wet mount technique	vaginal trichomoniasis	setting[b, c]			

[a] denotes covariates for which the overall pooled HSROC analysis could not be completed

[b] denotes covariates for which the parallel curve HSROC analysis could not be completed

[c] denotes covariates for which the crossing curve HSROC analysis could not be completed

N – number of studies

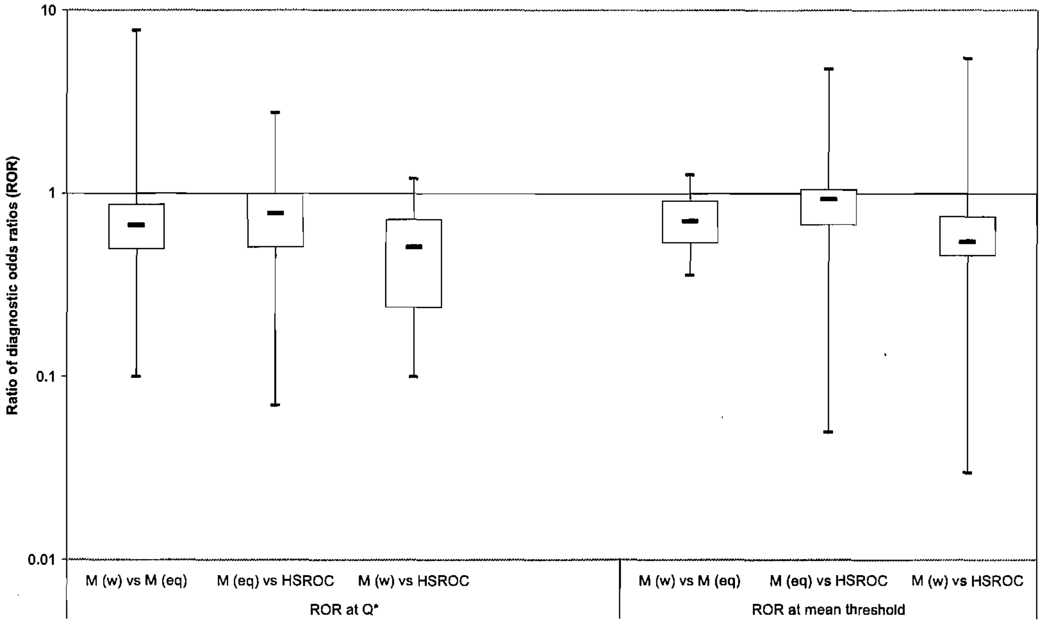
investigated topic was cancer (9 of 29 reviews), followed by heart disease (3 reviews), peripheral arterial disease (2 reviews) and dementia (2 reviews) (Table 24).

5.2.2 Comparison of primary analysis results

The HSROC analyses could not be completed for three of the 29 datasets.[123,129,191] The ROC plots for these three datasets reveal that SROC analyses for these datasets are probably not appropriate (Appendix 23), with data lying along the sensitivity=specificity line,[123] scattered mainly in the bottom left quadrant of the ROC space,[191] or showing all studies with specificity approaching 100%.[129] The comparisons of the two Moses methods are therefore based on 29 comparisons, while the comparisons with the HSROC method are based on 26 comparisons.

Figure 25 Similarity of DOR estimates between models

Box and whisker plot showing ratio of DORs between models: median, interquartile range (box) and range (whiskers), where weighted Moses model is compared to the unweighted Moses model (denominator) and each Moses model is compared to the HSROC model (denominator)



Maximum ROR	7.78	2.77	1.21	1.27	4.81	5.51
75th percentile	0.87	0.99	0.72	0.91	1.05	0.75
Median ROR	**0.67**	**0.78**	**0.51**	**0.71**	**0.94**	**0.55**
25th percentile	0.50	0.51	0.24	0.54	0.68	0.46
Minimum ROR	0.10	0.07	0.10	0.36	0.05	0.03

ROR – ratio of diagnostic odds ratios; Moses (w) – weighted Moses model; Moses (eq) – unweighted Moses model; HSROC – hierarchical SROC model; Q* - point where sensitivity=specificity; mean threshold – operating point estimated using mean threshold across studies

DOR estimates

Figure 25 shows that the weighted Moses model underestimates the results of the unweighted model on average by about 30% (ROR at Q* 0.67; ROR at average threshold

0.71). The IQRs are similar indicating that most comparisons are similarly spread around the median (similar levels of agreement) at Q* and at the average threshold but that the overall range in differences in results is much greater at Q*. This indicates that when DOR is estimated at Q* there is more scope for extreme differences between models compared to the DOR at the average threshold.

On average, both Moses models under-estimate the DOR in relation to the estimate from the HSROC model, with the weighted Moses model showing the biggest under-estimation of effects. For the unweighted model, the under-estimation at the median is less extreme at the average threshold compared to at Q*, with an ROR of 0.94 (Figure 25), indicating little bias on average. The width of the IQR is also narrower at the average threshold but the overall range in results is greater. This suggests that most of the data is less biased when DOR is estimated at the average threshold compared to at Q* (tighter IQR), but where there are observed biases they are more extreme (wider overall range).

The weighted model, on the other hand, on average underestimates the HSROC by 45 to 50% regardless of where DOR is estimated. A similar pattern in IQR and range to that for the unweighted model comparison can also be observed with a wider IQR at Q* (less agreement) but narrower range (less extreme disagreements).

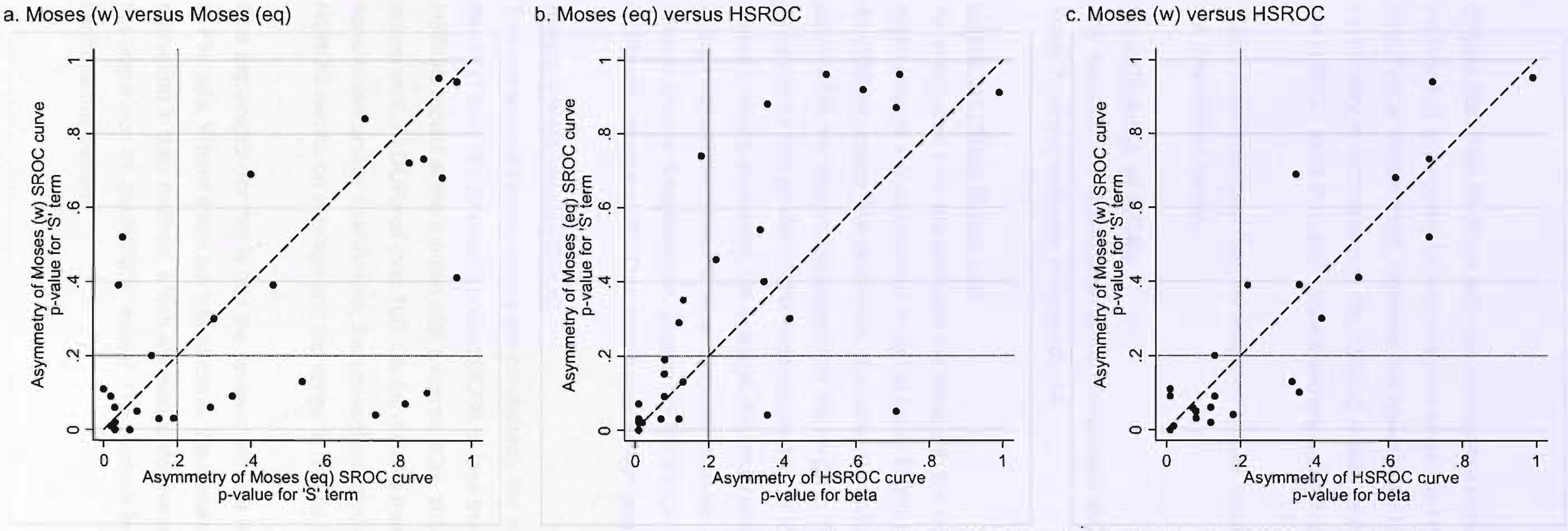
SROC curve shape

The extent to which the different models show similar evidence of asymmetry of SROC curves is demonstrated from the pairwise comparisons of P-values for the shape terms in (Figure 26).

The comparison of results from the two Moses models (Figure 26a) shows relatively poor agreement between the models, and a tendency for the weighted model to find more asymmetry than the unweighted model. Six out of 29 analyses with the weighted model found asymmetry to P<0.20 when the unweighted model found no such evidence (P>0.20). Only two of the unweighted analyses found curve asymmetry (P<0.20) when the weighted model did not (P>0.20).

For the comparisons of the Moses models with the HSROC model, agreement is better at lower P-values; i.e. where the HSROC model shows strong evidence of asymmetry, both Moses models also reach similar conclusions. The HSROC model finds asymmetry to P<0.20 for 14 of the 26 reviews for which the analyses could be completed. The P-values from the weighted Moses model agree more closely with the P-values from the HSROC model

Figure 26 Agreement in strength of evidence for asymmetry of SROC curves



'P-value' plots comparing the P-values for the shape terms between models. For each comparison the model represented on the x-axis indicates the reference case.

Central diagonal line indicates perfect agreement between methods

(Figure 26c) than do those from the unweighted model (Figure 26b). Both Moses models find evidence of asymmetry for two analyses when the HSROC model finds no evidence (over-detection of asymmetry)[j]. However the unweighted Moses model also under-detects asymmetry in comparison to the HSROC model for three analyses (i.e. it finds P>0.20 when the HSROC finds P<0.20)[k], while there are no such examples for the weighted Moses model.

These results suggest that the weighted Moses model is more sensitive to asymmetry than the unweighted model.

Stratification of RORs

The results of the stratification of the comparison of DORs between models are presented in Table 26 and graphically in Appendix 24.

Moses (w) versus Moses (eq)

As anticipated from the analyses in Chapter 4, the weighted Moses model more closely approximates the unweighted model at lower DORs and at smaller ranges in 'S' (Table 26). As DOR increases (the proportion of studies with exceptionally high sensitivities and/or specificities per review increases) and the range in 'S' increases (bigger variation in the threshold for test positivity) the under-estimation of the weighted compared to unweighted model's results increases. On average, the underestimation is less extreme when DORs are at their highest compared to when they are between 35 and 100, however the IQR is wider shower greater disagreement. Bias in the SE(lnDOR) is common even at moderately high DORs and ranges in 'S'. This data is also shown graphically in Appendix 24, Figure i and ii.

Moses models versus HSROC

The unweighted Moses results are consistently the most similar to those of the HSROC model (Table 26). At overall pooled DORs of less than 100 the results are identical to the HSROC model at the median with a narrow IQR, showing on average no bias and close agreement. At DORs of over 100, i.e. for reviews that include studies with exceptionally high sensitivities and/or specificities, the unweighted model considerably underestimates the HSROC results on average and the range in results is wider showing greater disagreement.

One explanation for this is that the reviews with the highest DORs will have the most zero FP or FN cells. Where there are lots of zeros, the Moses models will have added 0.5 as a correction in their method, which will lead to downward bias in the estimate of the odds ratio in comparison to the HSROC model. The analysis by presence of zero cells in Table 26 confirms this pattern.

[j] See Appendix 19, analyses 13 and 25 for the unweighted model and 7 and 8 for the weighted model.
[k] See Appendix 19, analyses 1, 4 and 23.

Table 26 Stratified comparison of DOR estimates between models

Number of reviews[a]	Moses (w) vs Moses (eq) n=29	Moses (eq) vs HSROC n=26	Moses (w) vs HSROC n=26
	Median ROR (p25, p75	median ROR (p25, p75)	median ROR (p25, p75)
ALL, n= 26 (29)	0.71 (0.54, 0.91)	0.94 (0.68, 1.05)	0.55 (0.46, 0.75)
by size of DOR[b]			
DOR < 35, n=11 (13)	0.86 (0.63, 0.96)	1.01 (0.88, 1.16)	0.75 (0.49, 0.95)
DOR 35-100, n=7 (7)	0.60 (0.54, 0.76)	1.00 (0.83, 1.05)	0.53 (0.42, 0.69)
DOR > 100, n=8 (9)	0.71 (0.55, 0.96)	0.67 (0.35, 0.74)	0.49 (0.15, 0.59)
by range in 'S'[c]			
range 3 to <6, n=7 (8)	0.76 (0.715, 0.89)	0.82 (0.68, 1.03)	0.75 (0.52, 0.79)
range 6 to <8, n=13 (14)	0.66 (0.53, 0.96)	0.88 (0.68, 0.99)	0.53 (0.49, 0.75)
range ≥8, n=6 (7)	0.55 (0.41, 0.86)	1.14 (1.05, 1.17)	0.51 (0.42, 0.73)
by % zero cells[d]			
<5%, n=9 (10)	0.81 (0.54, 0.98)	1.01 (0.88, 1.05)	0.78 (0.53, 0.88)
5 to 10%, n=9 (9)	0.60 (0.53, 0.66)	1.04 (0.95, 1.11)	0.57 (0.49, 0.73)
>10%, n=10 (8)	0.74 (0.58, 0.96)	0.67 (0.36, 0.72)	0.49 (0.16, 0.70)
by strength of evidence of asymmetry[e]			
beta, p<0.10, n=9	x	1.05 (1, 1.16)	0.53 (0.49, 0.57)
beta, 0.10≤p<0.35, n=6	x	0.84 (0.65, 1.11)	0.49 (0.45, 0.95)
beta, p≥0.35, n=11	x	0.82 (0.68, 0.95)	0.75 (0.5, 0.78)
by strength of evidence of threshold[f]			
theta, p<0.10, n=14	x	0.96 (0.81, 1.11)	0.61 (0.49, 0.79)
theta, 0.10≤p<0.35, n=5	x	0.59 (0.12, 0.83)	0.24 (0.07, 0.5)
theta, p≥0.35, n=7	x	0.99 (0.75, 1.04)	0.75 (0.46, 0.78)

Comparisons are between DORs estimated at the mean threshold. The weighted Moses model is compared to the unweighted Moses model (denominator) and each Moses model is compared to the HSROC model (denominator)

ROR – ratio of diagnostic odds ratios; Moses (eq) – unweighted Moses model; Moses (w) – weighted Moses model; median – ROR at the median; p25 – ROR at the 25th percentile; p75 – ROR at the 75th percentile

[a] The analyses per comparison is 29 for the Moses comparisons and 26 for the HSROC comparisons. The numbers in brackets indicate the numbers per subgroup for the Moses comparisons.

[b] The stratification by DOR is based on the HSROC overall pooled estimate; where the HSROC model did not run, it is based on the unweighted Moses model result.

[c] based on values for 'S' from Moses model

[d] number of zero false positive or false negative cells as a percentage of the total number of cells per analysis

[e] based on P-value associated with shape term from HSROC model

[f] based on P-value associated with threshold term from HSROC model

The weighted Moses model on average underestimates the HSROC results, but with lower median RORs and larger interquartile ranges (Appendix 24, Figure i). It also shows on average greater underestimation of DOR at overall pooled DORs of over 100.

The unweighted Moses model on average underestimates the HSROC DORs at smaller ranges in 'S' and slightly over-estimates the HSROC model at larger ranges in 'S'; the IQR narrows. The under-estimation of the HSROC DOR by the weighted Moses model, however, worsens as the range in 'S' increases and the IQR widens.

Where the HSROC model shows strong evidence of asymmetric SROC curves ($P<0.10$), the unweighted Moses model DORs quite closely approximate those from the HSROC model at the median, showing on average little bias. The IQR is quite narrow showing little disagreement (Table 26). As the evidence of asymmetry becomes less strong, the unweighted Moses model on average underestimates the DOR in comparison to the HSROC model and there is more disagreement. The weighted Moses model results are in the opposite direction, showing stronger underestimation of the HSROC DOR and less disagreement, the more asymmetric the SROC curves.

There does not appear to be any clear pattern in results according to the strength of threshold effects found by the HSROC model.

Summary

The weighted Moses model is strongly affected by the presence of bias in the SE(lnDOR). It underestimates the unweighted model by about 30%, and the underestimation is worse at higher DOR and wider ranges in 'S', both circumstances in which biased SE(lnDOR) would be expected.

The weigthed Moses model underestimates the DORs obtained from the HSROC model, and performs consistently worse than the unweighted model in the stratified analyses. However, the weighted model is also more likely to find curve asymmetry compared to the unweighted Moses model, and its P-values are in closer agreement with the HSROC model, compared to the unweighted model.

5.2.3 Comparison of heterogeneity investigations

The 29 reviews meeting the inclusion criteria provided information on 60 spectrum-related covariates (Table 25). The most commonly investigated characteristics were age (11/60), gender (10/60) and characteristics related to patient identification (n=10). Setting was investigated in 6 reviews. The remaining characteristics were largely more topic specific, such as the percentage of patients with "previous surgery",[123] the percentage with "type 1 diabetes",[114] or with "intermittent claudication".[190] For this section, three comparisons were undertaken: the two Moses models against the HSROC model; and for the HSROC model only, a comparison of results with and without an interaction of covaraite and shape.

Of the 60 investigated covariates, the HSROC model could not be completed for one covariate for the parallel curve version, four covariates for the crossing curve version and for five covariates for either model. These are denoted in Table 25. For 6 of the 10 covariates there were insufficient numbers of studies in at least one of the subgroups (less than 5 studies) and in the other four the studies exhibited exceptionally high specificities with varying sensitivities. The number of covariates per comparison therefore varies:

- Either Moses model versus HSROC with parallel curves, n=54
- Either Moses model versus HSROC with crossing curves, n=51
- HSROC parallel versus crossing, n=50

Table 27 Comparison of relative diagnostic odds ratios (RDORs) between models

	M (eq) vs HSROC		M (w) vs HSROC	
Parallel curve models	**Median (IQR)**	**Range**	**Median (IQR)**	**Range**
Ratio of RDORs at Q*	0.87 (0.58, 1.02)	0.02, 1.93	0.80 (0.54, 1.05)	0.05, 1.90
Crossing curve models				
Ratio of RDORs at Q*	0.66 (0.32, 1.63)	<0.01, 76.8	0.76 (0.26, 1.45)	<0.01, 35.63
Ratio of RDORs at average reference group threshold	0.70 (0.32, 1.26)	<0.01, 4.69	0.64 (0.39, 1.10)	<0.01, 11.81
Ratio of RDORs at average comparator group threshold	0.64 (0.39, 1.12)	<0.01, 18.16	0.66 (0.34, 0.91)	<0.01, 42.14

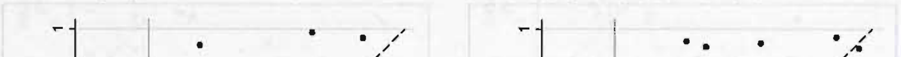
Each Moses model is compared to the HSROC model (denominator); IQR – interquartile range; RDOR – relative diagnostic odds ratio; RROR – ratio of RDORs

NB: The very extreme ranges for the crossing curve models have occurred in reviews with very small numbers of studies in one of the comparator groups leading to very big differences in RDORs between models

Comparison of relative RDORs and RDOR P-values

The comparison of RDORs between models shows that on average disagreement is common regardless of whether a shape interaction is included or not and regardless of the point at which the RDOR is estimated (Table 27 and Appendix 25). Both Moses models considerably underestimate the RDORs on average from the HSROC model and to a similar extent. The under-estimation is less for the parallel curves but nevertheless, they still on average underestimate the HSROC RDOR by 13% (unweighted Moses) and 20% (weighted Moses). For all estimates the IQR covers a wide range in values both over and under-estimating RDOR, showing considerable disagreement between methods. The range of disagreement is less for the comparisons of parallel curve models (with narrower IQRs).

Figure 27 Comparison of P-values for RDORs between parallel curve models


a. Moses (eq) versus HSROC

b. Moses (w) versus HSROC

'P-value' plots comparing the P-values for the relative diagnostic odds ratios (RDORs) between models. For each comparison the model represented on the x-axis indicates the reference case.

The visual comparison of P-values for RDORs from parallel curve models (Figure 27) suggests that neither Moses model has better overall agreement in terms of the strength of evidence for the effects of covariates on accuracy. However, taking P<0.20 as providing moderate to strong evidence of differences in accuracy, the weighted Moses model was more likely to find strong evidence of differences where the HSROC model finds none (7 out of 54 investigations found P<0.20 when HSROC found P>0.20, Table 28).

Table 28 Agreement in strength of evidence for differences in accuracy between models (RDOR P-values at P<0.20)

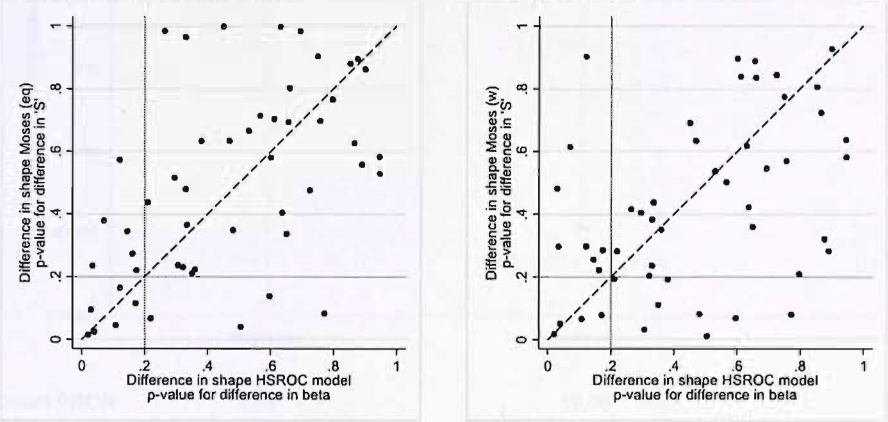
	HSROC P<0.2		HSROC P>0.2	
	Moses (eq) P>0.2	Moses (w) P>0.2	Moses (eq) P<0.2	Moses (w) P<0.2
Parallel curve models (n=54 comparisons)	5 (9%)	2 (4%)	3 (6%)	7 (13%)
Crossing curve models				
RDOR at Q* (n=51 comparisons)	7 (14%)	5 (10%)	6 (12%)	10 (20%)
RDOR at average reference group threshold	4 (8%)	8 (16%)	5 (10%)	8 (16%)
RDOR at average comparator group threshold	2 (4%)	7 (14%)	4 (8%)	9 (18%)

The final column of Table 28 shows that this trend was not strongly maintained for the RDORs from the crossing curve model with both weighted and unweighted models over and under-detecting differences in accuracy compard to the HSROC model.

Figure 28 Comparison of strength of evidence for differences in shape between models (comparison of P-values for shape interaction term)

a. Moses (eq) versus HSROC

b. Moses (w) versus HSROC



'P-value' plots comparing the P-values for the relative diagnostic odds ratios (RDORs) between models. For each comparison the model represented on the x-axis indicates the reference case.

Differences in SROC curve shape between groups

For the estimation of differences in curve shape between subgroups (Figure 28a and b) there is again little overall agreement between models. Taking P<0.20 as providing moderate to strong evidence of differences in shape (Table 29), there were little differences between the Moses models in the level of agreement with the HSROC model. Both unweighted and weighted models over and under-estimated differences in shape compared to the HSROC model.

Table 29 Disagreement in strength of evidence for differences in shape between models (comparison of P-values for shape differences at P<0.20)

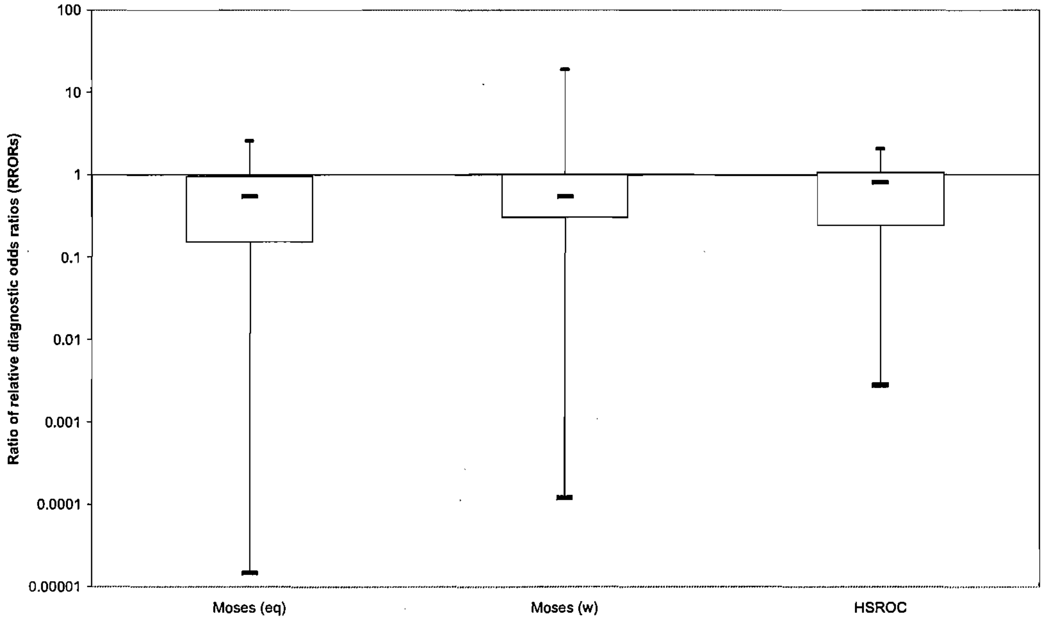
	HSROC P<0.2		HSROC P>0.2	
	Moses (eq) P>0.2	Moses (w) P>0.2	Moses (eq) P<0.2	Moses (w) P<0.2
No. of investigations (n=51)	6 (12%)	8 (16%)	4 (8%)	8 (16%)

Comparison of parallel versus crossing curve models

Figure 29 presents within model comparisons of RDORs at Q* with and without the interaction of covariate with shape, i.e. parallel versus crossing curve versions of the models. This shows that for each model on average, the parallel curve versions under-estimate the RDORs compared to the crossing curve versions, by up to 50% for the Moses models and 20% for the HSROC model. The interquartile ranges are wide for all three models, showing considerable scope for disagreement between methods.

Figure 29 Parallel versus crossing SROC curve models: Ratio of RDORs at Q*

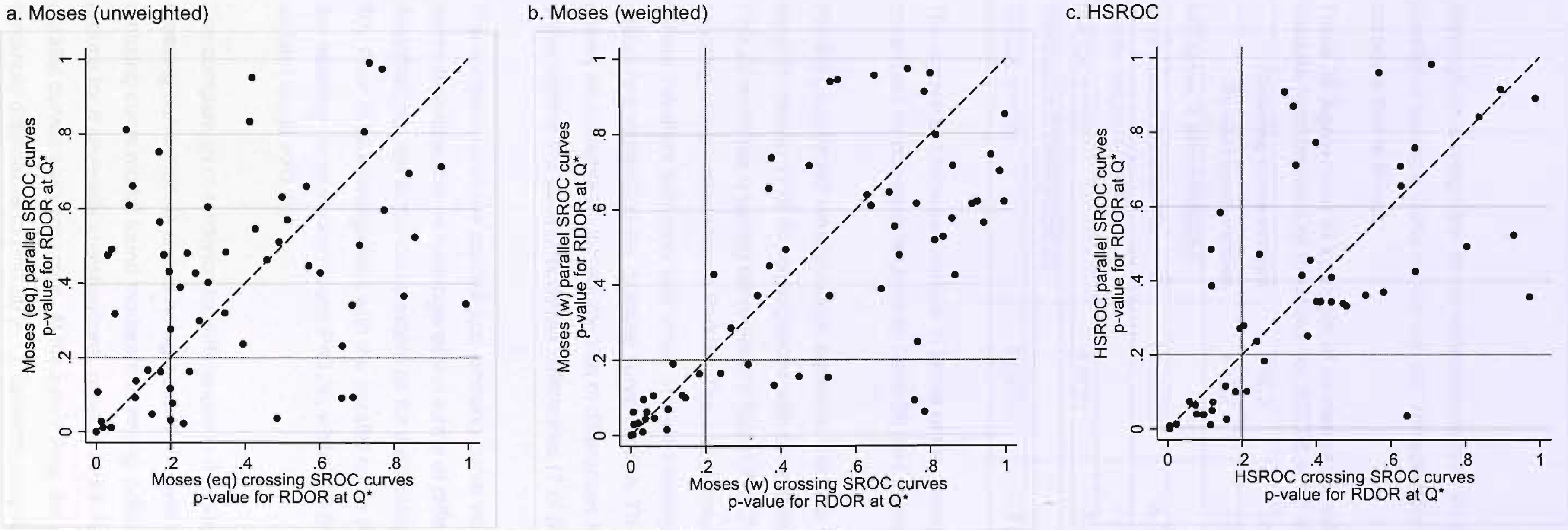
Box and whisker plot showing ratio of RDORs between models: median, interquartile range (box) and range (whiskers), where crossing curve version of each model is the reference case (denominator)



	Moses (eq)	Moses (w)	HSROC
Maximum RROR	2.58	19.05	2.07
75th percentile	0.96	1.01	1.07
Median RROR	**0.55**	**0.55**	**0.81**
25th percentile	0.15	0.31	0.24
Minimum RROR	<0.01	<0.01	<0.01

RROR – ratio of RDORs between models; Moses (w) – weighted Moses model; Moses (eq) – unweighted Moses model; HSROC – hierarchical SROC model
NB: The very extreme ranges for the crossing curve models have occurred in reviews with very small numbers of studies in one of the comparator groups leading to very big differences in RDORs between models

Figure 30 Comparison of P-values for RDOR between parallel (PA) and crossing curve (XG) models



'P-value' plots comparing the P-values for the relative diagnostic odds ratios (RDORs) between models. For each comparison the model represented on the x-axis (crossing curve version of each model) indicates the reference case. Central diagonal line indicates perfect agreement between methods.

Although on average the under-estimation is less for the HSROC comparison, the choice of parallel or crossing curve model can still considerably affect the magnitude of the difference in accuracy that is found.

Table 30 Agreement in strength of evidence for differences in accuracy between models (comparison of P-values for RDOR at Q* and threshold at P<0.20)

Crossing curve version	P<0.2	P>0.2
Parallel curve version	**P>0.2**	**P<0.2**
ACCURACY DIFFERENCES		
Moses unweighted (n=60)	11 (18%)	8 (13%)
Moses weighted (n=60)	0	7 (12%)
HSROC (n=50)	4 (8%)	3 (6%)
THRESHOLD DIFFERENCES		
HSROC (n=50)	6 (12%)	12 (24%)

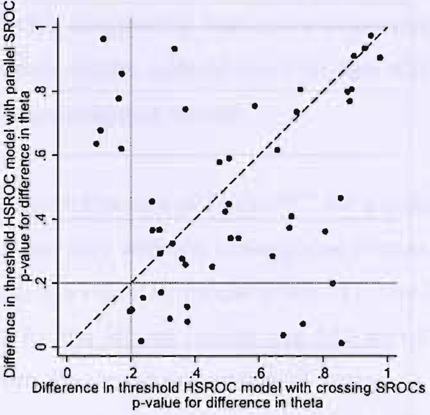
The agreement between models in terms of the strength of evidence for the effects of covariates on accuracy is given in Table 30 and Figure 30.

For the unweighted Moses model, agreement is poor both overall and at the more important lower P-values: 11 of 60 investigations with the parallel curve version of the model found P>0.20 when the crossing curve version found P<0.20, while 8 of 60 found P<0.20 while the crossing curve version found P>0.20. The unweighted Moses model with parallel SROC curves therefore both over and under detects heterogeneity in terms of differences in accuracy compared to the crossing curve version. The parallel curve weighted Moses model shows no evidence of under detection of differences in accuracy compared to the crossing curve version but does over detect differences (7 of 60 investigations).

The comparison of the parallel and crossing curve versions of the HSROC model shows some over-detection of heterogeneity in terms of differences in accuracy and some under detection, but not to the same extent as for the unweighted Moses model comparison (Table 30). Four of 50 investigations with the parallel curve version of the model found P>0.20 when the crossing curve version found P<0.20, while 3 of 50 found P<0.20 while the crossing cure version found P>0.20.

The comparison of evidence for differences in threshold between the HSROC parallel and crossing curve models shows no agreement at lower P-values (Table 30 and Figure 31). The crossing curve model found moderate to strong evidence of threshold differences between groups for 6 covariate investigations, none of which had P-values of less than 0.20 when parallel curves were assumed. At the same time, the parallel curve model found evidence of threshold differences to P<0.20 in 12 datasets, only two of which relatively closely agreed with the results when curves were allowed to have different shapes.

Figure 31 Comparison of P-values for threshold differences between HSROC parallel and crossing curve models



'P-value' plot comparing the P-values for the differences in threshold between the parallel curve and crossing curve (reference case) versions of the HSROC model.

Summary

The two Moses models underestimate the size of differences in accuracy between groups compared to the HSROC model; the difference is less when parallel curves are modelled but nevertheless remains. Both models find strong evidence for differences in accuracy when the HSROC model does not (over-detection of differences), the weighted model more so than the unweighted model. Both models also do not indicate evidence for differences in accuracy that are identified by the HSROC model (under detection of differences). For the detection of differences in curve shape between groups, the unweighted model most closely agreed with the HSROC models, but still both over and under-detected such differences.

The within model comparisons of parallel and crossing curve versions of the models showed that this choice will almost always affect a review's conclusions regarding the size of any differences in accuracy according to a given covariate, sometimes to quite a considerable extent. Differences in the strength of the evidence for differences also vary by choice of model. The effect on the size and strength of the difference in accuracy is less for the HSROC model but nevertheless occurs.

5.2.4 Selected illustrative examples – Moses versus HSROC

Primary Analyses

While the clinical implications of the under-estimation of diagnostic accuracy are relatively simple to interpret (tests will on average appear less accurate when analysed using the Moses methods), the implications of the over or under-detection of asymmetry are less intuitive. SROC curve asymmetry is introduced when the distribution of test results differs between diseased and nondiseased participants and means that accuracy (DOR) is not constant, i.e. it varies along the length of the SROC curve. This in turn means that the apparent accuracy of a test will vary according to the point at which the DOR is estimated.

The HSROC model found curve asymmetry to $P<0.20$ for 14 of the 26 reviews analysed (54%), suggesting that curve asymmetry is a relatively common occurrence. The weighted Moses model agreed more closely with the HSROC model results regarding asymmetry than the unweighted model.

Data on the use of FDG-PET for the diagnosis of lung cancer[107] show strong evidence of asymmetry with the unweighted Moses model (P=0.05) and no such evidence when analysed with the HSROC model (P=0.71). The DORs at Q* and at the average threshold are 127 and 72 for the Moses model and 142 and 107 for the HSROC model (Appendix 19). Quite apart from the under-estimation of accuracy, from the Moses model one would conclude that there are considerable differences in the distribution of test results between diseased and nondiseased, such that accuracy is not constant along the SROC curve. The HSROC model indicates that although there may be differences in the distribution of results (there is some variation in DOR) the differences are not statistically significant.

Heterogeneity analyses - Differences in RDORs

Where subgroup SROC curves are assumed to have the same shape (parallel curves), the HSROC model finds evidence of differences in accuracy between groups to $P<0.20$ for 18/54 (33%) analyses. The unweighted Moses model has a tendency to under-detect these differences, whilst the weighted model is more likely to over-detect differences.

An example of over-detection of differences in accuracy is provided by the review of MRA for the detection of peripheral arterial disease.[190] The weighted Moses model found some evidence to suggest that MRA is three times more accurate in studies of participants with a mean age of less than 65 than in those with on average older participants (RDOR 3.35, P=0.15). The HSROC model on the other hand found no evidence of differences (RDOR 0.89, P=0.91).

Heterogeneity analyses - Differences in shape

Differences in curve shape to $P<0.20$ were identified by the HSROC model for 12/51 (24%) investigations for which the analyses could be completed. Both Moses models over- and under-detected these differences.

An example of over-detection of differences in curve shape is provided by the review of straight or cross leg raising test for the detection of herniated discs.[123] The unweighted Moses model found strong evidence (P=0.07) that the SROC curves for patients having undergone previous surgery were different in shape to those who had not received previous surgery. This suggests that the distribution of test results between diseased and nondiseased differs between these subgroups. The HSROC model finds considerably less evidence to suggest these differences (P=0.22).

5.2.5 Evidence of spectrum effects – HSROC parallel versus crossing curves

If one takes the advanced models of meta-analysis as the best available tool for the synthesis of diagnostic test studies, the choice of a parallel (HSROC or BVN) or crossing curve (HSROC only) model is perhaps the most pertinent discussion to be had regarding the models assessed here.

Excluding the analyses that could not be completed for either model leaves 50 comparisons by spectrum-related covariates. The parallel curve model found evidence of differences in at least one parameter (accuracy or threshold) for 23 (46%) comparisons, compared to 25 (50%) for the crossing curve model (differences in accuracy, threshold or shape). Sixteen investigations showed evidence of differences in at least one parameter using both models, leaving a further 16 showing evidence of differences under only one model.

Of the 32 investigations showing strong effects from the covariate in question using either the parallel or crossing curve model, 7 were related to mean age, six to gender, 4 to setting, 10 to factors related to the identification of patients and 5 to particular clinical characteristics of the patients in question.

The crossing curve model was more likely to find differences in accuracy at Q*, at the average reference threshold or at the average comparator threshold (40% of investigations compared to 32% for the parallel curve version). The parallel curve model was more likely to find differences in threshold (24% compared to 14% for the crossing curve version), with little overlap in results between models (only one comparison showed evidence of differences in threshold under both frameworks). Differences in shape were identified in 24% of investigations.

Effects from spectrum-related covariates are therefore not uncommon however the choice of model is clearly key. Table 31 shows the differences in results between models. Where the two models both find differences in at least one parameter (n=16), both models suggest differences in accuracy in the majority of cases (15/16 for the parallel curve model and 14/16 for the crossing curve model; accuracy being the only parameter affected in 10 and 9 comparisons respectively). If the parallel curve model alone was employed differences in accuracy would be identified in a further one comparison and differences in threshold for six. However if the crossing curve model was employed, a further nine examples of differences in at least one parameter are identified. In six of the 9, this manifests as differences in accuracy. Allowing for differences in the distribution of test results between subsets of studies (crossing curve) therefore produces more evidence of differences in accuracy than assuming no such differences in distributions exist (parallel curves). Shape is the only parameter affected in a small number of examples (3/25).

Table 31 HSROC parallel versus crossing curve models: similarity of strength of evidence

Difference in parameters P<0.20*	Parallel curves N=50	Crossing curves[a] N=50	Difference to P<0.20			
			Both models		Parallel only	Crossing only
At least one parameter	23 (46%)	25 (50%)	16		7	9
			PA	XG		
Accuracy	16 (32%)	20 (40%)	15	14	1	6
Shape	-	12 (24%)	-	6		6
Threshold	12 (24%)	7 (14%)	6	3	6	2
accuracy alone	11 (22%)	12 (24%)	10	9	1	3
shape alone	-	3 (6%)	-	2		3
threshold alone	7 (14%)	1 (2%)	1	0	6	0
accuracy and shape only	-	3 (6%)	-	2		1
accuracy and threshold only	5 (10%)	0	5	0	0	0
accuracy, shape and threshold	-	5 (10%)	-	3		2
shape and threshold only	-	1 (2%)	-	0		0

* analyses for which either the parallel or crossing curve models would not complete are excluded

[a] difference in accuracy could be at Q*, at the average reference threshold or at the average comparator threshold

Variation in findings regarding differences in threshold and the added complexity from differences in shape complicate comparisons between models. Details of the results of the heterogeneity investigations are presented in Appendix 20 to Appendix 22. For illustrative purposes, two examples where differences in two or more model parameters were found are presented below.

Firstly, in a review of PET scanning for the detection of Alzheimer disease dementia,[194] the parallel curve version of the HSROC model finds no evidence of differences in accuracy according to whether healthy or diseased controls are recruited to the study (RDOR 1.91, P=0.39). When crossing SROC curves are modelled (i.e. the distribution of test results in diseased and nondiseased can vary between subgroups), some evidence of differences in accuracy by type of controls used is found (RDOR at Q* 5.70, P=0.12; RDOR at reference threshold 0.68, P=0.48; RDOR at comparator threshold 4.57, P=0.26). For this example the evidence for differences in curve shape and threshold between subgroups was not very strong (P=0.26 and P=0.37, respectively).

In a second example, stronger evidence for differences in curve shape (P=0.04) and threshold (P=0.09) between subgroups was identified.[184] For this review of sleep monitors for the diagnosis of sleep apnoea, the parallel curve HSROC model found no evidence for differences in accuracy by mean body mass index (BMI) above or below 30 (RDOR 1.54, P=0.48). However when the curves had different shapes, a more complex picture emerges. At Q*, studies of patients with a mean BMI of 30 or less were 11 times more accurate than those in patients with a mean BMI of greater than 30 (RDOR 11.07, P=0.12). The RDOR at the comparator group mean threshold was in the same direction (20.71, P=0.27), but at the reference group mean threshold sleep monitors were less accurate in studies with a lower mean BMI (RDOR 0.56, P=0.08). This example is similar to the example by reference test in Chapter 4; the SROC curves cross near to the centre of the data and additionally cross near to the comparator and reference group mean threshold points. The difference in accuracy depends on the point at which the RDOR is estimated.

5.3 Discussion

The purpose of this chapter was to determine the extent to which the findings in Chapter 4 could be generalised i.e. to examine the extent to which the meta-analytic models disagree and under what circumstances, and to determine whether spectrum-effects are more easily identified using any one of the methods. It should again be noted that both the HSROC and BVN models were applied to the TB data in Chapter 4 whereas for this chapter only the HSROC model was employed. This was because, where parallel curves are modelled the two models produce very similar results and also because the BVN model cannot easily incorporate an interaction of covariate with shape. As a result, some of the conclusions from Chapter 4 refer to the 'advanced models' whereas the discussion of the findings from this chapter refer only to the HSROC model.

The TB analyses found for the primary data analysis:

1. *considerable disagreement between the two Moses models*

This was supported by the re-analysis of review data for this chapter. The weighted Moses model, on average, consistently under-estimated the results of the unweighted model both for the DOR at Q* and at the average threshold. Stratification of the analyses showed the under-estimation to be exaggerated at higher pooled DORs and with wider ranges in 'S', i.e. in reviews of studies with exceptionally high specificity and/or variation in threshold. These are two of the characteristics that lead to (upward) bias in the SE(lnDOR). The third such characteristic is unequal numbers of diseased and nondiseased participants, but it was not possible to easily model this across multiple datasets. Weighting by the inverse of the SE leads to these studies receiving a very low weight in the weighted Moses analyses, so that the overall pooled DOR is lower in comparison to that of an unweighted analysis.

There was reasonable agreement between models regarding the presence of asymmetry in the SROC curves, however the weighted model found more curve asymmetry than the unweighted model.

2. *the unweighted Moses model results were most similar to those of the advanced models.*

This was also supported by the re-analysis of review data except for the detection of asymmetry. For estimation of DOR, the unweighted Moses model on average, only slightly under-estimated the HSROC results (ROR 0.94). The stratified analyses however, showed large under-estimations at the highest levels of DOR (i.e. over 100) and with increasing numbers of zero cells. The correction for zero cells (adding 0.5 to each of the four cells) will have led to downward bias in the estimate of the odds ratio in comparison to the HSROC model.

There also appeared to be a trend from under to over-estimation of the HSROC DOR by the unweighted Moses model as the range in 'S', or variation in threshold, increased. A similar pattern occurred with increasing asymmetry of the SROC curve with under estimations of DOR occurring where there was little or no evidence of asymmetry and over estimations of DOR occurring in the presence of asymmetry. This suggests that the Moses model cannot adequately deal with studies with very high specificities nor correctly model variation in threshold.

The weighted Moses model was more likely to find similar strength of evidence of asymmetry to the HSROC model, however the DORs for the reviews with asymmetric curves were less than half that of the HSROC model. The stratified analyses showed similar trends to those for the unweighted analysis, except by range in S. As the variation in threshold increased, the weighted model further underestimated the HSROC results.

The suggestions from the TB chapter regarding the addition of covariates to the models were as follows:

3. *where parallel SROC curves are modeled, the unweighted Moses model generally has results more similar to the advanced models than the weighted Moses model*

The reviews re-analysis data found some support for this finding but the differences between the unweighted and weighted models was small. At the median, the unweighted model showed slightly less bias in comparison with the weighted model, underestimating the HSROC model RDOR by 13% compared to 20% respectively. The interquartile ranges were almost identical, showing similar scope for disagreement with the HSROC results. The unweighted model was more likely than the weighted model to under detect differences in accuracy identified by the HSROC model, however the weighted model was more likely to over detect differences where the HSROC model found none.

4. *where curves have different shapes, neither Moses model consistently approximated the HSROC model results*

This finding was also supported by the reviews re-analysis data. Both Moses models on average considerably underestimated the RDORs of the HSROC model. The model that was closest to the HSROC varied by the point at which RDOR was estimated. The interquartile ranges were wide for all comparisons and also included over-estimations of the HSROC RDOR. The weighted model was also considerably more likely to over detect differences in curve shape between subgroups. This is a continuation of the feature noted in 2. above that the weighted model detects asymmetry more sensitively than the unweighted model. This data shows that it also detects asymmetry more sensitively than the HSROC model. However both models also under detected shape differences identified by the HSROC model.

5. *for the HSROC model, in some circumstances, the interaction of covariate with shape made little difference to the conclusions that would be drawn from the model regarding the importance of a covariate, but in others conflicting results arose*

There is again some evidence to support this observation. The average differences between parallel and crossing curve versions of the models were considerably less for the HSROC within model comparison than for the Moses comparisons, however on average, the parallel curve version of the model under-estimated the RDOR of the crossing curve model by 19% with an IQR from 0.24 to 1.07. There were reviews for which the choice of parallel or crossing curves made only a small difference to the RDOR, however in a considerable number, large differences were apparent. The agreement between parallel and crossing curve models in terms of strength of evidence for differences in accuracy was good, however, especially at lower P-values. Allowing for a shape interaction, however does lead to an increased number of covariates for which differences in accuracy (and in other parameters) is identified.

Both Moses models demonstrated much bigger differences in the magnitude of the differences in accuracy between the parallel and crossing versions of the model and in the strength of evidence of differences in accuracy. This shows that the choice of parallel or crossing curve model under the Moses framework, frequently has a large impact on conclusions regarding differences in accuracy.

A further finding from the TB analyses was common and sometimes considerable disagreement between the two Moses models regardless of whether parallel or crossing SROC curves were modeled. This was also examined for the reviews reanalysis dataset, but for simplicity, the data has not been presented. The observations seen for the TB data were supported by the reviews re-analysis with the weighted model on average consistently under-estimating the unweighted model.

The analyses in this chapter therefore provide considerable support for the general findings of Chapter 4. There are several implications of this data for the wider literature.

The detailed review of diagnostic test reviews in Chapter 3 showed that use of the weighted Moses model is common. Of 64 reviews using an SROC analysis, 20% (n=13) used an unweighted approach, 39% (n=25) a weighted model and a further 27 reviews did not specify any weighting schedule. Not all of the weighted reviews used the inverse of the variance of lnDOR as the weight, however under estimation of test accuracy in the literature due to the use of this weighting is clearly a problem as is over detection of asymmetric SROC curves.

Both Moses models can produce results very similar to the results of the HSROC model but on average they are much more likely to underestimate results both for primary analyses and for identification of differences in accuracy. Both models also over and under detect differences in accuracy and shape. This has huge implications for the majority of existing reviews of diagnostic tests. Taking the HSROC model as the benchmark, it is not too much of an exaggeration to say many hundreds of reviews have underestimated test accuracy and both over and under identified different aspects of heterogeneity.

These results not only potentially have real clinical significance but also may have consequences for our understanding of different biases in diagnostic test research. Empirical studies to identify and quantify sources of biases in diagnostic accuracy studies have used regression models adapted from the Moses models.[64,199] It is quite reasonable to assume that their results at the very least under-estimate the size of the biases in operation. It is likely that spectrum effects exist and that given the appropriate data can be detected, however use of either of the Moses methods to identify them will often lead to under-estimation of the size of any effect and to misleading indications of the strength of any effect.

The widespread use of Q* as the point at which to estimate DOR in itself introduces considerable bias. The Moses models are generally more biased when DOR is estimated at Q* compared to at the average threshold (wider IQR), although there are more extreme biases for DOR at the average threshold for the unweighted model. For the weighted Moses model compared to the unweighted model, the more extreme biases are at Q*, but on average the biases are similar at Q* and at the average threshold.

The second aim was to identify effects from spectrum-related variables. The analyses here were confined to spectrum-related variables and the presence of strong evidence of differences between subgroups suggests that such effects can be demonstrated using meta-analytic techniques. However, for many of the investigations only aggregated data such as the mean age or the percentage of men or women included could be examined. There may often be question marks over whether such variables are sufficiently good proxies for true spectrum-related variables. Furthermore, detection of true effects using aggregated data is problematic and often not applicable to individual patients.[200]

At the onset of the work on this thesis, it was hypothesised that an advantage of the advanced methods would be to identify any differential effects of spectrum-related covariates on sensitivity and specificity. Current thinking however is that given variations in threshold across studies it is more appropriate to compare SROC curves, i.e. to compare differences in accuracy threshold and shape, than it is to compare operating points.

The question remaining for the advanced methods is whether an interaction of covariate with shape should be routinely modelled or not. These results show that taking the approach of fitting the simplest model and ignoring any potential differences in distributions of test results between subsets of studies (difference in shape) can give a different answer to an approach where shape differences are directly modelled. These differences in results between models depends on the extent to which the distributions of test results between diseased and nondiseased differ according to the covariate in question. If the two or more subsets of studies exhibit similar patterns in these distributions, the associated SROC curves will have similar shapes and the crossing curve version of the model will produce results more akin to those of the parallel curve model. If the pattern in the distribution of results differs so that the SROC curve for one subset of studies is perhaps more asymmetric than the other, the parallel and crossing curve versions of the model would be expected to produce different results.

For the dataset used here, differences in at least one parameter were identified by both parallel and crossing curve models for around a third of all of the covariate investigations (16/50) and for most of these, differences in the accuracy parameter were found. Although a similar number of additional investigations with strong evidence for differences by covariate were identified using each model (7 for the parallel curve version and 9 for the crossing curve version), those identified by the crossing curve model may be more clinically significant. Of the 7 investigations for which significant differences were found with the parallel curve model alone, one indicated differences in accuracy and six indicated differences in threshold. Of the 9 investigations for which significant differences were found only with the crossing curve model, six indicated differences in accuracy, six in shape and two in threshold. Differences in accuracy suggest that the subgroups are operating on two different SROC curves, differences in shape that the relationship between the distribution of test results in diseased participants and nondiseased participants differs by covariate, and differences in threshold that the studies operate at different points on the curves.

The most appropriate approach to modeling, e.g. whether both models should always be carried out or whether one should start with the simplest approach and progress to more complex modeling if required, needs further work. The evidence presented suggests that although interpretation of results may be more complex when crossing curves are modelled, there is possibly a greater risk of missing covariate effects if only parallel curves are constructed. This question can perhaps only be addressed by simulation studies although these in themselves would be complex to design.

The strength of this review was the number of datasets available for reanalysis. This enabled the further investigation of observations identified from a single dataset in Chapter 4 so that the findings can be strengthened and generalised. Further investigation might identify certain circumstances under which the Moses methods more closely approximate those of the HSROC method, however the ease of use of the HSROC method is now such that it or the BVN model, should be the preferred approach. The main issue that requires further investigation is the circumstances under which the parallel and/or crossing curve models should be employed. Some of the extreme results from this dataset also emphasise that there are circumstances under which meta-analysis should not be undertaken and that this should be carefully assessed before any pooling is attempted.

6 Discussion

Chapter 1 introduced the concept of diagnostic accuracy as the means by which diagnostic tests are evaluated and also introduced that diagnostic tests can operate differently according to spectrum-related characteristics. Actual clinical examples of variations in accuracy by spectrum were presented and the mechanism of the effect explained. Namely, characteristics such as disease severity or symptoms in diseased persons and conditions similar to that of the target disorder in nondiseased persons can affect the response of an individual to a given test. The mix or distribution of these characteristics amongst the participants of any given study affects the distribution of test results in diseased and nondiseased persons and thereby the sensitivity and specificity of the test in question. The distribution of spectrum-related characteristics is unlikely to be constant across studies, therefore sensitivity and specificity will vary to a greater or lesser extent between studies.

Diagnostic accuracy studies are also subject to a host of other potential sources of variation including those related to test, methodology and threshold. It is rare for diagnostic accuracy studies to be sufficiently large in size or to recruit a sufficiently broad spectrum of participants to allow the influence of spectrum to be teased out from other potential sources of variation. Systematic reviews, and particularly meta-analysis, may therefore be the best available tool to identify the extent to which various the sources of heterogeneity, including spectrum, can affect test accuracy. Various methods of meta-analysis may be employed, however random effect models that specifically allow for threshold effects and for variation in test accuracy (DOR) with threshold are preferred.

Chapter 2 discussed in more detail the sources of heterogeneity other than spectrum and explained four methods of meta-analysis that allow for variability in threshold and for variation in DOR with threshold. These are the Moses model, unweighted and weighted by the inverse variance of the log of the DOR, and the so-called 'advanced models', the bivariate normal model (BVN) and the hierarchical SROC (HSROC) model.

Primary analyses with all four models produces an SROC curve which can be interpreted in terms of its DOR (a global measure of test accuracy) and shape (or degree of asymmetry). The advanced models also produce an estimate of threshold, indicating likely position on the SROC curve. The shape of the SROC curve depends on the distribution of test results in diseased and nondiseased persons. If the distribution, or variance, of test results around the mean is the same in diseased and nondiseased persons there will be no asymmetry in the SROC curve and it can be represented by a single constant DOR. If the variance in test results differs between diseased and nondiseased, one distribution perhaps being wider and flatter as might occur where a study recruits a considerable proportion of patients with advanced disease, the SROC curve will be asymmetric and the DOR will vary along it.

Sources of variation in test results are investigated by extending the models to allow for covariates. At the simplest level one assumes that the variances of test results in diseased and nondiseased participants do not differ according to the covariate, i.e. the shape of the SROC curves are the same (parallel curve models). A second level allows for an interaction of covariate with SROC shape. This means that the variances in test results of diseased and nondiseased persons can differ between groups; the SROC curves can therefore have different shapes and will cross at some point along their length (crossing curves). All four models allow the effect from covariates to be estimated in terms of the differences in accuracy between groups (relative DOR) and differences in shape[I] (the distribution of test results between diseased and nondiseased groups varies according to the covariate in question). The advanced models also allow differences in threshold to be estimated.

Only the 'advanced' BVN and HSROC models - which without the addition of covariates are in fact different parameterisations of the same model - possess the characteristics of an 'optimal' meta-analytic method, i.e. that a model should:

- be bivariate in its parameterisation and should allow interpretation in terms of sensitivity and specificity,
- use appropriate weighting to allow the different levels of uncertainty or precision associated with the sampling variability in TPR and FPR to be addressed,
- allow for the threshold relationship or correlation between sensitivity and specificity,
- use a random effects approach to allow for the almost inevitable heterogeneity that arises in a systematic review of a diagnostic test or tests.

It was not known to what extent the less optimal Moses methods might approximate the results of the advanced methods for the detection of spectrum effects.

Chapter 3 reported a methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy in a large sample of reviews published up until 2002 and in a smaller more recent sample of reviews that have used the advanced methods of meta-analysis. This showed that less than optimal methods of meta-analysis have been commonly employed. None of the reviews in the main sample employed the advanced methods of meta-analysis and less than half (48%; 64/133) of those using meta-analysis employed SROC type methods that allow for a threshold effect; the remainder pooled individual accuracy indices.

Of the 131 reviews carrying out quality assessment, 51% (n=67) considered patient spectrum in some way. Where sources of heterogeneity were investigated (102/133 meta-analyses), spectrum-related variables were commonly included (68%; 69/102) and 'statistically

[I] Differences in shape cannot be easily modelled within the standard BVN model framework and have not been modelled in this thesis, however the model is being developed to allow for differences in shape.

significant' results reported (41/69; 59%), although often the results were not reported in detail but referred to narratively. The small number of reviews (n=14) that reported their results in detail and also looked at covariates related to test, spectrum and quality showed similar percentages with statistically significant results, although the methods used to investigate heterogeneity varied. A statistically significant impact from spectrum-related factors was identified in 57% (n=8) of investigations, from test-related in 57% and from quality-related covariates in 43% (n=6) of investigations.

The small number of reviews identified that used the advanced models of meta-analysis showed overall improved systematic review methods, as would be expected from reviews coming from academic centres of excellence. The reviews were also more likely to have considered spectrum-related characteristics, although again this was sometimes restricted to consideration of the presence of an adequate description of patients. This is very likely due to lack of recording or reporting in the primary studies. Of the five reviews that examined spectrum-related characteristics, three found statistically significant effects, supporting the finding from the main dataset that when spectrum variables are reported to have been considered they are often found to have a significant effect.

Of the five reviews using advanced methods that also looked at spectrum-related covariates, three used the BVN model to examine effects on sensitivity and specificity; one used the HSROC model to examine effects on accuracy, threshold and shape and the last used the BVN model for the main analysis but appeared to develop a separate random effects meta-regression model to examine the effect of the covariates on the natural log of the DOR. For their reviews of diagnostic test accuracy, the Cochrane Collaboration recommend that the comparison of operating points should only be undertaken where there is an explicit constant threshold, even though similar threshold type effects could arise through differences in test interpretation between observers, characteristics of the sample and differences in the execution of tests. Where the explicit threshold for positivity varies between studies, the comparison of operating points should not be undertaken as the operating points have no direct interpretation; they are average points based on the average of the thresholds. The differential effects of a covariate on sensitivity and specificity in these circumstances are therefore cannot be identified.

Chapter 4 reported a detailed case study comparing the four meta-analytic methods on a large dataset. The methods' performance regarding primary analysis of the dataset and heterogeneity investigations with three selected covariates were compared. The three covariates (index test blinding, test type, and reference test used) were specifically chosen to reflect increasing levels of complexity in results. Four main observations requiring further investigation emerged from the analyses.

Firstly, there was common and sometimes considerable disagreement between the unweighted and weighted Moses models both for the primary analysis and for the heterogeneity investigations regardless of whether parallel or crossing SROC curves were modelled. For almost all comparisons, the weighted Moses model under-estimated the results of the unweighted model, in terms of accuracy or differences in accuracy. There was also some suggestion that the two models performed somewhat differently in terms of detection of differences in the distribution of test results between diseased and nondiseased (shape).

Further investigation showed that the disagreement between the two Moses models in terms of both accuracy and shape was primarily due to bias in the standard error of the log of the DOR or the SE(lnDOR), whose inverse was used as the weight for the weighted model. The SE(lnDOR) was shown to be biased for several studies in the dataset with very high DORs so that they had higher standard errors than might have been expected from their sample sizes. Weighting by the inverse of the standard error meant that these studies received a very low emphasis in the weighted Moses analysis, leading to overall under-estimation of effects in comparison to the unweighted analysis.

Secondly, the unweighted Moses model results were more similar to the HSROC model results than the weighted model but nevertheless still generally under-estimated effects.

The third observation was that for the primary analyses and for the investigation of heterogeneity with no interaction of covariate with shape (parallel curve models), the BVN model and the HSROC model produce almost identical results, as had previously been shown mathematically by Harbord and colleagues.[79]

Finally the inclusion of the shape interaction term in the HSROC model sometimes led to different conclusions regarding the effect of a covariate and sometimes not. This appeared to be related to the studies lying around the edges of the ROC plot. Studies with extreme values in sensitivity and/or specificity, or studies for which sensitivity estimates were greater than specificity or were similar in magnitude to specificity, had the biggest individual effects on accuracy and on shape. The group to which these studies were allocated according to covariate in turn impacted on the difference in model parameters between groups and the complexity of the differences between models.

Chapter 5 further explored these findings using data obtained from a large sample of previously published systematic reviews of diagnostic tests and using only spectrum-related covariates. The key findings of Chapter 4 were strongly supported by this re-analysis.

First of all, weighting the Moses model by the inverse variance of the lnDOR led to consistently lower results compared to the unweighted model. On average the weighted model underestimated the results of the unweighted model by around 30%, with considerable disagreement between models. The stratified analyses suggested that this underestimation is due to bias in the SE(lnDOR) and hence it is likely that the weighted model results are misleading. The circumstances under which biased SE(lnDOR) might be expected are: extreme values of sensitivity and specificity, often with zero FNs or FPs, unequal sample sizes of diseased and nondiseased patients, and variation in the threshold for test positivity leading to variation in the proportion of patients who are test positive. These circumstances are common in diagnostic meta-analysis, therefore bias in the SE is always a concern.

The comparison of the results of the primary analyses also showed that the weighted model is more sensitive to the presence of asymmetry, i.e. is more likely to suggest differences in the distribution of test results between diseased and nondiseased and therefore variation in DOR along the SROC curve and also that when DOR is estimated at Q* there is more scope for extreme differences between models compared to the DOR at the average threshold.

The implications of these findings are considerable. Chapter 3 found that use of the weighted Moses model is common. Not all of the identified reviews used the inverse of the variance of lnDOR as the weight, however under-estimation of test accuracy in the literature due to the use of this weighting is clearly a big problem as is over detection of asymmetric SROC curves.

Secondly, although the unweighted Moses model results were generally more similar to the HSROC model than the weighted Moses model, it cannot be relied upon to approximate the results of the 'optimal' HSROC model. For the primary analyses, when DOR was estimated at the average threshold there was on average little bias in the unweighted method compared to the HSROC method, however there was still considerable scope for disagreement between models and furthermore the biases could be quite extreme. At high DORs, the unweighted Moses model underestimated the HSROC model on average by 33% and in addition the differences ranged from under- to over-estimation as the range in 'S' increased, suggesting that the Moses model cannot adequately deal with studies with very high specificities nor correctly model variation in threshold.

For the investigation of heterogeneity, the unweighted Moses model consistently underestimated the differences in accuracy that were observed with the HSROC model. The underestimation was less when parallel SROC curves were modelled. The two models also differed in terms of their indication of the strength of evidence for differences in both accuracy and shape so that the unweighted Moses model found evidence for such differences when the HSROC model did not, and also did not detect differences that were identified by the HSROC model.

These results have considerable implications for the majority of existing reviews of diagnostic tests, casting some degree of doubt on the results of the many hundreds that have used the Moses model and especially the weighted Moses model to analyse their data. The presence of extreme differences in results between the Moses and HSROC model are particularly concerning, especially as one cannot necessarily predict the circumstances in which this might occur. Chapter 3 showed that up until 2002 at least, around half of diagnostic meta-analyses on the DARE database (i.e. reviews that had passed certain quality standards) employed SROC methods other than the HSROC method. The remaining half used methods that do not even allow for threshold effects.

Finally, the suggestion from Chapter 4 that allowing for differences in the distribution of test results between diseased and nondiseased by covariate (shape differences) sometimes affects the conclusions that would be drawn from an analysis and sometimes not was also supported by the reanalysis of review data in Chapter 5. There were differences in RDOR between parallel and crossing curve HSROC models however, the agreement in terms of strength of evidence for differences in accuracy was good, especially at lower P-values. This implies that although the magnitude of differences between groups may vary between models, the inclusion of a shape interaction term does not necessarily change the strength of evidence for differences in accuracy.

A key issue for the crossing curve model is the variation in RDOR along the curves. RDOR is most commonly estimated at Q* however this is not necessarily representative of the majority of the data. The alternatives presented here were to estimate RDOR at the average threshold of the each subgroup of studies, however where there are strong differences in shape so that the curves cross near to the centre of the data and where the expected operating points of the subgroups are some distance apart, the direction of effect can change according to where RDOR is estimated. It is potentially highly misleading to rely on estimates of DOR or RDOR at Q* alone.

It is not clear whether potential differences in the distributions of test results (differences in shape) should be routinely modelled or whether the more simple parallel curve approach will generally suffice. Differences in the distributions of test results were identified for 24% of heterogeneity investigations undertaken for Chapter 5. It is not known whether similar findings would occur for analyses of test or quality-related variables. It might be that in circumstances under which one might anticipate differences in the distribution of test results, such as for spectrum-related characteristics, both simple and more complex models should be constructed.

For example in Chapter 4, the comparison of the TB dataset according to the type of reference test used is likely to show differences in the distributions of test results because of the differing definitions of disease according to whether culture alone is used to indicate the presence of TB or whether a combined reference standard (culture plus clinical examination and other tests such as chest x-ray) is used. This is because culture is a far from perfect gold standard; a proportion of patients who are culture negative will in fact be found to have TB. Using a combined reference standard to indicate the presence of TB includes these patients as disease positive (along with a few patients who have clinical signs and symptoms similar to TB but do not in fact have the disease). This will lead to a larger number of diseased patients than would occur if culture alone was used. Because PCR amplifies the presence of mycobacterial DNA it is less likely to be able to do this in samples that failed to grow the mycobacteria (culture negative samples). The distribution of PCR test results will therefore be affected as the number of false negatives will be increased compared to if culture alone was used as the reference test.

The optimal approach for the investigation of heterogeneity requires further investigation however the question can perhaps only be fully addressed by simulation studies.

The final observation to make is the frequency of findings of strong evidence of effects from the spectrum-related variables that were investigated. Strong evidence of effects on at least one model parameter were identified by the parallel or crossing curve HSROC model for over half of the investigations conducted (32/50). This could have considerable implications for the use of tests in practice. For example, the skin test for the detection of TB infection is interpreted differently according to whether the patient has had a prior BCG vaccination.

It was notable also that both for the analyses in Chapter 5 and the review of reviews in Chapter 3, the spectrum-related variables investigated were not necessarily truly representative of the case mix of the patients; prevalence, for example, or age being commonly considered. One of the main challenges in the investigation of heterogeneity in systematic reviews are limitations in the primary study data. It can be particularly problematic to identify, let alone record and publish true spectrum-related characteristics in primary studies. Characteristics related to patient presentation and previous test results are likely to be the most relevant, however variables that are easier to measure and record such as age and sex are often used instead as proxies.

The STARD initiative[30] (Standards for Reporting of Diagnostic Accuracy) to promote the completeness and quality of reporting of diagnostic accuracy studies should help to improve the reporting of spectrum-related variables in the future, however meta-analysts must take care in setting their review question and defining their inclusion criteria in addition to being aware of the limitations in the data that is available. One of the key stages of any review is to describe the characteristics that describe the clinical problem to be addressed,[201] i.e. which clinical presentations would be recognised as suggesting the clinical problem? Overrestrictive inclusion criteria to certain subgroups of patients make it impossible to investigate key spectrum issues further down the line.

Although it can be argued that on an individual basis age and sex can be good proxies for true spectrum characteristics, when variables such as these are aggregated across participants, 'ecological bias' can occur, i.e. where there is insufficient data on which to fully investigate interactions between a covariate and a treatment effect or test accuracy. For example if all studies in a review demonstrate similar mean age of participants, a meta regression will fail to detect effects from age, however that does not mean that age does not influence effects.[200]

One solution to this problem is the use of individual patient data (IPD) analysis where the reviewer obtains raw study data directly from the original authors. This method is seen by many as the gold standard for meta-analyses to identify treatment effects[150,202] as it minimises bias and increases the power of statistical analysis and reanalysis, and its use has increased over the years.[200,202,203] The application of IPD analysis to diagnostic accuracy reviews is rare, however its potential benefits have been recognised.[150,204] Simulation work in the field of RCTs has shown that the statistical power of meta-regression techniques is dramatically and consistently lower than that of IPD analysis.[53] Nevertheless, care must be taken in the design and analysis of IPD studies. A matched comparison of subgroup analyses undertaken using IPD analysis and conventional analysis of RCTs demonstrated that although reviews using IPD analysis were more likely to investigate patient and diseased-related characteristics than those using conventional analysis, direct modelling of the raw data was rarely reported.[203] More commonly, "two-stage" analyses were undertaken such that the individual patient data was stratified by trial. This approach does not fully utilise the potential statistical power of the data available. Considerable time and resources are also required and IPD analysis should not be undertaken lightly.

The ideal solution for the assessment of diagnostic accuracy is for within study comparisons. Even with improved recording and reporting of study and patient characteristics, systematic reviews may never be the best way to get evidence of spectrum effects due to dilution from other confounding factors. It is possible that more important test- or methodology-related characteristics might affect accuracy in such a way as to dominate any spectrum effects. In RCTs, for example, lack of allocation concealment during the randomisation process introduces so much bias as to supersede any differences in patients. Within study comparisons require diagnostic studies to be sufficiently large, prospective, well-designed and multi-centre, evaluating a number of diagnostic tests (or variations on a test), thereby allowing test accuracy to be established as well as allowing the investigation of the influence of patient and other characteristics on accuracy.

In the meantime, although questions remain regarding the optimal approach to take with the advanced methods, such as the inclusion of interactions of covariate with shape in heterogeneity investigations, the results presented here lend further support for moves to increase the use of the advanced methods. The October 2007 launch of the new Cochrane Collaboration database for reviews of diagnostic test accuracy in the Cochrane Library gives reviewers much needed guidance on conducting diagnostic systematic reviews and meta-analyses. The complexity and challenges of conducting diagnostic accuracy reviews is recognised and reinforced by the requirement for the review author team to consist of authors with certain areas of expertise, including content expertise, review expertise and statistical expertise. Both the Moses and the advanced methods can be employed, however the handbook will include guidance on using the advanced methods. This will help to spread knowledge and understanding of the advanced methods.

Research implications

- Simulation studies are needed to find out which methods actually perform best and, if possible, the circumstances under which parallel or crossing curve models are more appropriate, i.e. are there circumstances or types of covariate for which which, as a general rule, one might expect differences in the distributions of test results in diseased and nondiseased?
- Large scale diagnostic accuracy studies should be performed to allow within study comparison of accuracy in different subgroups

Policy and practice implications

- Reviewers should be encouraged to use the more optimal advanced methods of meta-analysis in place of the Moses method and to carefully consider potential sources of heterogeneity including spectrum.
- The potential importance of spectrum effects in terms of the practical use of tests should be emphasised to clinicians. Clinicians also need a better understanding of summary ROC methods and their outputs, such as DOR and RDOR, and hw these can be applied to their clinical practice.
- Investigators conducting primary studies of diagnostic tests shoud be encouraged to appropriately record actual spectrum characteristics insteadof using proxies and to follow the STARD guidelines for reporting of their studies.

Appendices

Appendix 1 Calculation of diagnostic accuracy statistics

i) Contingency table (2 x 2 table)

		Reference standard		
		+ve	-ve	
		Diseased	Nondiseased	
Index test result	+ ve	True positives a	b False positives	Total test positive
	- ve	False negatives c	d True negatives	Total test negative
		Total diseased	Total nondiseased	

ii) Diagnostic accuracy indices

Sensitivity	Proportion of diseased who have positive test results	True positives / Total diseased a / (a + c)
Specificity	Proportion of nondiseased who have negative test results	True negatives / Total nondiseased d / (b + d)
Positive predictive value (PPV)	Proportion with positive test result who actually have the disease	True positives / Total test positive a / (a + b)
Negative predictive value (NPV)	Proportion with negative test result who really don't have the disease	True negatives / Total test negative d / (c + d)
Positive likelihood ratio (LR+ve)	Likelihood of a person with disease having a positive test result than a person without disease	(True positives / Total diseased) / (False positives / Total nondiseased) sensitivity / (1 – specificity)
Negative likelihood ratio (LR-ve)	Likelihood of a person with disease having a negative test result than a person without disease	(False positives / Total diseased) / (True negatives / Total nondiseased) (1 – sensitivity) / specificity
Diagnostic odds ratio (DOR)	The ratio of the odds of a positive test result in a patient with disease compared to a patient without disease	(True positives x True negatives) / (False positives x false negatives) LR +ve / LR -ve

Appendix 2 Specification of Rutter and Gatsonis HSROC model

The two-level random effects model is formulated in terms of the probability (π_{ij}) that a patient in study i with disease j has a positive test result, where j=0 for a patient without disease and j=1 for a patient with disease.[206] In the first level of the model the precision of the estimates of the proportion test positive according the numbers diseased and not diseased in each study is taken into account. In the second level the pattern of estimates of accuracy is modelled using the following non-linear regression equation:

$$\text{logit}(\pi_{ij}) = \left(\theta_i + \alpha_i \, dis_{ij}\right)\exp\left(-\beta \, dis_{ij}\right)$$

where π_{ij} is the proportion test positive. The model yields parameter estimates $\widehat{\theta}$ (the mean of the implicit threshold), $\widehat{\alpha}$ (the mean log diagnostic odds ratio) and $\widehat{\beta}$ which allows for asymmetry in the underlying SROC curve by allowing the logDOR to vary with implicit threshold. If the threshold and log diagnostic odds ratio parameters are fitted as random effects, associated variances are also estimated assuming Normal distributions of θ_i and α_i.

A summary ROC curve can be constructed by computing values of sensitivity across the range of specificities using the following equation:

$$sensitivity = \frac{1}{1 + \exp\left(-\widehat{\alpha}\exp(-0.5\widehat{\beta}) - \ln\left(\dfrac{1 - specificity}{specificity}\right)\exp(-\widehat{\beta})\right)}$$

Appendix 3 Specification of bivariate normal model

The BVN model as expressed by Reitsma and colleagues[90] considers individual studies (I = 1, ..., k) with sensitivity ($p_{A,i}$) determined in N_A individuals with the target disorder and specificity ($p_{B,i}$) determined in N_B individuals who do not have the target disorder.

The first level of the model incorporates the precision with which sensitivity and specificity have been measured in each study.

In the second level a random effects approach is used, assuming that the true logit sensitivities for the individual studies are normally distributed around some common mean value $\mu_{A,i}$ with a between study variability of σ^2_A. The same random effect assumption is made for true logit specificities, with mean value $\mu_{B,i}$ and between study variability of σ^2_B. The potential correlation σ_{AB} between sensitivity and specificity is explicitly included into the analysis.

Combining two normal distributions that can be correlated leads to the following bivariate normal model:

$$\begin{pmatrix} \mu_{A,i} \\ \mu_{B,i} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \sum\right) \text{ with } \sum = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix}$$

The model yields parameter estimates for:

- mean logit sensitivity (μ_A), mean specificity (μ_B) and their 95% confidence intervals
- estimates of between study variability in sensitivity (σ^2_A) and specificity (σ^2_B), and
- an estimate of the covariance between sensitivity and specificity (σ_{AB})

Appendix 4 Spectrum-related items used in reviews undertaking quality assessment (n=131)

Review	Quality assessment tool	Complete /appropriate spectrum	Sample described	Setting described	Participant sampling	Subgroup analysis	Spectrum related item(s)
Adams, 1998[205]	Haynes 1995 Author's own	√	√	√	-	-	Grade A: Studies with *broad generalisability* to a variety of patients and no significant flaws in research methods: sample size >70, patients drawn from *clinical relevant sample* with *clinical symptoms completely described*, diagnoses defined by an appropriate reference standard, PET studies technically of high quality and evaluated independently of references diagnosis Grade B: Studies with *narrower spectrum of generalisability*, with only a few flaws that are well described: > 70 patients, more limited spectrum of patients, free of other method flaws that promote interaction between test results and disease determination, prospective study Grade C: Studies with several method flaws: small sample size, incomplete reporting, retrospective studies of diagnostic accuracy Grade D: no credible reference standard, test results and determination of final diagnosis not independent, source of patient cohort could not be determined or influenced by test result, opinions without substantiating data
Anand, 1998[83]	Holleman 1995	-	-	-	-	-	Not considered
Attia, 1999[206]	Authors' own	-	-	-	-	-	Not considered
Bader, 2001[207]	Authors' own	-	-	-	-	-	Not considered
Badgett, 1997[82]	Modified Holleman 1995	-	-	-	-	-	Not considered
Badgett, 1996[90]	Authors' own	√*		-	-	-	Did the population include a *continuous spectrum* of patients that *included normal* patients?
Bafounta, 2001[208]	Irwig 1994; Cochrane 1996	√*	√	-	-	-	*Spectrum* of included patients *well described* (with a spectrum of melanoma lesions *and lesions commonly confused* with melanoma)
Balk, 2001[180,209]	Authors own	√*	-	√	-	-	For *generalisability* assessment categories of populations/settings: I - included all pts with *signs/symptoms suggestive of* ACI, such as chest pain, shortness of breath, jaw pain, acute pulmonary edema etc II - chest pain as inclusion criteria III - included pts with chest pain but excluded those with clinical of ECG findings diagnostic of AMI IV - all pts hospitalised or that used additional criteria to enrol highly selected subpopulations or retrospective studies Setting described
Bastian, 1998[210]	Holleman 1995 as inclusion criteria	-	-	-	-	-	Not considered
Bastian, 1997[211]	Holleman 1995	-	-	-	-	-	Not considered
Becker, 1996[212]	Becker 1989	√*	-	-	-	√	The subjects studied *represented the complete spectrum* of patients with suspected DVT or PE, including those with and *without disease.* *Results of tests should be stratified* by the extent and severity of DVT or PE. The reproducibility of the D-dimer results should be evaluated in a *setting where the test is likely to be used.*
Bell, 1998[213]	Cochrane 1996	√	√	-	-	-	*Description* of the study with respect to major risk factors, which may affect the *generalisability* of the results to other populations
Berger, 2000[106]	Authors own	√		-	√		Setting - studies divided into those in which *a (random) sample of the popl was invited* for screening and those in which patients *were referred for* gallbladder investigation because of abdominal symptoms Spectrum - patients in hospital-based studies were classified as 'mild disease' if so described or if *elective referrals*, and as 'serious disease' if pts were so described or they were *emergency referrals or hospitalised pts.* The definition 'no disease' was applied to all studies based in general population Patient characteristics
Berry, 1999[214]	Authors own[215]	-	√	-	-	-	Are the study group's clinical, pathological and co-morbid details described? i.e. severity and chronicity of symptoms, sex ratio, age range and mean age, type and location of disease for those receiving gold standard, presence/absence of co-morbid

Review	Quality assessment tool	Complete /appropriate spectrum	Sample described	Setting described	Participant sampling	Subgroup analysis	Spectrum related item(s)
							conditions
Berry, 2002[96]	Authors own	√	√	-	-	-	Various patient selection biases considered, including referral bias, patient filtering bias, *patient cohort bias* (description of clinical, pathologic and co-morbid details)
Blakely, 1995[216]	Modified Sackett 1991	√	-	√	-	-	*Site of patient enrolment* (radiology vs surgery) Disease spectrum of patients enrolled
Bonis, 1997[217]	Mulrow 1989	-	-	-	-	-	Not considered
Bradley, 1998[218]	Authors own	-	-	-	-	-	Not considered
Buntinx, 1997[108]	Authors own	-	√	√	-	-	*Setting* *Age distribution and sex ratio*
Cabana, 1995[219]	Authors own (as inclusion criteria)	-	-	-	-	-	Not considered
Campens, 1997[220]	Authors own	-	-	√	√	-	Details concerning *patient selection* and setting were reported for each study
Cher, 2001[98]	Authors own	-	-	-	-	-	Not considered
Chien, 1997[99]	Author's own	-	-	-	-	-	Not considered
Conde-Agudelo, 1998[221]	Authors' own	-	-	-	-	-	Not considered
Cuzick, 1999[222]	Authors own (not a formal assessment)	√	-	-	-	-	Selection of control groups
Da Silva, 1995[223]	Authors own	√*	-	-	√	-	ideal popl: *consecutive* infants enrolled *prospectively* who presented with *clinical signs suggestive* of sepsis admitted to a neonatal intensive care unit 2nd best: consecutive infants who had in the past been evaluated for sepsis in a neonatal intensive care unit , enrolled from hospital records worst: nonconsecutive
De Bernardinis, 1999[94]	Authors' own	-	-	-	-	-	Not considered
de Bruyn, 2001[93]	Authors own	√	-	-	-	-	Appropriate spectrum (not further defined)
de Vries, 1996[91]	Authors own	-	-	-	-	-	Not considered
Deville, 2000[123]	Cochrane 1996	√*	√	√	-	-	*Spectrum of disease and non-disease given*; Enough information to identify *setting*. Duration of illness before diagnosis; Previous tests/referral filter; *Comorbid conditions* in diseased and nondiseased.
Devous, 1998[224]	Authors own	-	-	-	-	-	Not considered
Dinnes, 2001[225]	Authors own	-	-	-	-	-	Not considered
Divakaran, 2001[226]	Existing checklists: Dunn 1995, Guyatt 1992, Cochrane 1996	-	-	-	-	-	Not considered
Ebell, 2000[227]	Authors own (as inclusion criteria)	-	-	-	-	-	Not considered
Fahey, 1995[125]	Authors own	-	-	√	-	-	*Clinical use*: follow-up (i.e. prompted by previous Pap test result) vs. screening;
Fiellin, 2000[228]	Authors own	-	√	-	-	√	*Adequate description* of spectrum if included information on: demographics (age and sex distirbution); comorbidity (medical and psychiatric); and eligiblity criteria and number of eligible and screened subjects (i.e. participation rate) *Analysis of pertinent clinical subgroups* - as test accuracy can vary according to clinical or dempographic characteristcis
Fischer, 2001[229]	Adams 1996	√*	√	-	-	-	Patients drawn from a *clinically relevant sample* (not selected to include only severe disease) with clinical *symptoms fully*

Review	Quality assessment tool	Complete /appropriate spectrum	Sample described	Setting described	Participant sampling	Subgroup analysis	Spectrum related item(s)
							described.
Fleischmann, 1998[109]	Authors own (ref Begg 1998)	-	-	-	-	-	Not considered
Fowlie, 1998[230]	Authors own	√	-	-	-	-	Did the population studied include an *appropriate spectrum* of babies to whom the test would be applied in practice?
Garzon, 2001[231]	Authors own (as inclusion criteria)	-	-	-	-	-	Not considered
Gianrossi, 1990[113]	Wachter 1988	-	-	-	-	-	Not considered
Gottlieb, 1999[232]	Authors own (as inclusion criteria)	-	-	-	-	-	Not considered
Gould, 2001[107]	Adapted Kent 1992	-	-	-	-	-	Not considered
Gronseth, 2000[233]	Authors own	-	-	√	-	-	Setting
Hallan, 1997[234]	Authors own	-	√	√	-	-	Prevalence and degree of disease, clinical setting
Heffner, 1995[235]	Irwig 1994	√	√	-	-	-	*Generalisability*: assessed studies for *reporting of characteristics* including patient age, presenting complaints, pneumonic pathogen, comorbid lung disease, comorbid underlying conditions, drug therapy resulting in immunosuppression, duration and severity of illness, and blood values for pH, glucose, and LDH obtained concomitant with pleural fluid values.
Heffner, 1997[236]	Authors own	√	√	-	√	-	Assessment of *generalisability* was assessed by noting whether *sufficient clinical information*, such as age, gender, and underlying medical conditions was provided to allow the reader to determine if the *study results could be generalised* to their population Cohort assembly (presence of an adequate spectrum of patients and the detail by which the assembly of the cohort was described)
Hider, 1999[237]	New Zealand National Health Committee	-	-	-	-	-	Not considered
Hobbs, 1999[238]	Reid 1995	-	-	-	-	-	Not considered
Hoffman, 2000[116]	Authors own	√	-	-	-	-	*Spectrum* of study patients: judged on age, race, sex, digital rectal exam findings, urinary symptoms, presence of benign prostatic hyperplasia, and cancer stage, plus explicit mention of eligibility criteria
Hooft, 2001[239]	Cochrane 1996	√	-	-	-	-	*Appropriate* clinical setting and patients spectrum
Hrung, 1999[240]	Author's own	-	-	-	-	-	Not considered
Huicho, 2002[241]	Modified Mulrow 1989	√	√	√	-	-	Were the subjects symptomatic? If so were they assessed? Did the author *describe the age, sex and symptoms* of their cohorts or at least *state cohort was 'unselected'?* Did investigators assemble *population-based cohorts* or did they assemble their cohorts from patients who had been referred for a urine culture? What was the age of the cohort? What % of cohort was male? *Where did the examinations take place*: hospital, clinic, in the field or at laboratory?
Huicho, 1996[242]	Mulrow 1989	√	√	√	-	-	Were pts symptomatic? If so, were they assessed? Did authors *describe age, sex, and symptomatology* of their cohorts or did they at least state that their cohorts were *'unselected'?* Did investigators assemble *popl-based cohorts* or did they assemble their cohorts from *pts who had been referred* for a stool culture (fecal microbiologic study or other test) What was the age of the cohort; what % were male; what other health or comorbid conditions characterised the cohort? *Where* did the examinations take place? Hospital, clinic, in the field or at laboratory?

Review	Quality assessment tool	Complete /appropriate spectrum	Sample described	Setting described	Participant sampling	Subgroup analysis	Spectrum related item(s)
Ioannidis, 2001[209,243]	Authors own? (reference Irwig 1994)	√*	-	√	-	-	For *generalisability* assessment categories of populations/settings: I - included *all pts with signs/symptoms suggestive* of ACI, such as chest pain, shortness of breath, jaw pain, acute pulmonary edema etc II - chest pain as inclusion criteria III - included pts with chest pain but excluded those with clinical of ECG findings diagnostic of AMI IV - all pts hospitalised or that used additional criteria to enrol highly selected subpopulations or retrospective studies Setting described
Ioannidis, 2001A[189,209]	As above	√*	-	√	-	-	*Setting* described Plus separate 4-category scale to group populations and settings (see below). For generalisability assessment categories of populations/settings: I - included *all pts with signs/symptoms suggestive* of ACI, such as chest pain, shortness of breath, jaw pain, acute pulmonary edema etc II - chest pain as inclusion criteria III - included pts with chest pain but excluded those with clinical of ECG findings diagnostic of AMI IV - all pts hospitalised or that used additional criteria to enrol highly selected subpopulations or retrospective studies Also considered differences in prevalence of ACI or AMI as way to determine baseline risk in the 4 population categories
Kearon, 1998[101]	Authors own - used as inclusion criteria	-	-	-	-	-	Not considered
Kim, 2001[121]	Irwig 1994	-	-	-	-	-	Not considered
Kinkel, 1999[244]	Authors own	-	-	-	-	-	Not considered
Kittler, 2002[104]	Authors own	-	-	-	-	-	Not considered
Klompas, 2002[245]	Refer to previous articles in series (JAMA) but give no reference	-	-	-	-	-	Not considered
Koelemay, 2001[190]	Authors own	-	√	-	-	-	Clear definition of study population
Koelemay, 1996[246]	Authors own	-	-	-	-	-	Not considered
Koumans, 1998[85]	Authors own	√	√	-	√	-	4. *Clinical description of sample* (whether description of source and characteristics of study sample was complete) 5. Assembly of population (*adequate spectrum*; sufficient *description of assembly*; independent application of reference test)
Kowalski, 2001[102]	Authors own	-	-	-	-	-	Not considered
Kwok, 1999[95]	Authors own	-	√	-	-	-	Clear definition of selection criteria and *presentation of participant characteristics*
Lacasse, 1999[247]	Authors' own	-	-	-	-	-	Spectrum not included in VA, but authors stated that by only including studies with consecutive pts also ascertained that the pt sample included an appropriate spectrum of pts.
Lau, 1999[248]	Authors own	-	-	-	-	-	Not considered
Law, 1998[110]	Authors own	-	√	-	-	-	Subject: *score 1 point for description of* each of age (mean, range or SD), gender and ethnicity or socio-economic status Sample from general population
Lederle, 1999[249]	Holleman 1995	√	-	-	-	-	Patients *suspected of* having the target condition
Liedberg, 1996[250]	Authors own	-	-	-	-	-	Not considered
Lindbaek, 2002[251]	Cochrane 1996	-	-	-	-	-	Not considered
Littenberg, 1995[252]	Authors own	√	-	-	-	-	*Patient sources* examined to assess referral bias Inclusion/exclusion criteria examined to assess *generalisability*

Review	Quality assessment tool	Complete /appropriate spectrum	Sample described	Setting described	Participant sampling	Subgroup analysis	Spectrum related item(s)
Loy, 1996[118]	Author's own	-	-	-	-	-	Not considered
Lysakowski, 2001[191]	Adapted Lijmer 1998	√	-	-	-	-	*Homogeneity of study population* (same vs different pathologies)
MacKenzie, 1996[253]	Authors own	-	-	-	-	-	Not considered
Mango, 1998[254]	Authors own	-	-	-	-	-	Not considered
Mayer, 1997[255]	Sackett 1991	√	-	√	-	-	*Spectrum of pigmented skin lesions, study setting, patient demographics*, prevalence of melanoma, proportion of pigmented skin lesions in which no dermatoscopic diagnosis could be made.
McCrory, 1999[122]	Authors own	-	√	-	√	-	*Description of disease spectrum* Avoidance of bias in *sample selection*
McGee, 1999[256]	Authors own?	√	-	-	-	-	*Pts suspected of* having volaemia
McNaughton Collins, 2000[257]	Reid 1995	-	-	-	-	-	Not considered
Metlay, 1997[258]	Authors own	√	-	-	-	-	Patients *suspected of* having CAP (Level I)
Mol, 1997[259]	Authors own	-	-	-	-	-	Not considered
Mol, 1998[260]	Author's own	-	-	-	-	-	Not considered
Mol, 1998[120]	Authors' own	-	-	-	-	-	Not considered
Mol, 1999[261]	Authors own	-	-	-	-	-	Not considered
MSAC, 1999[262]	Authors own	-	-	-	-	-	Not considered
Mullins, 2000[263]	Authors own	-	√	-	√	√	Sufficient *description of selection process*; sufficient *description of patients*; sufficient description of non-enrolled patients; description of extent of disease such that *results could be stratified by location or severity*; reporting of *non-PE diagnoses*
Muris, 1994[264]	Authors own	√	-	√	-	-	The study is done in a *setting relevant* for a general practitioner There is a *sufficient variation (spectrum)* in quantity and severity of diseases Intra-observer variability of recorded symtoms measured (relates to test variation?)
Muris, 1992[265]	Authors own	-	-	√	-	-	Setting relevant to GP
Mustafa, 2002[266]	Jaeschke, 1994	√	-	-	-	-	*Broad spectrum* included
Nallamothu, 2001[193]	Authors own	-	-	-	-	-	Not considered
Nanda, 2000[127]	Authors own	-	-	-	-	-	Not considered
Nuovo, 1997[267]	Authors own	√*	-	√	√	-	Did patient sample include an *appropriate spectrum* of mild and severe, treated and untreated disease in addition to patients with different but *commonly confused disorders*? Was *study setting and filter through which patients passed* adequately described? Are results applicable to primary care patients?
Oosterhuis, 2000[92]	Authors own	-	-	√	-	-	No selection bias (e.g. where B12 and/or MCV ordered as part of regular treatment)
Owens, 1996[128]	Authors own	-	√	-	√	-	*Adequacy of description* of clinical population *Appropriateness of assembly* of study sample
Owens, 1996[87]	Authors own	√	√	-	-	-	Clinical description - was the *study population described adequately*? Cohort assembly - was the *spectrum of patients adequate*
Pasternak, 2001[268]	Authors own	-	√	-	-	-	*Comparability of controls*
Patel, 2000[269]	Irwig 1994	-	-	-	-	-	Not considered

Review	Quality assessment tool	Complete /appropriate spectrum	Sample described	Setting described	Participant sampling	Subgroup analysis	Spectrum related item(s)
Paul, 2000[100]	Authors own	-	-	-	-	-	Not considered
Pearl, 1996[270]	Kent 1992	√			-	-	*Representative sample* without selection bias
Rao, 1995[271]	Authors own	√*	-	-	-	-	*More than 50% of controls had actual diagnoses* (e.g. other vasculitides, pulmonary renal syndromes etc) as opposed to being healthy controls
Rao, 1999[272]	Jaeschke 1994	√	-	-	-	-	Did the pt sample include an *appropriate spectrum* of pts to whom the diagnostic test will be applied in clinical practice?
Rathbun, 2000[273]	Jaeschke 1994	√*	√	-	√	-	Does study include *a consecutive series* of patients *with suspected* PE? Does the study examine a *broad spectrum* of patients (including patients *with and those without* PE) and a broad *spectrum of patient characteristics* (such as: age; sex; high, intermediate or low clinical suspicion of PE; comorbid conditions that may confuse the diagnosis; and size or PE on angiography)?
Reed, 1996[274]	Author's own	-	-	-	-	-	Not considered
Ross, 1999[275]	Authors own;derived from Irwig 1994 and Flemons 1996	√*	√	√	√	-	Patients *both with and without disease*? (1 point); inclusion criteria reported? (1 point); patient *selection process described*? (1 point); statement of *where patients were recruited from*? (1 point); *wide spectrum* of patient's SA severity? (1 point); *patient characteristics described*? (1 point); patients eligible but not enrolled, described? (1 point);
Scheid, 2001[276]	McKibbon 1995	-	-	-	-	-	Not considered
Schwimmer, 2000[277]	Authors own (based on US medical payer source criteria)	√	√	√	-	-	Selection/exclusion criteria presented *Patient characteristics*: age range gender *Institution characteristics:* special expertise How patients directed to PET (referral pattern)
Scouller, 2000[112]	Authors own (based on several references)	√	-	√	√	√	Case-control design avoided Recruitment of *consecutive* patients *Subject selection method* recorded *Gender and/or age comparability* stated between those positive and negative on reference standard Spectrum of race stated *Stratification of results* by gender, race or age Also recorded *recruitment site of study* (clinical or community setting, pts with or without known alcohol problems), and classified the spectrum of alcohol intake (not part of VA)
Smith-Bindman, 2001[278]	Authors own (as inclusion criteria)	-	-	-	-	-	Not considered
Solomon, 2001[279]	Holleman 1995	√	-	-	-	-	*Relevance* of the patient
Sonnad, 2001[280]	Single criterion used	-	-	-	-	-	Not considered
Spencer-Green, 1997[281]	Mulrow et al	√*	-	√	-	-	*Appropriate study popl*: i.e. included a cohort of pts with scleroderma or systemic sclerosis *Source of pts described*: to satisfy this criterion, papers had to identify from where their patient populations and sera were derived *Wide spectrum* of case patients included: required that a description of a spectrum of clinical or laboratory features of the case patients be included. The description of some evaluation for the presence or absence systemic involvement satisfied this criterion *Inclusion of comorbid disease*: papers satisfying this criterion used as non-SSc controls patients with other connective tissue disorders including systemic lupus erythematosus, rheumatoid arthritis, Sjogrens syndrome, dermatomyositis, or linear scleroderma, or primary or secondary Raynaud's phenomena. *Comorbid diseases included* in case group: papers met this criterion if in their description of SSc patients, the authors did not specifically exclude comorbid diseases
Stengel, 2001[282]	Authors own plus CEBM	-	-	-	-	-	Not considered

Review	Quality assessment tool	Complete /appropriate spectrum	Sample described	Setting described	Participant sampling	Subgroup analysis	Spectrum related item(s)
Swart, 1995[283]	Authors own	-	-	√	-	-	Disease prevalence (< or >= 35%) *Setting* (academic or non-academic)
Taylor-Weetman, 2002[284]	Authors own	√	-	-	-	-	Is the *population representative* of general UK patients?
van Beek, 2001[285]	Authors own (as inclusion criteria)	-	-	-	-	-	Not considered
van den Hoogen, 1995[286]	Author's own	√*	√	-	√	-	Clinical description: 10 points if *sufficiently detailed clinical description* of subjects; 5 points incomplete clinical description; 0 points no description other than "low back pain" Study population: 10 points for *prospective enrolment*, explicit inclusion/exclusion criteria and adequate patient spectrum; 5 points retrospective design without inclusion criteria or with limited patient spectrum; 0 points other studies. Studies received extra 10 points if from general practice or general population Study population: studies in which *both diseased and nondiseased subjects* participated scored 10 points others scored 0
van der Wurff, 2000[287]	Authors own[287]	-	√	-	-	-	*Description of study population* *Description of inclusion and exclusion criteria*
Varonen, 2000[198]	Cochrane	-	-	-	-	-	Not considered
Vasbinder, 2001[288]	Authors own	√	-	-	-	-	Inclusion criteria: limited to studies where *reason for referral was clinical suspicion* of renovascular hypertension, i.e. appropriate spectrum
Visser, 2000[119]	Authors own plus Kent 1992 (latter not described)	-	-	-	-	-	Not considered
Vroomem, 1999[289]	Sackett 1991	√*	√	-	√	-	Patient description: both *demographic and clinical characteristics should be described* Study population: *prospective* design, adequate description of selection criteria and *adequate patient spectrum* (variance in disease severity and comorbidity such that the popl was representative of a clinical patient population) *Diseased and nondiseased included*
Watson, 2002[290]	Irwig 1994	-	-	-	-	-	Not considered
Wells, 1995[291]	Authors own	-	-	-	-	-	Not considered
White, 2000[292]	Authors own (as secondary inclusion criteria)	-	-	-	-	-	Not considered
Whited, 1998[293]	Holleman, 1995 (as inclusion criteria)	-	-	-	-	-	Not considered
Whitsel, 2000[114]	Authors own	-	-	-	-	-	Not considered
Wiese, 2000[129]	Irwig 1994	√	-	-	-	-	*Appropriate spectrum* included (not stated to be part of quality assessment, but was reported in Results)
Wijnberger, 2001[294]	Authors own	-	√	-	-	-	*Scored clinical criteria*: min/max gestational age, inclusion of multiple pregnancies, diabetic pregnancies, women with ruptured membranes, and use of corticosteroids
Williams, 2002[295]	Authors own	-	-	-	-	-	Not considered

Appendix 5 Summary details per review of spectrum-related heterogeneity investigations

	Spectrum-related variables investigated									Method of investigating heterogeneity		Statistically significant effect from		
Review	Age	Sex	Prevalence	Clinical indication/ eligibility	Disease severity/ stage	Symptoms/ risk status	Setting/ source of pts	Sampling/study design	Comorb-idities		Results clearly reported	Spectrum	Test	Quality
Badgett, 1997[82]							√			Pooled Se/Sp	Y	Y	n/i	n/i
Bafounta, 2001[208]			√				√			Pooled Se/Sp	N	N	N	N
Balk, 2001[180,209]			√							Not clear; covariate to regr?	N	N	N	n/i
Bastian, 1998[210]			√	√						Pooled Se/Sp	P	N	N	n/i
Bafounta, 2001[208]							√			Pooled Se/ES	Y	Y	n/i	n/i
Berger, 2000[106]					√		√			Covariate to logistic regr	Y	Y	n/i	Y
Berry, 2002[96]				√						Covariate to SROC regr model	N	N	N	N
Buntinx, 1997[108]	√					√				Pooled Se	Y	Y	N	n/i
Carlson, 1994[103]					√					Pooled Se	Y	Y	n/i	n/i
Cher, 2001[98]			√			√		√		Covariate to SROC regr model	Y	Y	Y	Y
Chien, 1997[99]						√				Pooled LRs	P	Y	n/i	N
Conde-Agudelo, 1998[221]	√									Median Se/FPR	Y	Y	Y	n/i
D'Arcy, 2000[84]	√	√	√							Study exclusion in sens analysis	N	N	N	N
de Bruyn, 2001[93]				√						Covariates to SROC regr model	N	N	N	N
de Vries, 1996[91]			√	√						Covariates to SROC regr model	N	N	Y	N
De Bernardinis, 1999[94]				√		√				Pooled ES	Y	Y	Y	Y
Deville et el., 2000[123]				√						Covariates to SROC regr model	Y	N	Y	Y
Di Fabio, 1996[124]				√						ANOVA	Y	Y	N	N
Dijkhuizen, 2000[100]	√					√				Pooled Se/Sp	P	Y	Y	N
Fahey, 1995[125]							√			Pooled Se/Sp; multiple linear regr	Y	N	N	N
Faron, 1998[296]						√				Pooled LRs	Y	Y	Y	n/i
Fleischmann, 1998[109]	√	√		√			√			Covariates to SROC regr model	Y	Y	Y	N
Gianrossi, 1990[113]	√	√	√	√						Multiple linear regr (Se/Sp as dependent variable)	P	Y	Y	Y
Gould, 2001[107]					√					Separate SROC models (rep as log ORs)	P	N	N	Y
Hallan, 1997[234]							√			Separate SROC models	Y	Y	n/i	n/i
Heffner, 1995[235]							√			IPD	Y	A	n/i	n/i
Hoffman, 2000[116]				√			√	√		Median log DOR in subgroups	Y	N	N	N
Hofman, 2000[297]	√			√						Correl of log DOR with covariates	Y	Y	n/i	N
Huicho, 2002[241]	√									Covariates added to multiple regr model	N	Y	Y	N
Hurley, 2000[126]				√						Separate SROC models	Y	N	N	N
Ioannidis, 2001[209,243]							√			Pooled DOR/AUC	Y	N	n/i	n/i

	Spectrum-related variables investigated									Method of investigating heterogeneity		Statistically significant effect from		
Review	Age	Sex	Prevalence	Clinical indication/ eligibility	Disease severity/ stage	Symptoms/ risk status	Setting/ source of pts	Sampling/study design	Comorb-idities		Results clearly reported	Spectrum	Test	Quality
Ioannidis, 2001A[189,209]							√			Pooled Se/Sp/DOR	Y	N	n/i	n/i
Kearon, 1998[101]						√				Pooled Se/Sp	Y	Y	n/i	n/i
Kim, 2001[121]	√		√	√						Covariates to regr model	P	Y	N	Y
Kinkel, 2000[117]			√	√			√			Covariates to SROC regr model	P	Y	N	Y
Kinkel, 1999[244]					√					Covariates to regr model	N	N	N	N
Kittler, 2002[104]			√				√			Covariates to SROC regr model	P	Y	Y	N
Koelemay, 2001[190]			√							Covariates to SROC regr model	N	N	Y	N
Koumans, 1998[85]		√	√							Pooled Se/Sp	P	N	N	n/i
Kowalski, 2001[102]						√				Covariates to GEE regr model	P	Y	Y	Y
Kwok, 1999[95]		√	√	√						Covariates to regr model	N	Y	N	N
Lacasse, 1999[247]								√		Pooled Se/Sp	N	N	N	N
Law, 1998[110]	√						√			Correl with LRs	N	Y	n/i	Y
Leitich, 1999[97]				√		√				Pooled Se/Sp	Y	A	A	n/i
Loy, 1996[118]	√		√	√			√			Covariates to SROC regr model	P	N	N	N
McCrory, 1999[122]			√							Covariates to log regr	P	Y	n/i	Y
Mol, 1998[120]					√			√		Covariates to log regr	P	Y	n/i	n/i
Nallamothu, 2001[193]	√	√	√							Covariates to SROC regr model	P	N	N	N
Nanda, 2000[127]						√				Mean Se/Sp	Y	A	A	A
Oosterhuis, 2000[92]				√						Pooled Se	Y	Y	Y	Y
Orr, 1995[111]	√	√	√							Covariates added to linear regr on Se/Sp	N	N	N	N
Owens, 1996[128]						√				Pooled log OR	Y	N	Y	Y
Owens, 1996[87]	√									Pooled log OR	Y	Y	Y	N
Peters, 1996[115]							√			IPD	P	Y	n/i	n/i
Rao, 1995[271]					√					Pooled Se/Sp	Y	Y	n/i	N
Reed, 1996[274]				√						Covariates to SROC regr model	N	N	N	N
Revah, 1998[298]						√				Pooled Se/Sp	Y	Y	n/i	n/i
Scheidler, 1997[76]					√					Separate SROC models	Y	N	N	n/i
Scouller, 2000[112]	√	√		√						Covariates to SROC regr model	N	N	Y	n/i
Smith-Bindman, 1998[299]				√						Pooled Se/Sp; Separate SROC models	P	Y	N	N
Smith-Bindman, 2001[278]								√		Pooled Se/Sp	Y	N	N	N
Spencer-Green, 1997[281]				√						Pooled Se/Sp	P	Y	N	N
Stengel, 2001[282]	√									Separate SROC models	P	N	N	N
Swart, 1995[283]			√				√			Pooled Se/Sp	P	N	N	Y
Tugwell, 1997[105]					√					Pooled Se/Sp	Y	Y	n/i	n/i

	Spectrum-related variables investigated									Method of investigating heterogeneity		Statistically significant effect from		
Review	Age	Sex	Prevalence	Clinical indication/ eligibility	Disease severity/ stage	Symptoms/ risk status	Setting/ source of pts	Sampling/study design	Comorb-idities		Results clearly reported	Spectrum	Test	Quality
Visser, 2000[119]	√		√				√	√		Covariates to SROC regr model	P	Y	Y	Y
White, 2000[292]					√					Covariates to SROC regr model?	P	Y	N	n/i
Whitsel, 2000[114]	√	√	√	√				√		Covariates to SROC regr model	Y	Y	N	N
Wiese, 2000[129]			√							Pooled Se/Sp; correl of Se with covariates	Y	Y	Y	Y

Se – sensitivity; Sp – specificity; Es – effect size; regr – regression; LR – likelihood ratio; FPR; false positive rate; OR – odds ratio; IPD – individual patient data; correl – correlation; DOR – diagnostic odds ratio; AUC – area under the curve; Y – yes; N – no; P – partially; n/i – not investigated

Appendix 6 Reviews using advanced methods of meta-analysis: Methods

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. accuracy studies	No. of patients
Bipat, 2003[132]	cervical cancer staging according to: parametrial invasion bladder invasion rectal invasion lymph node involvement	CT MRI	histopathology	MEDLINE and EMBASE Jan 1985 to May 2002	none	Authors' own	MRI 38 CT 11 Both 8	Not reported
Bipat, 2004[133]	rectal cancer staging	endoluminal ultrasound CT MRI	histopathology	MEDLINE, EMBASE, Cochrane, CANCERLIT Jan 1985-Dec 2002	English only	Authors' own	90 studies; 299 datasets	
Bipat, 2005[137]	pancreatic adenocarcinoma	Ultrasound CT MRI	histopathology surgical findings follow-up	MEDLINE, EMBASE, Cochrane, CANCERLIT	English German	Authors' own	For diagnosis Helical CT 23 Conventional CT 20 MRI 11 Ultrasound 14 For resectability Helical CT 32 Conventional CT 12 MRI 7 Ultrasound 6	For diagnosis helical CT 959 conventional CT 1473 MRI 583 ultrasound 2909 For resectability helical CT 1823 conventional CT 1467 MRI 516 ultrasound 1233
Bipat, 2005a[134]	colorectal liver metastases	CT MRI PET	histopathology	MEDLINE EMBASE Jan 1990 to Dec 2003	English German French	QUADAS	61 Nonhelical CT 58 Helical CT 53 1.0T MRI 34 1.5T MRI 102 PET 26	3187 Nonhelical CT 1915 Helical CT 621 1.0T MRI 173 1.5T MRI 391 PET 1058

Glas, 2003[80]	primary bladder cancer	cytology and urinary based tumour markers	cytoscopy	MEDLINE EMBASE 1990-Nov 2001	English German	Authors' own	42 BTA 6 BTA stat 8 BTA TRAK 5 NMP22 14 Telomerase 10 FDP 2	BTA 715 BTA stat 1300 BTA TRAK 829 NMP22 2290 Telomerase 855 FDP 157
Halligan, 2005[141]	detection of colorectal polyps	CT colonography	colonoscopy	MEDLINE 1994 and 2003	No language restrictions	STARD and QUADAS	Category 1: 7 Category 2: 7	Category 1: 2610 Category 2: 1834
Koelemay, 2004[138]	symptomatic carotid artery disease	CTA	arteriography/intra-arterial digital subtraction angiography	PubMed, MEDLINE, PREMEDLINE, EMBASE, and CINAHL 1990 to July 2003	None	Authors' own	28	864
Kwee, 2007[142]	follow-up of intracranial aneurysms treated with Guglielmi detachable coils	MRA	digital subtraction angiography	PubMed/MEDLINE and Embase to Jan 2007	English, German, French	Adapted QUADAS	16	616
Shaheen, 2007[130]	prediction of chronic hepatitis C virus-related fibrosis	aspartate aminotransferase-to-platelet ratio index (APRI)	liver biopsy	Medline, EMBASE, and Cochrane Library (01/1997-12/2006)	No language restrictions	QUADAS	22	4266
Thangaratinam, 2007[131]	screening tool for congenital heart disease in asymptomatic newborns	pulse oximetry	echocardiography	MEDLINE, EMBASE, Cochrane Library, MEDION	No language restrictions	Authors' own	8	35 960 newborns
Van Westreenen, 2004[135]	preoperative staging of patients with esophageal cancer	FDG-PET	pathology or surgery	PubMed, Embase, and Cochrane to June 2003	No language restrictions	Cochrane Methods Working Group checklist	N-stage 12 M-stage 11	N-stage 421 M-stage 452
Whiting, 2006[139]	early diagnosis of multiple sclerosis in patients presenting with suspected disease	MRI	clinically defined MS	12 databases from inception until September or November 2004.	No language restrictions	QUADAS	40 (Most analyses restricted to 15 cohort studies)	

Williams, 2007[140]	diagnosis of renal artery stenosis in patients with hypertension	renal duplex sonography - peak systolic velocity, renal-aortic ratio, acceleration time, acceleration index	intraarterial angiography	MEDLINE and EMBASE 1966-2005	No language restrictions	Authors' own	88	9974 arteries

Appendix 7 Reviews using advanced methods of meta-analysis: Synthesis methods

Study	Method of study synthesis	Heterogeneity investigation	Data presentation – overall analysis	Data presentation – heterogeneity investigations
Bipat, 2003[132]	BVN (reference Van Houwelingen, 1993 and 2002)	Yes, but only possible for MRI for parametrial invasion and lymph node involvement Covariates investigated for sensitivity and specificity: sample size (>50 vs ≤50) publication period (1985–1991 vs 1992–1997 vs 1998–2002) methodological shortcomings (added simultaneously): patient selection, unblinded interpretation of test results, verification bias, and retrospective collection of data These criteria were adjusted for by adding covariates simultaneously to the bivariate approach. Also subgroup analysis comparing 4 different aspects of MRI techniques	No summary results tables presented. Summary sensitivities and 95%CIs reported in text . Some summary specificities reported in text. ROC plots according to outcome measure, plotted per test Forest plots according to outcome measure, plotted per test	No results tables presented. Actual data not reported in text Forest plots per outcome for MRI by covariate.

Bipat, 2004[133]	BVN (reference Van Houwelingen, 1993 and 2002)	Covariates investigated for sensitivity and specificity: year of publication (continuous variable), sample size (≤50 vs>50), and study design characteristics: patient selection, unblinded interpretation of test results, verification bias, and retrospective collection of data Subsequently developed multivariable regression models with multiple covariates for each stage per test to identify the most important characteristics. Characteristics were retained when $P<0.10$ For each test, a model adjusted for significant variables was obtained using the regression formula logit-sens= alpha + beta(logit-spec) and an SROC curve estimated. Logit sensitivities and specificities were compared across imaging techniques using a final model adjusted for significant covariates. Subgroup analysis for MR and EUS comparing different aspects of test techniques performed	SROC curves based on the final regression model for evaluation of perirectal tissue invasion and lymph node involvement per test presented Summary estimates of sensitivity and specificity with 95%CIs tabulated for four outcomes in staging of rectal cancer.	Regression coefficients for covariates reaching statistical significance in backward regression analysis presented in tabular format. Sensitivity and specificity estimates from subgroup analyses on MRI and endoluminal US for perirectal tissue invasion presented SROC curves for different subgroups in the evaluation of perirectal tissue invasion and for results of individual datasets given.

Bipat, 2005[137]	BVN (reference Van Houwelingen, 1993 and 2002)	Covariates investigated for sensitivity and specificity: year of publication (continuous variable), sample size (≤50 vs>50), department of origin (radiology vs others) and the study design characteristics: patient selection, unblinded interpretation of test results, verification bias, and retrospective collection of data, reporting of study popl, reporting of test, reporting of ref test. Subsequently developed a multivariable regression model to identify the most important characteristics. Characteristics were retained when P<0.10 Logit sensitivities and specificities were then compared across imaging techniques. Subgroup analysis comparing different aspects of test techniques and lesion sizes also performed	Summary sensitivity and specificity per test presented in tabular format for diagnosis and resectability. ROC plots for diagnosis and resectability outcomes Forest plot of overall pooled sensitivity for each test presented.	Regression coefficients for covariates reaching statistical significance in backward regression analysis presented in tabular format.
Bipat, 2005a[134]	BVN (reference Van Houwelingen, 1993 and 2002)	Covariates investigated for sensitivity: Year of publication (1995 or earlier vs later than 1995), sample size (≤50 vs>50) reporting of study popl, reporting of test, reporting of ref test. Study design characteristics also investigated not reported but presumably as for previous reviews Subsequently developed a multivariable regression model to identify the most important characteristics. Characteristics were retained when P<0.10 Logit sensitivities were then compared across imaging techniques. Subgroup analysis comparing different aspects of test techniques and lesion sizes also performed	Summary sensitivity per test presented in tabular format. Forest plot of overall pooled sensitivity for each test presented.	Regression coefficients for covariates reaching statistical significance in backward regression analysis presented in tabular format. Sensitivity for some subgroups according to test technique presented in tabular format for helical CT and MR.

Glas, 2003[80]	BVN (reference Van Houwelingen, 2002	Multivariable analysis performed to explain variation in sensitivity and specificity. Covariates were selected if a specific variable correlated with sensitivity or specificity at P<0.10. Not clearly listed. Appear to have included Study design Type of control group Clear description of study popl Clear description of reference test and marker test Consecutive pt selection Verification by the reference standard Independent assessment of marker test and reference test BCG therapy Hematuria Distribution of tumour differentiation of the diseases Method of urine collection	Summary sensitivity and specificity per test presented in tabular format along with correlation between sensitivity and specificity. ROC plots of studies per test Forest plots of overall pooled sensitivity and specificity for each test presented.	Narrative discussion
Halligan, 2005[141]	HSROC (reference Macaskill 2004)	None. Authors report significant heterogeneity and suggest sources but insufficient studies to investigate.	No summary results tables presented. Summary sensitivities, specificities and 95%CIs derived from HSROC model and reported in text. Forest plots of sensitivities and specificities per study (but not for pooled analysis) reported according to polyp size ROC plots with HSROC curves reported for category 1 and category 2 polyps. No curve derived for category 3 polyps.	NA

Koelemay, 2004[138]	BVN (reference Van Houwelingen, 1993 and 2002)	Covariates were selected and added to model if a specific methodological or clinical variable showed a positive Spearman correlation with the sensitivity or specificity with a probability value <0.1. Included year of publication consecutive enrolment prospective design clear description of technique clear definition of cutoff levels blind assessment of CT angiography and arteriography Patient demographics, symptoms and interval between CTA and arteriography could not be included due to incomplete reporting.	No summary results tables presented. Summary sensitivities and 95%CIs reported in text . Some summary specificities reported in text. ROC plot of studies and pooled sensitivity and specificity with confidence ellipse	No results tables presented. Actual data not reported in text
Kwee, 2007[142]	BVN (reference Reitsma 2005)	None. Authors report insufficient studies to use meta-regression to examine the causes of the heterogeneity.	Data for individual studies and pooled analysis presented in tabular format according to outcome ROC plot of studies and pooled sensitivity and specificity with confidence ellipse per test	NA
Shaheen, 2007[130]	BVN (reference Reitsma 2005) Also conducted Moses regression models to estimate AUC and pooled DORs using DerSimonian and Laird regression model	Does not appear that heterogeneity investigation undertaken within BVN model framework Random effects meta-regression model (referenced to Schmid 2004) to investigate effect on lnDOR of: sample size median age %men methodological quality inclusion of HIV/HCV co-infected patients prevalence of significant fibrosis/cirrhosis location of the study histopathologic scoring system quality of reference standard	Summary sensitivities and specificities at different threshold presented in tabular format for both prediction of significant fibrosis and cirrhosis ROC plots of individual studies and ROC curve Forest plot of DORs Plot of predictive values against prevalence of significant fibrosis No BVN plot presented	No results tables presented. Data reported in text

Thangaratinam, 2007[131]	BVN (reference Reitsma 2005)	None. Authors report significant heterogeneity but insufficient studies to investigate.	Data for individual studies and pooled analysis presented in tabular format for data at commonest threshold (Sao_2<95%). Note 8 datasets listed but not identified according to study ROC plot of studies and pooled sensitivity and specificity with confidence ellipse for data at commonest threshold (Sao_2<95%). Also plot of TPR and FPR for other threshold levels of Sao_2. Studies not identified.	NA
Van Westreenen, 2004[135]	BVN (reference Van Houwelingen, 1993)	None. Exclusion of two outlying studies mentioned in discussion	Data for individual studies and pooled analysis presented in tabular format according to outcome No plots presented	NA
Whiting, 2006[139]	HSROC used to assess the duration of follow-up on overall accuracy and threshold (ref Rutter 2001)	Random effects meta-analysis used to estimate DOR for cohort and case-control studies. HSROC model used to assess effect on accuracy and threshold from duration of follow-up. Separate ROC plots according to MRI criteria	Study data reported in tabular format ROC plot according to cohort and other study designs HSROC curve	ROC plot of case-control and cohort studies ROC plot according to duration of follow-up with single SROC curve ROC plots according to MRI criteria

Williams, 2007[140]	HSROC A function of the estimated model parameters was used to obtain the expected operating point on the SROC cue (Rutter 2001) (reference Rutter 1995, 2001, Macaskill 2004)	Covariates added to the model to assess whether test accuracy, threshold or shape was associated with population or design characteristics: articles reporting no. of pts undergoing both test and reference test articles reporting no. of failed US US method described Exclusion of analyses of occlusion Severity of renal artery stenosis Blinded ref test interpretation Blinded US interpretation Angiographic views specified Accessory arteries included/excluded Prospective design Vessel diameter measures during angiography Consecutive enrolment Clinical spectrum included Hypertension and other features Hypertension with or without chronic renal failure Hypertension moderate or unspecified Hypertension and peripheral vascular disease Transplant recipient Peripheral vascular disease No details stated	ROC plots and HSROC curves per test and according to whether data were paired or unpaired. Estimated sensitivity, 1-specificty, LR+ and LR- presented per test in tabular format	Narrative discussion

Appendix 8 Reviews using advanced methods of meta-analysis: Results

Review	Results of main analysis using advanced methods	Heterogeneity investigations using advanced methods	Author comment on advanced method used
Bipat, 2003[132]	Sensitivity and specificity (%) with 95%CIs for Parametrial invasion Sensitivity: MRI 74 (68–79) CT 55 (44–66), P< 0.01. specificities: reported to be comparable. Lymph node involvement Sensitivity: MRI 60 (52–68) CT 43 (37–57), P < 0.05. specificities: reported to be comparable. Bladder invasion Sensitivity: MRI 75 (66–83) CT 64 (39–82), difference not statistically significant Specificity: MRI 91 (83–95 CT 73 (52–87) for CT (P=0.03 Rectum invasion sensitivity: MRI 71 (53–83) CT 45 (20–73), difference not statistically significant specificities: reported to be comparable.	Text reports that the covariates investigated had no influence on both the sensitivity and specificity estimates	Advantages of BVN: a. more convenient than Moses method b. produces summary estimates of sensitivity and specificity as outcomes, which are more familiar to clinicians. c. both the error of estimation of the sensitivity and specificity values in each study and the heterogeneity between studies due to different population or threshold settings are taken into account. d. also possible to evaluate the effects of study characteristics on sensitivity and specificity separately.
Bipat, 2004[133]	Sensitivity and specificity (%) with 95%CIs for Muscularis propria invasion EUS: 94 (90, 97) 86 (80, 90) CT: NA NA MRI: 94 (89, 97) 69 (52, 82) Perirectal tissue invasion EUS: 90 (88, 92) 75 (69, 81) CT: 79 (74, 84) 78 (73, 83) MRI: 82 (74, 87) 76 (65, 84) Adjacent organ invasion EUS: 70 (62, 77) 97 (96, 98) CT: 72 (64, 79) 96 (95, 97) MRI: 74 (63, 83) 96 (95, 97) Lymph node involvement EUS: 67 (60, 77) 78 (71, 84)	Covariates included in the final models were as follows: Muscularis propria invasion EUS: publication year, sample size MRI: none Perirectal tissue invasion EUS: consecutive pt selection CT: publication year MRI: prospective data collection Adjacent organ invasion EUS: publication year, sample size CT: none MRI: publication year Lymph node involvement EUS: publication year, prospective data collection	The model accounts for: a. the heterogeneity between studies caused by different threshold settings b. the error of estimation of the sensitivity values in each study that represents the size of the population The random model also accounts for the residual heterogeneity that may remain even after adjusting for study characteristics and main techniques

Review	Results of main analysis using advanced methods	Heterogeneity investigations using advanced methods	Author comment on advanced method used
	CT: 55 (43, 79) 74 (67, 80) MRI: 66 (54, 83) 76 (59, 87)	CT: complete verification MRI: publication year, blind interpretation of results	
Bipat, 2005[137]	Sensitivity and specificity (%) with 95%CIs for Diagnosis (sensitivity and specificity respectively) Helical CT. 91 (86, 94), 85 (76, 91) Conventional CT. 86 (81, 89), 79 (60, 90) MRI 84 (78, 89)*, 82 (67, 92) US 76 (69, 82)*, 75 (51, 89) Resectability (sensitivity and specificity respectively) Helical CT. 81 (76, 85), 82 (77, 87) Conventional CT. 82 (74, 88), 76 (61, 86) MRI 82 (69, 91), 78 (63, 87) US 83 (68, 91), 63 (45, 79)* * statistically significant difference compared to helical CT	Significant predictors (diagnosis) Helical CT – sufficient description of patient popl (sens P<0.05 and spec P<0.01) Conventional CT– blinded interpretation of results (sens P<0.01) MRI – sufficient description of patient popl (sens P<0.01) US – sufficient description of patient popl (sens P<0.01 and spec P<0.01) Significant predictors (resectability) Helical CT – year of publication (spec P=0.01), departmet of origin (sens P<0.01 and spec P<0.01), sufficient description of diagnostic test (sens P<0.01 and spec P<0.01) Conventional CT– size of patient popl (sens P<0.01)	The model accounts for: c. the heterogeneity between studies caused by different threshold settings d. the error of estimation of the sensitivity values in each study that represents the size of the population
Bipat, 2005a[134]	Sensitivity (%) with 95%CIs for Non-helical CT. sensitivity 52.3 (52.1, 52.5) CT. sensitivity 63.8 (54.4, 72.2) 1.0-T MRI sensitivity 66.1 (65.9, 66.3) 1.5 T MRI sensitivity 64.4 (57.8, 70.5) FDG PET sensitivity 75.9 (61.1, 86.3) FDG-PET had significantly higher sensitivity than the other three tests (P<0.001, P=0.003, P<0.001 respectively)	Significant predictors (per patient data): Nonhelical CT – reference standard (P<0.002), blinded reference test interpretation (P<0.002) Helical CT – no predictors MRI – no predictors FDG PET – blinded index test interpretation (P<0.002)	The model accounts for: e. the heterogeneity between studies caused by different threshold settings f. the error of estimation of the sensitivity values in each study that represents the size of the population g. the residual heterogeneity that may remain even after adjustment for study design characteristics

Review	Results of main analysis using advanced methods	Heterogeneity investigations using advanced methods	Author comment on advanced method used
Glas, 2003[80]	Sensitivity and specificity (%) with 95%CIs for Diagnosis (sensitivity and specificity respectively) Cytology 55 (48, 62), 94 (90, 96) BTA 50 (30, 65), 79 (70, 86) BTA stat 70 (66, 74), 75 (64, 84) BTA TRAK 66 (62, 71), 65 (45, 81) NMP22 67 (60, 73), 78 (72, 83) Telomerase 75 (71, 79), 86 (71, 94)* *sensitivity and specificity significantly correlated (P<0.05)	Cytology Sensitivity correlated with year of publication and design (P<0.1) Cohort studies: 48 (39, 57) Case-control: 61 (52, 69) From 1990 to 2000 sensitivity decreased linearly from 80% to 52% and specificity decreased from 97% to 92%. BTA – Study design affected sensitivity: cohort 73 (60, 83) case-control 33 (26, 41); Specificity affected by blinded test interpretation: blinded 59 (46, 71), nonblind (83, 76, 88) BTS stat – sensitivity correlated with design: cohort 77 (71, 82) case-control 66 (60, 4171 NMP22 - positive correlation of sensitivity and specificity with method of patient selection, but based on only 2 studies. Telemerase: publication year – sensitivity increased from 67 to 95% and specificity decreased from 95% to 62% BTA trak – no correlations observed	Method is more convenient than Moses method Sensitivity and specificity can be interpreted as a pair Random effects nature allows systematical and coincidental differences between studies
Halligan, 2005[141]	Sensitivity and specificity (%) with 95%CIs for Detection of large polyps alone (category 1) sensitivity 93 (73, 98) specificity 97 (95, 99) HSROC curve very close to top left hand corner of plot Detection of medium and large polyps (category 2) sensitivity 86 (75, 93) specificity 86 (76, 93), HSROC curve further from top left hand corner of plot	Tried to compare studies with and without a modified reference standard but too few studies to allow meaningful analysis	HSROC model a. allows for explicit and implicit variation in threshold between studies. b. estimates the average threshold and diagnostic odds ratio, as well as variability, and it allows summary ROC curves to have either a symmetrical or an asymmetrical shape. c. allows calculation of the average operating point, which is the point on the summary ROC curve that represents the sensitivity and specificity results at the average threshold, together with 95% CIs.

Review	Results of main analysis using advanced methods	Heterogeneity investigations using advanced methods	Author comment on advanced method used
			When interpreting the results of these models, it is important to consider both these figures and the variability in sensitivity and specificity along this curve, as depicted in the ROC plot across the range of study values.
Koelemay, 2004[138]	Sensitivity and specificity (%) with 95%CIs for For detection of a 70% to 99% stenosis: sensitivity 85 (79 to 89) specificity 93 (89 to 96). For detection of 100 stenosis: In nearly all studies, the sensitivity and specificity were 100 for detection of an occlusion. A zero cell correction was not carried out due to the resulting downward bias in summary estimates due to low occlusion rates. Fixed effect pooling resulted in a sensitivity of 97 (to 99) and a specificity of 99 (98 to 100).	Incomplete reporting meant that only year of publication and design-related characteristics could be included one at a time. Diagnostic accuracy was reportedly not influenced by any covariates except for a higher specificity in prospective studies compared with retrospective studies. Data not reported	Advantages of BVN: a. estimates and incorporates the possible correlation between logit sensitivity and specificity within studies due to possible differences in threshold between studies. b. uses a random effects approach for both sensitivity and specificity, allowing for heterogeneity beyond chance due to clinical or methodological differences between studies. c. acknowledges the difference in precision by which sensitivity and specificity have been measured in each study.
Kwee, 2007[142]	Nonenhanced time-of-flight MRA (TOF-MRA) for the detection of residual flow (within the aneurysmal neck and/or coil mesh) Sensitivity 83.3 (70.3–91.3) Specificity 90.6 (80.4–95.8) Contrast-enhanced MRA (CE-MRA) for the detection of residual flow were Sensitivity 86.8 (71.4–94.5) Specificity 91.9 (79.8–97.0), respectively. There were no statistically significant differences in pooled sensitivity and specificity between TOF-MRA and CE-MRA (F test P=0.66 and P=0.82, respectively). All pooled estimates were subject to heterogeneity (P<0.05),	NA	Advantages of BVN: a. assumes a bivariate normal distribution for the logit transformed sensitivity and specificity values across studies, allowing for heterogeneity beyond chance due to clinical or methodological differences between studies. b. incorporates and estimates the correlation that might exist between estimates of sensitivity and specificity within studies.
Shaheen, 2007[130]	Sensitivity and specificity (%) with 95%CIs for Prediction of significant fibrosis threshold 0.5 (n=16): sensitivity 81 (76-86) specificity 50 (47-52) threshold 1.5 (n=15): sensitivity 35 (30-41) specificity	APRI accuracy for detecting significant fibrosis not affected by study-related or patient-related factors (P-values given). Age of study population (P=0.1), sex (P=0.96), prevalence of significaint fibrosis (P=0.46), inclusion of HIV/HCV co-infected patients (P=0.60)	Pairs of sensitivity and specificity for diagnostic thresholds are jointly analyzed, incorporating any correlation that might exist between these measures using a random effects approach

Review	Results of main analysis using advanced methods	Heterogeneity investigations using advanced methods	Author comment on advanced method used
	91 (89-92) Prediction of significant cirrhosis threshold 1.0 (n=9): sensitivity 76 (68-82) specificity 71 (69-73) threshold 2.0 (n=8): sensitivity 49 (43-55) specificity 91 (90-93)	For detection of cirrhosis, APRI accuracy was greater in studies containing higher proportion of men (P=0.001), younger participants (P=0.04), and HIV/HCV co-infected patients (P=0.03). The other covariates were not significant (data not given).	
Thangaratinam, 2007[131]	Sensitivity and specificity (%) with 95%CIs for sensitivity 63 (39 to 83) specificity 99.8 (99 to 100)	NA	Authors note that the model produces: a random effect estimate of mean sensitivity and specificity with 95% CIs, the amount of between-study variation for sensitivity and specificity separately, and the strength and shape of the correlation between sensitivity and specificity. Only the first is presented in the results. Advantages of BVN: a. accounts for the heterogeneity between studies caused by different threshold settings. b. acknowledges the difference in precision by which sensitivity and specificity have been measured in each study c. accounts for the residual heterogeneity due to clinical or methodological differences between studies.
Van Westreenen, 2004[135]	Sensitivity and specificity (%) with 95%CIs for Detection of locoregional metastases sensitivity 0.51 (0.34 to 0.69) specificity 0.84 (0.76 to 0.91) Detection of distant metastases, sensitivity 0.67 (0.58 to 0.76) specificity 0.97 (0.90 to 1.0)	NA	Model assumes a bivariate normal distribution for the logit-transformed sensitivity and specificity values across studies, allowing for additional heterogeneity between studies due to differences in study characteristics
Whiting, 2006[139]	None for HSROC	HSROC analysis shows that cohort studies with longer follow-up produced higher estimated specificity and lower estimated sensitivity (P=0.074)	
Williams, 2007[140]	Test (sensitivity and 1-specificity respectively) peak systolic velocity 0.85 (0.76, 0.90), 0.08 (0.05, 0.13) renal-aortic ratio 0.80 (0.62, 0.91), 0.12 (0.05, 0.25) acceleration time 0.74 (0.55, 0.87), 0.15 (0.07, 0.29) acceleration index 0.78 (0.67, 0.86), 0.11 (0.67, 0.86)	Peak systolic velocity the approach to failed sonographic examinations was associated with the cutpoint for test positivity but not with accuracy. Studies explicitly showing no PSV failures had a higher expected sensitivity (0.95) and hence a lower expected specificity (0.76) than those where PSV failures were excluded, PSV failures were included, or no indication of what investigators did	The model a. takes into account the uncertainty in estimates of both sensitivity and specificity within each study b. includes a random effect for both test accuracy and threshold thereby taking into account unexplained heterogeneity between studies c. allows test accuracy to vary with threshold through the inclusion of a scale (shape) parameter that provides for

Review	Results of main analysis using advanced methods	Heterogeneity investigations using advanced methods	Author comment on advanced method used
		with PSV failures was given (sens 0.81 and spec 0.93, difference P=0.004) acceleration index – test accuracy increased as test threshold increased. For every 0.5-m/s2 increase in test threshold, DOR increased an average of 3.8 times (1.4, 10.5, P=0.01). Other popl and study design characteristics had no significant effect on test performance	asymmetry in the SROC curve. This shape parameter is assumed to be constant across studies (fixed effect)

Appendix 9 Rationale for choice of topic, type of TB and test(s) for the case study

Potential for spectrum to affect test accuracy in tuberculosis

A variety of spectrum-related factors could potentially affect the accuracy of tests for diagnosing TB. In most developed countries, TB mostly affects older people, recent immigrants from developing countries, members of ethnic minorities, and the immunocompromised (mainly HIV).

Patient age can confound the diagnosis of TB, it being much more difficult to diagnose in children and more common in older people. The particular problems amongst children, are that disease is often asymptomatic, children rarely produce sputum, so that gastric aspirates are often used for mycobacterial testing, and they are also less likely to be AFB smear positive, i.e. the bacterial load in children is substantially lower in children than in adults (for both sputum and gastric aspirate specimens). [300,301]

The presence of mycobacterial infections other than tuberculosis (MOTT) or non-tuberculous mycobacteria (NTM), including the atypical forms such as *M. avium* species is also a key factor. *M. avium complex* disease occurs either as a disseminated disease largely in patients with human immunodeficiency virus (HIV) infection, or as a pulmonary disease in immunocompromised patients. The rapidly growing atypical mycobacteria, including *M. fortuitum*, *M. chelonae* and *M. abscessus* cause cutaneous, pulmonary and postsurgical wound infections.[302] The rates of infection with NTM vary across the world, with rates pulmonary NTM reported at between 1 and 15 per 100,000.[303] Generally similar rate save been reported in Europe, Japan and Australia, with a particularly high rates in South Africa.[303] A study of non-HIV positive patients in Leeds found an increase in incidence of NTM infections as a proportion of total number of recorded mycobacterial infections from 8% in 1995, to 14% in 1996, 18% in 1997, 15% in 1998 and 14% in 1999.[304] Patients infected with these mycobacteria are more likely to have false-positive results on testing.

Study setting and place of birth are further key factors that might affect test accuracy due to the variation in prevalence of both *M.TB* and other nontuberculous mycobacteria across the world. Immigrants to the UK and children born to immigrant families are more likely to have TB on arrival in the UK or to contract the disease from family members returning from visits to their countries of origin, due to higher prevalence of the disease in those countries.

Tuberculosis in immunocompromised individuals, especially those with HIV infection, may have unusual features, such as atypical pulmonary manifestations or false-negative microbiological results, which can cause diagnostic difficulties.[305] HIV infection substantially increases the risk of developing TB once infected with the bacillus, and also shortens the time to development of the disease. [305] Those with double infection have an estimated 10% risk of developing active TB each year.[306,307] HIV-positive patients may be at 10 times greater risk of multi-drug resistant TB (MDR-TB) than HIV-negative patients.[308] Other immunocompromised populations at risk for developing TB are those with diabetes mellitus,[309] those on immunosuppressive medication post organ transplantation[310] and populations receiving treatment with TNF-alpha antagonists for rheumatoid arthritis and other autoimmune diseases.[311]

Overall, the prevalence and distribution of these factors within a given study sample could strongly affect the accuracy of the test under investigation and variation between studies may contribute significantly to the heterogeneity observed in a systematic review.

Choice of pulmonary TB

A total of 368 datasets comparing a rapid diagnostic test with a reference standard were included in the HTA systematic review.[78] These covered eight different types of TB (plus a group of studies using miscellaneous specimens from various sites) and nine groups of tests. The vast majority of the evidence identified (146 datasets) was for tests for the detection of pulmonary tuberculosis and therefore this was chosen as the topic for the case study.

Choice of test(s)

Of the 146 available datasets in pulmonary TB, 110 related to nucleic acid amplification tests (NAATs); 59 evaluated commercially produced NAATs and 51 were of 'in-house' NAATs, i.e. tests developed and used within a single laboratory. The NAAT test studies therefore provided the largest single source of studies from the project.

The commercial nature of over half the test evaluations is also unique to the NAAT tests; very few serological or biochemical tests are commercially produced and although the fully automated liquid culture tests are generally commercially produced they are not evaluated with standard accuracy outcomes (e.g. sensitivity and specificity). The benefit of limiting the case study to one or more commercial tests is that the test methods and thresholds for positivity used are more standardised than for in-house tests, thereby largely eliminating this potential source of variation between studies.

In general, the studies of the commercial NAAT tests recruited more patients and were better reported than those of the inhouse tests. A meta-analysis of studies with larger sample sizes is preferable to one with many small and underpowered studies. Better reporting of study characteristics also makes it easier to judge the quality of the included studies. The mean number of patients recruited was much lower for the studies of inhouse tests (153, SD 128, range 14 to 833) compared to commercial tests (n=362, SD 506; range 22 to 3794).[78]

For these reasons the two most commonly investigated tests: the Roche Amplicor® mycobacterium tuberculosis test[312] (30 datasets) and the Gen-Probe Amplified

Mycobacterium tuberculosis Direct Test (MTD®)[313] (21 datasets) were selected for inclusion in the case study.

Appendix 10 Meta-analytic methods used for the TB case study

Aim

For the group of available studies evaluating two nucleic acid amplification tests (MTD and Amplicor) for the detection of pulmonary tuberculosis, to examine the extent to which the effect of spectrum on test accuracy can be identified or masked by currently available methods of meta-analysis as introduced in Chapter 2, i.e.

- the Moses and Littenberg summary ROC (SROC) method
- the Rutter and Gatsonis hierarchical summary ROC (HSROC) model
- the bivariate normal model.

Inclusion criteria

Population

Studies of adults or children with any form of active pulmonary tuberculosis were eligible for inclusion. Patients with any co-morbidity (including HIV infection) were included. Studies exclusively conducted in patients with non-tuberculous mycobacterial infection were excluded on the basis that these infections are rare and inclusion of them was outwith the resource constraints of the review.

Studies with more than one specimen per patient were included only where accuracy data could be extracted on a *per patient* as opposed to a *per specimen* basis or where the difference in number of specimens compared to number of patients was less than 10%. Studies of specimens 'spiked' with mycobacteria were excluded as they did not use clinical samples.

Diagnostic tests

Any study that compared one of two NAAT tests for detection of active pulmonary tuberculosis with a reference standard was included. The two eligible tests were:

- the Roche Amplicor® mycobacterium tuberculosis test (including either the original manual version and the subsequently developed automated 'Cobas Amplicor' test,
- the Gen-Probe Amplified Mycobacterium Tuberculosis Direct (MTD®) test (including either the original (AMTD) and the enhanced (EMTD) version.

Reference standards

Reference standards for tests for detecting active TB can be broadly defined as follows:

A: culture and/or microscopy smear test

B: very high clinical suspicion of TB ± response to therapy

C: clinical suspicion of TB, but it is not certain one way or the other

Studies may use one or more of these reference tests either alone or in combination with each other as a reference strategy. Strategy A alone, although previously considered good practice, is now recognised as an inadequate reference standard especially in patients with acid fast bacilli (AFB) smear negative tuberculosis. Although culture specificity is high (a positive culture result is highly indicative of the presence of mycobacteria), sensitivity is much poorer as culture can miss true cases of TB. Unfortunately, clinical diagnosis, whilst improving sensitivity, has a relatively low specificity for TB diagnosis. The definition of strategies B and C can also vary significantly, i.e. in terms of what signs and symptoms are considered to suggest the presence of TB infection. We accepted any of these categories as eligible reference tests and examined any impact on accuracy in the analyses by designating culture plus high clinical suspicion with or without additional investigations as an ideal reference strategy, i.e. definition of disease being either positive culture or high clinical suspicion.

Study setting

No restrictions on study setting were applied and studies from all countries were eligible for inclusion.

Study design

Only 'cohort' or case series type studies that compared a diagnostic test with an established reference standard in patients *suspected of having* tuberculosis were eligible for inclusion in the review. These could be either prospective or retrospective in nature.

'Case-control' type studies where the performance of a test is compared in two or more groups of patients potentially ranging from those with confirmed active TB infection through to those with diseases other than TB or even no known disease (healthy controls) were excluded. This type of design is known to be significantly more susceptible to bias than cohort studies especially when healthy control patients are included; the artificial selection of patients leading to an unrepresentative case mix of patients.

Outcome measures

The evaluation of diagnostic tests has largely focused on the establishment of test accuracy, and this was the main focus of this review. Studies that examined the effect of diagnosis on diagnostic thinking, patient management or subsequent patient outcomes were also eligible for inclusion but none were identified. Studies focusing on the establishment of technical efficacy alone were excluded.

At a minimum, accuracy studies were required to report sufficient information to allow the construction of a 2x2 contingency table. This information was used to calculate relevant accuracy statistics. Studies reporting only summary accuracy statistics without sufficient raw data to allow the construction of a 2x2 table were excluded. For studies using discrepant analysis (where false positive and/or false negative results usually against culture are resolved by examining clinical data for those patients), pre-discrepant analysis results were used wherever possible as this can be a potential source of bias.[314]

To limit the amount of potential variation resulting from varying definitions for an abnormal result between studies, data were extracted at the manufacturer's designated optimum cut-offs points where possible. Only one dataset per test comparison was included.

Literature search

Literature was identified from several sources including electronic databases and other sources. A comprehensive database of relevant articles was constructed using Reference

Manager. All databases were searched from 1975 to August 2003. Reference lists of included studies and relevant review articles were scanned to check for additional studies not identified from other sources.

A highly sensitive strategy to identify studies of tests evaluated in patients with active TB infection was used in the wider systematic review of all tests.[78] Due to the high volume of studies in TB infection, tuberculosis-related terms were combined firstly with terms relating to the tests under evaluation, and secondly in combination with a sensitive methodological filter developed to identify diagnostic accuracy studies. Due to time frame and resource constraints, searches were restricted to English language only.

Study inclusion

Studies were selected for inclusion in the review in a two-stage process. In the first instance, the literature search results (titles and abstracts) were screened independently by two reviewers to identify all citations that appeared to meet our inclusion criteria. Full manuscripts of all selected citations were retrieved. Where it was not possible to determine study eligibility from the title and/or abstract the full manuscript were obtained. A checklist for study inclusion was piloted and subsequently completed for every full paper retrieved. Any disagreements over study inclusion were resolved by consensus or if necessary by arbitration by a third reviewer.

Quality assessment

The methodological quality of all included studies was appraised using a formal quality assessment tool developed by the University of York (also funded by the HTA Programme).[21]

Use of a formal quality assessment tool allows the exploration of study design aspects either for which empirical evidence of bias exists[64,199] or that are generally accepted as important for diagnostic test studies. A list of quality assessment criteria used and guide to their interpretation is provided on page 212.

Study quality was assessed independently by two reviewers. Any disagreements were resolved by consensus or if necessary by arbitration by a third reviewer.

Data extraction

The extraction of study findings were conducted in duplicate using a pre-designed and piloted data extraction form to minimise any errors. Data were recorded onto a Microsoft Access database. Information on study participants, study design, tests and reference test details, test performance (2x2 contingency tables) and on potential sources of bias were extracted. Any disagreements between reviewers were resolved by consensus or if necessary by arbitration by a third reviewer.

Data synthesis

For each test comparison, the sensitivity, specificity and their exact 95% confidence intervals were calculated. Statistical heterogeneity of sensitivities and specificities was initially assessed using the chi-squared test and by plotting sensitivity against the false-positive rate (one minus specificity) on a ROC plot and visually considering the scatter of points.

Three methods of data synthesis were employed as described in Chapter 2 section 2.3:

1. the Moses and Littenberg summary ROC (SROC) method, both unweighted (or equal weight) and weighted by inverse variance of lnDOR. The analyses were performed using STATA version 8. The model output estimates mean accuracy (DOR) at the Q* point, and an estimate of asymmetry in the SROC curve (P-value associated with the 'S' term). The mean value of S across the primary studies was used to estimate the average sensitivity and specificity of a point on the SROC curve that lies closer to the centre of the data - the average threshold point - as described in section 2.3.2.
2. the Rutter and Gatsonis hierarchical summary ROC (HSROC) model was carried out using the PROC NLMIXED command in SAS version 8.02. The model estimates mean accuracy (DOR) at the Q* point, mean threshold and the shape of SROC curve, along with their 95% confidence intervals. The model output was used to estimate DOR at the average threshold, and was also transformed to estimate sensitivity and specificity as described by Harbord and colleagues.[79]
3. the bivariate normal (BVN) model analyses were performed with the PROC NLMIXED command in SAS version 8.02. The average sensitivity and specificity, with their 95% confidence intervals were estimated. The model output was transformed to estimate DOR at the average threshold, and was also transformed to estimate mean accuracy (DOR) at the Q* point, mean threshold and the shape of SROC.[79]

Heterogeneity investigations

For each method of meta-analysis, sources of heterogeneity were investigated by adding the following covariates to the standard models:

- test used, e.g. MTD vs Amplicor and for each test, standard versus enhanced versions
- reference standard used: culture plus clinical suspicion with or without additional tests vs culture without clinical diagnoses
- index blinded vs not blinded/unknown

Covariates were added to the models in two ways. At the most simple level, no interaction of covariate with curve shape is allowed (parallel curve models). This, by definition, assumes that the SROC curves for the two groups are parallel; the RDOR, or difference in DOR between groups, is therefore constant at all thresholds.

A further level of complexity is added where the covariate is allowed to interact with curve shape ('crossing' curve models). This can occur only for the Moses and HSROC models; an interaction of covariate with shape cannot be modeled under the bivariate normal parameterisation. Where a covariate interacts with shape, the SROC curves for the subgroups may have different shapes and therefore will cross at some point along the curves. The RDOR will not remain constant but vary systematically with threshold, to a greater or lesser extent along the length of the curves.

The addition of covariates to the models produce the following parameters to assess differences between groups:

4. difference in accuracy - the relative diagnostic odds ratio (RDOR). All three models naturally produce an estimate of RDOR at Q*, with 95% CIs. For the parallel curve models, the RDOR will be constant all along the length of the curves, however for the crossing curve models, RDOR will vary. I have estimated RDOR using the average threshold value (or for the Moses methods mean S) for the reference group and for the comparator group, this gives RDORs near to the average operating points of the two subgroups.
5. difference in threshold. Only the advanced methods produce an estimate of differences in threshold between groups; this, with its 95%CI is estimated both for the parallel and 'crossing' curve models.
6. difference in shape. This can be estimated only for the crossing curve models and only for the Moses models and the HSROC model, not for the BVN model.
7. difference in sensitivity and specificity. This was estimated for all models, both in their parallel and crossing curve forms. Only the advanced models provide confidence intervals for the differences in sensitivity and specificity, as these do not fall naturally from the Moses models.

Quality assessment criteria used - QUADAS tool (Whiting and colleagues, 2004[21])

Item		*Yes*	*No*	*Unclear*
1.	*Was the spectrum of patients representative of the patients who will receive the test in practice?*	*Yes*	*No*	*Unclear*
2.	*Were selection criteria clearly described?*	*Yes*	*No*	*Unclear*
3.	*Is the reference standard likely to correctly classify the target condition?*	*Yes*	*No*	*Unclear*
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?*	*Yes*	*No*	*Unclear*
5.	*Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?*	*Yes*	*No*	*Unclear*
6.	*Did patients receive the same reference standard regardless of the index test result?*	*Yes*	*No*	*Unclear*
7.	*Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?*	*Yes*	*No*	*Unclear*
8a.	*Was the execution of the index test described in sufficient detail to permit replication of the test?*	*Yes*	*No*	*Unclear*
8b.	*Was the execution of the reference standard described in sufficient detail to permit its replication?*	*Yes*	*No*	*Unclear*
9a.	*Were the index test results interpreted without knowledge of the results of the reference standard?*	*Yes*	*No*	*Unclear*
9b.	*Were the reference standard results interpreted without knowledge of the results of the index test?*	*Yes*	*No*	*Unclear*
10.	*Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?*	*Yes*	*No*	*Unclear*
11.	*Were uninterpretable/ intermediate test results reported?*	*Yes*	*No*	*Unclear*
12.	*Were withdrawals from the study explained?*	*Yes*	*No*	*Unclear*

Appendix 11 Primary study details

id	Study	Test	Reference standard	Index interp blinded?	N	tp	tn	fp	fn	Sens	Spec	DOR	ESS	%weight
1	Abe, 1993[173]	AMTD	C alone	?	135	29	98	5	3	0.91	0.95	189	98	0.01
2	Abu-Amero, 2002[155]	Cobas Amplicor	C alone	?	628	51	562	0	13	0.79	1.00	4292	233	0.00
3	Al Zahrani, 2000[161]	Amplicor	C+Clin+T+X	Y	489	24	430	0	33	0.42	1.00	630	204	0.00
4	Al Zahrani, 2000[161]	AMTD	C+Clin+T+X	Y	385	20	336	0	27	0.43	1.00	502	168	0.00
5	Alcala, 2001[170]	AMTD	C+Clin	?	365	54	267	36	8	0.87	0.88	50	206	0.04
6	Arimura, 1996[315]	Amplicor	C alone	?	76	19	48	2	7	0.73	0.96	65	68	0.01
7	Bemer-Melchoir, 2000[174]	Cobas Amplicor	C+Clin	?	207	21	161	0	23	0.48	1.00	296	141	0.00
8	Bennedson, 1996[316]	Amplicor	C alone	Y	3794	251	3333	42	168	0.60	0.99	119	1491	0.21
9	Bergmann, 1996[317]	Amplicor	C alone	?	502	22	465	6	9	0.71	0.99	189	116	0.02
10	Bergmann, 1999[318]	EMTD	C+Clin	?	489	20	458	6	5	0.80	0.99	305	95	0.02
11	Cartuyvels, 1996[319]	Amplicor	C alone	?	536	9	508	15	4	0.69	0.97	76	51	0.02
12	Catanzaro, 2000[320]	EMTD	C+Clin	Y	338	60	259	7	12	0.83	0.97	185	227	0.03
13	Cavusoglu, 2002[175]	AMTD	C+Clin+T+X	?	63	30	28	2	3	0.91	0.93	140	63	0.01
14	Chedore, 1999[11]	EMTD	C+Clin	?	618	194	414	8	0	1.00	0.98	18969	534	0.00
15	Chin, 1995[321]	Amplicor	C+T+X	?	227	9	204	2	12	0.43	0.99	77	76	0.01
16	Cohen, 1998[322]	Amplicor	C alone	Y	85	20	54	4	7	0.74	0.93	39	74	0.02
17	D'Amato, 1995[92]	Amplicor	C+Clin+X	?	365	17	333	4	11	0.61	0.99	129	103	0.02
18	Devallois, 1996[151]	Amplicor	C alone	?	372	20	350	0	0	0.98	1.00	28741	79	0.00
19	Ehlers, 1996[297]	AMTD	C alone	?	261	39	203	11	8	0.83	0.95	90	154	0.03
20	Eing, 1998[157]	Cobas Amplicor	C alone	?	833	25	801	4	3	0.89	1.00	1669	108	0.01
21	Gleason Beavis, 1995[323]	Amplicor	C alone	Y	270	11	249	6	4	0.73	0.98	114	57	0.01
22	Gomez-Pastrana, 2001[165]	Amplicor	C+Clin+T+X	Y	88	11	59	4	14	0.44	0.94	12	72	0.02

id	Study	Test	Reference standard	Index interp blinded?	N	tp	tn	fp	fn	Sens	Spec	DOR	ESS	%weight
23	Hoffner, 1996 (a)[169]	AMTD	C alone	Y	274	70	166	26	12	0.85	0.86	37	230	0.05
24	Hoffner, 1996 (b)[159]	AMTD	C alone	?	309	15	290	2	2	0.88	0.99	1088	64	0.01
25	Kambashi, 2001[171]	AMTD	C+Clin+T	Y	92	63	17	3	9	0.88	0.85	40	63	0.01
26	Kang, 2002[176]	Amplicor	C+Clin+T+H	?	47	11	28	0	6	0.64	0.98	101	44	0.00
27	La Rocco, 1994[168]	AMTD	C+Clin	?	246	56	184	3	3	0.95	0.98	1145	179	0.01
28	Lim, 2000[324]	Cobas Amplicor	C+Clin+T+X	Y	441	11	411	5	14	0.44	0.99	65	94	0.02
29	Lim, 2002[325]	Cobas Amplicor	C+Clin+T+X	?	128	15	107	1	5	0.75	0.99	321	68	0.01
30	Lockman, 2003[326]	Amplicor	C alone	?	112	50	35	3	24	0.68	0.92	24	100	0.02
31	Middleton, 2003[167]	AMTD	C alone	?	773	86	449	232	6	0.93	0.66	28	324	0.04
32	Mitarai, 2001 (a)[172]	Amplicor	C+Clin+T+H+X	?	116	25	48	1	42	0.37	0.98	29	113	0.01
33	Mitarai, 2001 (b)[327]	Amplicor	C+Clin+T+H+X	?	780	197	449	12	122	0.62	0.97	60	754	0.07
34	Neu, 1999[177]	Amplicor	C alone	?	30	2	25	1	0	0.83	0.94	85	11	0.00
35	Osumi, 1995[166]	AMTD	C+Clin	Y	24	0	17	3	2	0.17	0.83	1	11	0.00
36	Piersimoni, 2002[160]	EMTD	C+Clin	?	402	72	315	2	13	0.85	0.99	872	268	0.01
37	Piersimoni, 1998[328]	AMTD	C+Clin	?	219	13	172	29	5	0.72	0.86	15	66	0.02
38	Reischl, 1998[329]	Cobas Amplicor	C+Clin	?	807	81	691	14	21	0.79	0.98	190	356	0.05
39	Sato, 1998[164]	Amplicor	C alone	?	72	32	22	13	5	0.86	0.63	11	72	0.02
40	Sato, 1998[164]	AMTD	C alone	?	72	31	25	10	6	0.84	0.71	13	72	0.02
41	SeThoe, 1997[330]	Amplicor	C alone	Y	179	26	142	6	5	0.84	0.96	123	103	0.02
42	Shim, 2002[331]	Cobas Amplicor	C+Clin+T+H+X	?	331	26	276	3	26	0.50	0.99	92	175	0.02
43	Smith, 1999[154]	AMTD	C alone	?	153	9	142	0	0	0.95	1.00	5415	37	0.00
44	Smith, 1999[154]	EMTD	C alone	?	153	9	139	3	0	0.95	0.98	757	37	0.00
45	Vuorinen, 1995[162]	Amplicor	C alone	?	256	22	228	2	4	0.85	0.99	627	93	0.01
46	Vuorinen, 1995[162]	AMTD	C alone	?	256	22	227	3	4	0.85	0.99	416	93	0.01

id	Study	Test	Reference standard	Index interp blinded?	N	tp	tn	fp	fn	Sens	Spec	DOR	ESS	%weight
47	Wang, 1999[153]	AMTD	C alone	?	230	71	156	2	1	0.99	0.99	5538	198	0.00
48	Wang, 1999[153]	Cobas Amplicor	C alone	?	230	69	152	6	3	0.96	0.96	583	198	0.01
49	Yam, 1998[156]	Cobas Amplicor	C alone	?	387	38	341	0	6	0.86	1.00	4045	159	0.00
50	Yee, 2002[332]	Cobas Amplicor	C alone	?	85	12	69	1	3	0.80	0.99	276	49	0.01
51	dos Anjos Filho, 2002[163]	Amplicor	C alone	?	98	34	45	10	9	0.79	0.82	17	97	0.03

1. Test
2. Reference test used: C – culture; Clin – clinical diagnosis; H – Histology; T – treatment trial; X – x-ray
3. Index test interpreted blinded?: Y – yes; N – no; ? – can't tell
4. N: total number of patients tested with a given test
5. TP – true positives; TN – true negatives, FP – false positives; FN false neagtvies
6. Sens (95%CI) [tp/dis]: sensitivity (95% confidence interval) [number true positive/total number of diseased]
7. Spec (95%CI) [tn/nodis]: specificity (95% confidence interval) [number true negative/total number without disease]
8. DOR (95%CI): diagnostic odds ratio (95% confidence interval)
9. ESS – effective sample size
10. %weight – percentage weight accorded per study for weighted Moses analysis (weighting by inverse variance lnDOR)

Appendix 12 Bias in the standard error of the log DOR

Deeks and colleagues[81] explain the mechanism by which the standard error of the log diagnostic odds ratio, or SE(lnDOR), operates as follows.

The asymptotic estimator for the standard error is as follows:

$$\text{SE(lnDOR)} = \sqrt{\frac{1}{TP}+\frac{1}{FN}+\frac{1}{FP}+\frac{1}{TN}}$$

Where TP is true positive, FN is false negative, FP is false positive, TN is true negative. If:

DOR = Φ = (TP x TN)/(FP x FN);

n_1 = number not diseased = TN + FP;

n_2 = number with disease = TP + FN;

r = odds of testing negative in nondiseased = TN/FP,

the asymptotic estimator for the standard error can be re-expressed as:

$$\text{SE(lnDOR)} = \sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\left(r+\frac{1}{r}+2\right)+\left(\frac{\phi-1}{n_2}\right)\left(\frac{1}{r}-\frac{r\phi}{}\right)}$$

The three functions contained in this equation have the following properties.

The sample size dependent term (SSdep) which inversely relates to effective sample size[m] $(4n_1n_2)/(n_1+n_2)$, appropriately reflecting unequal numbers in diseased and nondiseased groups:

$$f(n_1,n_2)=\frac{1}{n_1}+\frac{1}{n_2}=\frac{n_1+n_2}{n_1n_2}$$

The proportion testing positive dependent term (PTPdep):

$$g(r)=r+{}^{1}\!/_{r}+2$$

The SE(lnDOR) is minimised when the numbers of true negatives and false positives are equal ($r=1$). For fixed vales of n_1 and n_2, shifting the threshold changes r and alters the standard error in a multiplicative manner:

[m] Effective sample size (ESS) is the sample size needed in equal-sized groups to achieve the available power where there are groups of unequal sizes. It will generally be less than the total number of subjects in the unequal groups. (http://www.uvm.edu/~dhowell/methods/Glossary/Glossary.html)

- where r is 1, the PTPdep term is 4,
- where r is 9 or 0.11, the PTP term is 11 and
- where r is 0.5 or 99.5, the PTP term is 201

The DOR dependent term (DORdep)

$$h(\phi, r, n_2) = \left(\frac{\phi - 1}{n_2}\right)\left(\frac{1}{r} - \frac{r}{\phi}\right)$$

$$= \left(\frac{\phi - 1}{n_2}\right)\left(\frac{FP}{TN} - \frac{FN}{TP}\right)$$

The SE(lnDOR) increases or decreases according to an additive term dependent on the DOR. The term is zero when DOR = 1 (i.e. for a test with no diagnostic value) and when sensitivity = specificity. For a fixed value of r, DORdep is positive if sensitivity is greater than specificity and negative otherwise. The magnitude of the term decreases with increasing numbers of diseased. For example, for a constant r of 49 and constant sensitivity (0.70) and specificity (0.90), with numbers of TN and FP (90 and 10) also remaining constant:

- where number of diseased=100, the DORdep term is -0.06
- where number of diseased=300, the DORdep term is -0.0

Only the first of the three terms will operate appropriately under the particular characteristics of diagnostic meta-analyses, i.e.

i. high DOR, with number of fps and fns often small
ii. explicit or implicit variation in threshold leading to variation in the proportion that are test positive
iii. unequal sample sizes for diseased and nondiseased

This has implications for estimation of confidence intervals for DORs, detection of bias graphically and statistically, and for weighting schemes when pooling data.

Appendix 13 Deletion residual analysis for Moses (eq) model: studies with ≥ 5% effect on at least one parameter (sorted by effect on DOR)

									Individual study data					Pooled analysis results minus each study				
														DOR		S		sens, spec
														b'line		b'line		b'line
id	Author	1 sens>spec	2 min sespdiff	3 sens ≥ 0.934	4 spec ≥ 0.995	5 zero cell FN	6 zero cell FP	7 exceptions	D	S	Sensitivity	Specificity	Sens minus Spec	121.1	% change	-0.17, P=0.21	change %	0.81, 0.98
14	Chedore, 1999[11]	Y	-	Y	-	Y	-	-	9.85	2.08	0.997	0.980	+0.018	80.2	-34%	-0.29, P=0.03	73%	0.79, 0.98
47	Wang, 1999[153]	-	Y	Y	-		-	-	8.62	-0.09	0.986	0.987	-0.001	101.5	-16%	-0.21, P=0.12	24%	0.80, 0.98
39	Sato, 1998[164]	Y	-	-	-	-	-	-	2.38	1.33	0.865	0.628	+0.236	139.6	+15%	-0.13, P=0.36	-24%	0.80, 0.98
40	Sato, 1998[164]	Y	-	-	-	-		-	2.56	0.73	0.838	0.714	+0.124	136.4	+13%	-0.14, P=0.32	-19%	0.80, 0.98
31	Middleton, 2003[167]	Y	-	Y	-	-	-	-	3.32	2.00	0.935	0.659	+0.275	131.8	+9%	-0.14, P=0.32	-15%	0.80, 0.98
51	dos Anjos Filho, 2002[163]	-	Y	-	-	-	-	-	2.83	-0.17	0.791	0.818	-0.027	132.5	+9%	-0.15, P=0.28	-12%	0.81, 0.98
48	Wang, 1999[153]	-	Y	Y	-		-	-	6.37	-0.10	0.958	0.962	-0.004	112.7	-7%	-0.19, P=0.18	10%	0.80, 0.98
18	Devallois, 1996[151]	-	Y	Y	Y	Y	Y	-	10.27	-2.84	0.976	0.999	-0.022	112.5	-7%	-0.16, P=0.22	-7%	0.80, 0.98
27	La Rocco, 1994[168]	-	Y	Y	-		-	-	7.04	-1.19	0.949	0.984	-0.035	113.1	-7%	-0.18, P=0.19	7%	0.80, 0.98
2	Abu-Amero, 2002[155]	-	-	-	Y	-	Y	-	8.36	-5.69	0.792	0.999	-0.207	126.8	+5%	-0.13, P=0.36	-25%	0.81, 0.98
35	Osumi, 1995[166]	-	-	-	-	-	-	X	0.00	-3.22	0.170	0.830	-0.667	128.1	6%	-0.19	13%	0.81, 0.98
25	Kambashi, 2001[171]	Y	-	-	-	-	-	-	3.68	0.21	0.875	0.850	+0.025	127.8	+6%	-0.16, P=0.26	-8%	0.80, 0.98
44	Smith, 1999[154]	-	Y	Y	-	Y	-	-	8.63	-0.74	0.950	0.976	-0.026	113.4	-6%	-0.18, P=0.19	8%	0.80, 0.98
23	Hoffner, 1996 (a)[169]	-	Y	-	-	-	-	-	3.62	-0.09	0.854	0.865	-0.011	127.9	+6%	-0.16, P=0.26	-8%	0.80, 0.98
49	Yam, 1998[156]	-	-	-	Y	-	Y	-	8.31	-4.75	0.856	0.999	-0.143	123.2	+2%	-0.14, P=0.31	-18%	0.80, 0.98

									Individual study data					Pooled analysis results minus each study				
id	Author	1 sens>spec	2 min sespdiff	3 sens ≥ 0.934	4 spec ≥ 0.995	5 zero cell FN	6 zero cell FP	7 exceptions	D	S	Sensit ivity	Specif icity	Sens minus Spec	DOR b'line 121.1	% change	S b'line -0.17, P=0.21	cha nge %	sens, spec b'line 0.81, 0.98
43	Smith, 1999[154]	-	Y	Y	Y	Y	Y	-	8.60	-2.71	0.950	0.997	-0.046	114.7	-5%	-0.16, P=0.22	-3%	0.80, 0.98
5	Alcala, 2001[170]	-	Y	-	-	-	-	-	3.91	-0.09	0.871	0.881	-0.010	126.2	+4%	-0.16, P=0.25	-6%	0.80, 0.98
3	Al Zahrani, 2000[161]	-	-	-	Y	-	Y	-	6.45	-7.07	0.422	0.999	-0.576	123.0	+2%	-0.16, P=0.28	-6%	0.81, 0.98
20	Eing, 1998[157]	-	-	-	Y	-	-	-	7.42	-3.18	0.893	0.995	-0.102	118.3	-2%	-0.16, P=0.24	-5%	0.80, 0.98
32	Mitarai, 2001 (a)[172]	-	-	-	-	-	-	X	3.34	-4.39	0.370	0.980	-0.606	121.8	1%	-0.19	12%	0.81, 0.98
Studies from categories 1-6 but without big effects on the analysis												-						
1	Abe, 1993[173]		Y					-	5.24	-0.71	0.906	0.951	-0.045	119.5	-1%	-0.17, P=0.22	+2%	0.80, 0.98
4	Al Zahrani, 2000[161]				Y		Y	-	6.22	-6.81	0.427	0.999	-0.571	122.1	+1%	-0.16, P=0.26	-4%	0.81, 0.98
7	Bemer-Melchoir, 2000[174]				Y		Y	-	5.69	-5.87	0.478	0.997	-0.519	120.8	0%	-0.17, P=0.23	+1%	0.81, 0.98
13	Cavusoglu, 2002[175]		Y					-	4.94	-0.34	0.909	0.933	-0.024	120.6	0%	-0.17, P=0.22	+1%	0.80, 0.98
26	Kang, 2002[176]						Y	-	4.61	-3.47	0.639	0.983	-0.344	121.8	+1%	-0.17, P=0.21	+2%	0.81, 0.98
34	Neu, 1999[177]					Y		-	4.44	-1.22	0.833	0.944	-0.111	123.31	+2%	-0.17, P=0.23	-2%	0.80, 0.98

Shaded cells indicate studies whose removal has ≥10% effect on at east one parameter
DOR – diagnostic odds ratio; b'line – baseline value for analyses including all 52 datatsets; % change - percentage change in DOR or S from baseline; sens, spec – average sensitivity and specificity

1 sens>spec: sensitivity greater than specificity
2 min sespdiff: minimal difference between sensitivity and specificity
3 sens ≥ 0.934: sensitivity ≥ 0.934

4 spec ≥ 0.995: specificity ≥ 0.995
5 zero cell FN: zero false negative results
6 zero cell FP: zero false positive results
7 studies with lowest sensitivities

Appendix 14 Deletion residual analysis for HSROC model: studies with ≥ 5% effect on at least one parameter (sorted by effect on DOR)

									Individual study data						Pooled analysis results minus each study					
id	Author	1 sens>spec	2 min sespdiff	3 sens ≥ 0.934	4 spec ≥ 0.995	5 zero cell FN	6 zero cell FP	7 exceptions	DOR median 129	Sens-itivity	Spec-ificity	Sens minus Spec	Total n Median 256	ESS Median 24	theta b'line -0.79	% change	DOR b'line 139.06	% change	beta b'line 0.35, P=0.06	% change
14	Chedore, 1999[11]	Y	-	Y	-	Y	-	-	18969	0.997	0.98	0.018	618	533	-0.66	-17%	109.78	-21%	0.51, P=0.01	46%
47	Wang, 1999[153]	-	Y	Y	-	-	-	-	5538	0.986	0.987	-0.001	230	198	-0.75	-6%	122.47	-12%	0.41, P=0.03	19%
39	Sato, 1998[164]	Y	-	-	-	-	-	-	11	0.865	0.628	0.236	72	72	-0.9	14%	152.99	10%	0.28, P=0.14	-19%
18	Devallois, 1996[151]	-	Y	Y	Y	Y	Y	-	28741	0.976	0.999	-0.022	372	79	-0.79	0%	125.89	-9%	0.35, P=0.06	0%
40	Sato, 1998[164]	Y	-	-	-	-	-	-	13	0.838	0.714	0.124	72	72	-0.88	11%	151.06	9%	0.30, P=0.11	-14%
51	dos Anjos Filho, 2002[163]	-	Y	-	-	-	-	-	17	0.791	0.818	-0.027	98	97	-0.85	7%	149.1	7%	0.32, P=0.09	-9%
31	Middleton, 2003[167]	Y	-	Y	-	-	-	-	28	0.935	0.659	0.275	773	324	-0.9	13%	148.18	7%	0.29, P=0.12	-15%
43	Smith, 1999[154]	-	Y	Y	Y	Y	Y	-	5415	0.95	0.997	-0.047	153	37	-0.79	0%	130.26	-6%	0.34, P=0.06	-1%
27	La Rocco, 1994[168]	-	Y	Y	-	-	-	-	1145	0.949	0.984	-0.035	246	179	-0.78	-2%	130.16	-6%	0.38, P=0.05	8%
48	Wang, 1999[153]	-	Y	Y	-	-	-	-	583	0.958	0.962	-0.004	230	198	-0.78	-2%	130.93	-6%	0.38, P=0.04	10%
23	Hoffner, 1996 (a)[169]	-	Y	-	-	-	-	-	37	0.854	0.865	-0.011	274	230	-0.84	6%	146.06	5%	0.33, P=0.08	-6%
44	Smith, 1999[154]	-	Y	Y	-	Y	-	-	757	0.95	0.976	-0.026	153	37	-0.78	-2%	132.41	-5%	0.37, P=0.05	7%
5	Alcala, 2001[170]	-	Y	-	-	-	-	-	50	0.871	0.881	-0.01	365	206	-0.84	6%	144.47	4%	0.33, P=0.08	-5%
25	Kambashi, 2001[171]	Y	-	-	-	-	-	-	40	0.875	0.85	0.025	92	63	-0.84	5%	143.75	3%	0.33, P=0.08	-4%
49	Yam, 1998[156]	-	-	-	Y	-	Y	-	4045	0.856	0.999	-0.143	387	159	-0.82	3%	134.32	-3%	0.31, P=0.10	-11%

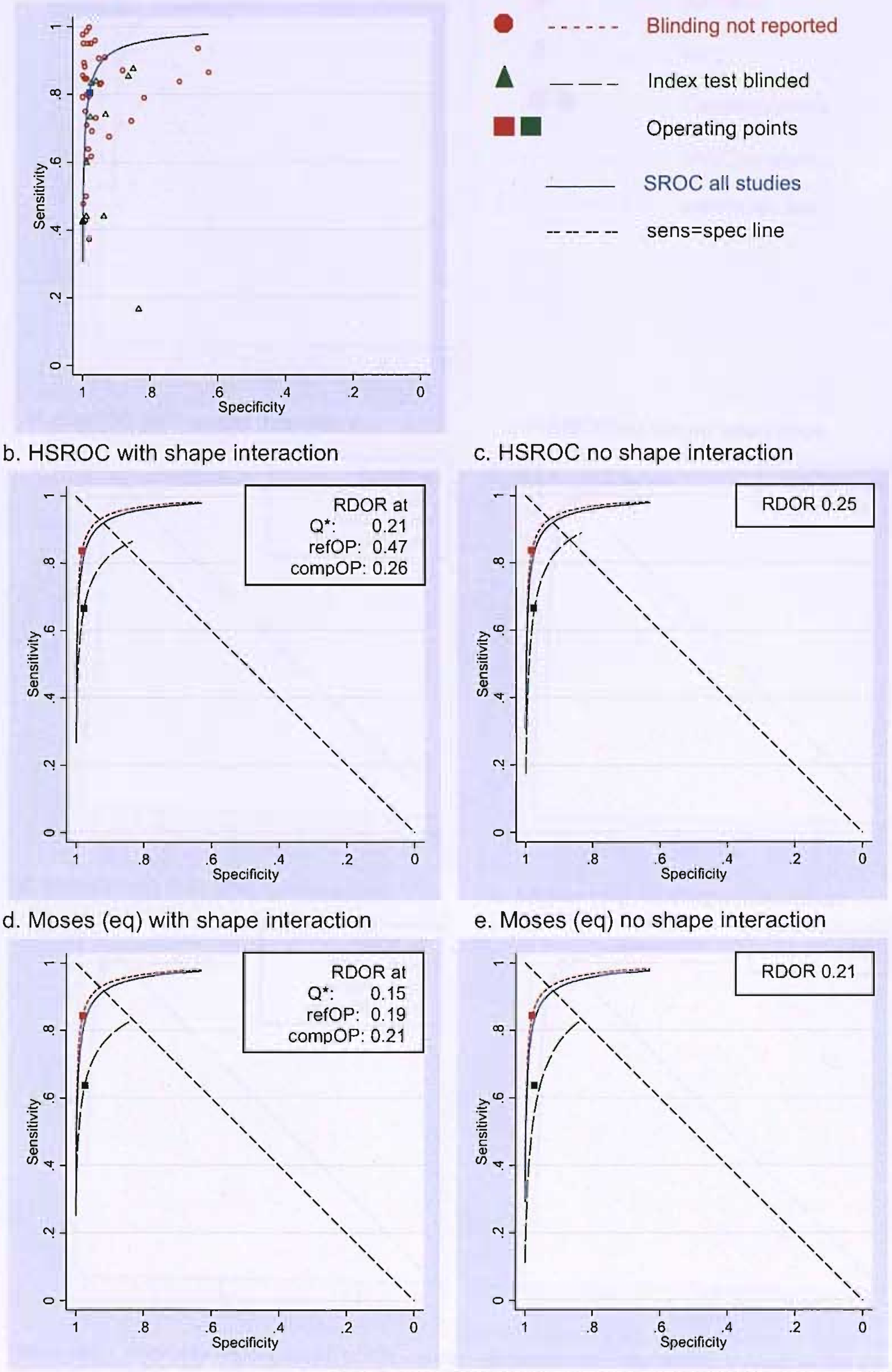
									Individual study data						Pooled analysis results minus each study					
id	Author	1 sens>spec	2 min sespdiff	3 sens ≥ 0.934	4 spec ≥ 0.995	5 zero cell FN	6 zero cell FP	7 exceptions	DOR median 129	Sens-itivity	Spec-ificity	Sens minus Spec	Total n Median 256	ESS Median 24	theta b'line - 0.79	% change	DOR b'line 139.06	% change	beta b'line 0.35, P=0.06	% change
2	Abu-Amero, 2002[155]	-	-	-	Y	-	Y	-	4292	0.792	0.999	-0.207	628	233	-0.83	5%	136.96	-2%	0.29, P=0.12	-17%
4	Al Zahrani, 2000[161]	-	-	-	Y	-	Y	-	502	0.427	0.999	-0.571	385	168	-0.77	-3%	139.97	1%	0.33, P=0.08	-6%
3	Al Zahrani, 2000[161]	-	-	-	Y	-	Y	-	630	0.422	0.999	-0.576	489	204	-0.77	-3%	139.98	1%	0.32, P=0.09	-8%
32	Mitarai, 2001 (a)[172]	-	-	-	-	-	-	X	29	0.370	0.980	-0.606	116	113	-0.72	-10%	141.08	1%	0.39, P=0.04	12%
Studies from categories 1-6 but without big effects on the analysis																				
1	Abe, 1993[173]	-	Y	-	-	-	-	-	189	0.906	0.951	-0.045	135	98	-0.8	1%	137.81	-1%	0.36, P=0.06	2%
7	Bemer-Melchoir, 2000[174]	-	-	-	Y	-	Y	-	296	0.478	0.997	-0.519	207	141	-0.77	-3%	140.27	1%	0.33, P=0.08	-4%
13	Cavusoglu, 2002[175]	-	Y	-	-	-	-	-	140	0.909	0.933	-0.024	63	63	-0.81	2%	138.12	-1%	0.35, P=0.06	2%
20	Eing, 1998[157]	-	-	-	Y	-	-	-	1669	0.893	0.995	-0.102	833	108	-0.8	0%	133.42	-4%	0.34, P=0.07	-2%
26	Kang, 2002[176]	-	-	-	-	-	Y	-	101	0.639	0.983	-0.344	47	44	-0.79	-1%	140.48	1%	0.34, P=0.07	-4%
34	Neu, 1999[177]	-	-	-	-	Y	-	-	85	0.833	0.944	-0.111	30	11	-0.79	0%	137.32	-1%	0.36, P=0.06	3%

Shaded cells indicate studies whose removal has ≥10% effect on at east one parameter
DOR – diagnostic odds ratio; b'line – baseline value for analyses including all 52 datatsets; % change - percentage change in DOR or S from baseline; sens, spec – average sensitivity and specificity

1 sens>spec: sensitivity greater than specificity
2 min sespdiff: minimal difference between sensitivity and specificity
3 sens ≥ 0.934: sensitivity ≥ 0.934
4 spec ≥ 0.995: specificity ≥ 0.995
5 zero cell FN: zero false negative results

6 zero cell FP: zero false positive results
7 studies with lowest sensitivities

Appendix 15 Plots according to index test blinding (blinding not reported as reference case)



a. All studies

b. HSROC with shape interaction

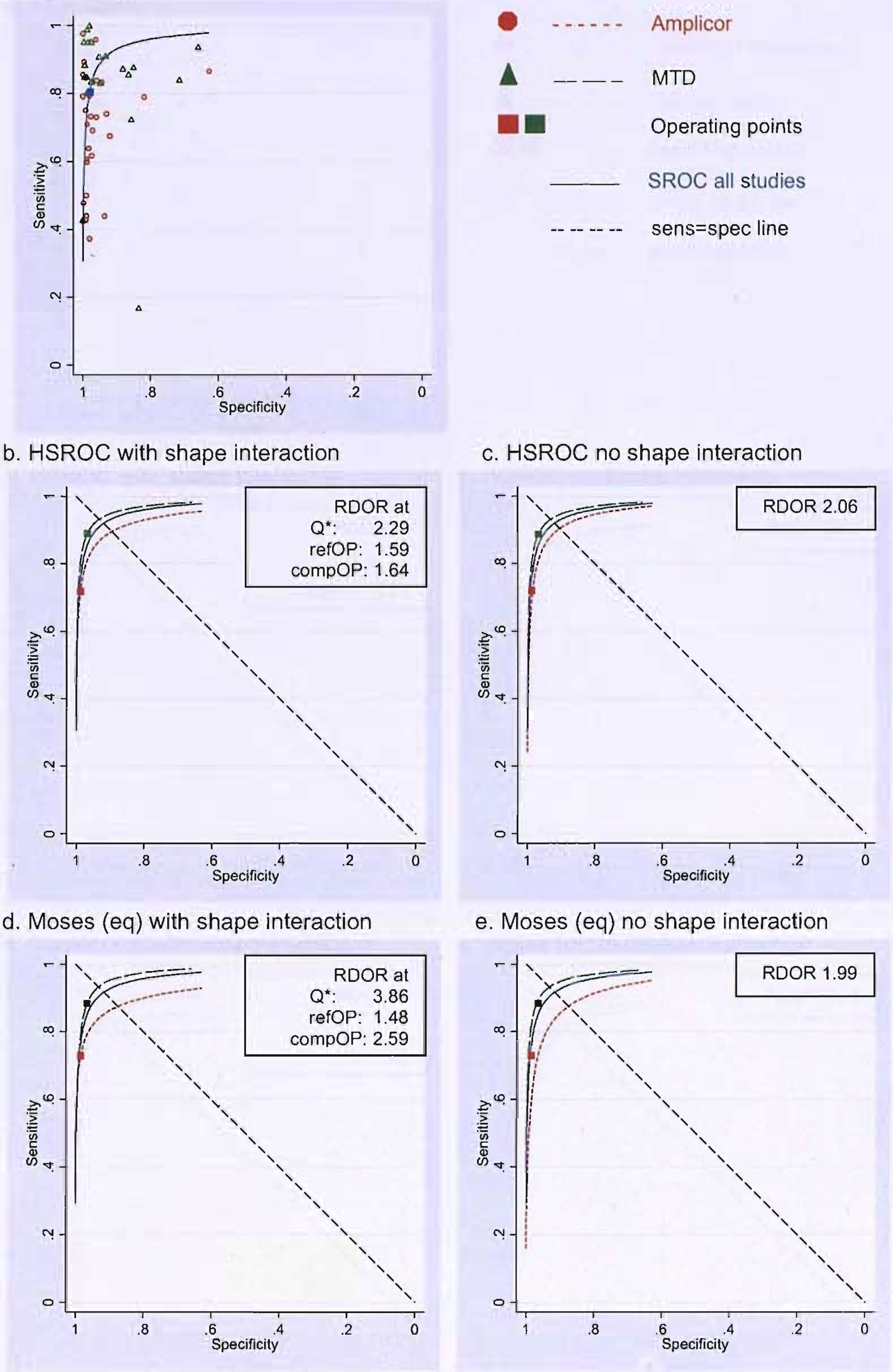
c. HSROC no shape interaction

d. Moses (eq) with shape interaction

e. Moses (eq) no shape interaction

Moses (eq) – unweighted Moses model; RDOR – relative diagnostic odds ratio (index test linding not reported is reference case (denominator); Q* - point where sensitivity=specificity (denoted by diagonal line); refOP – operating point estimated at average threshold in reference group; compOP - operating point estimated at average threshold in comparator group

Appendix 16 Plots according to test type (Amplicor as reference case)



a. All studies

b. HSROC with shape interaction

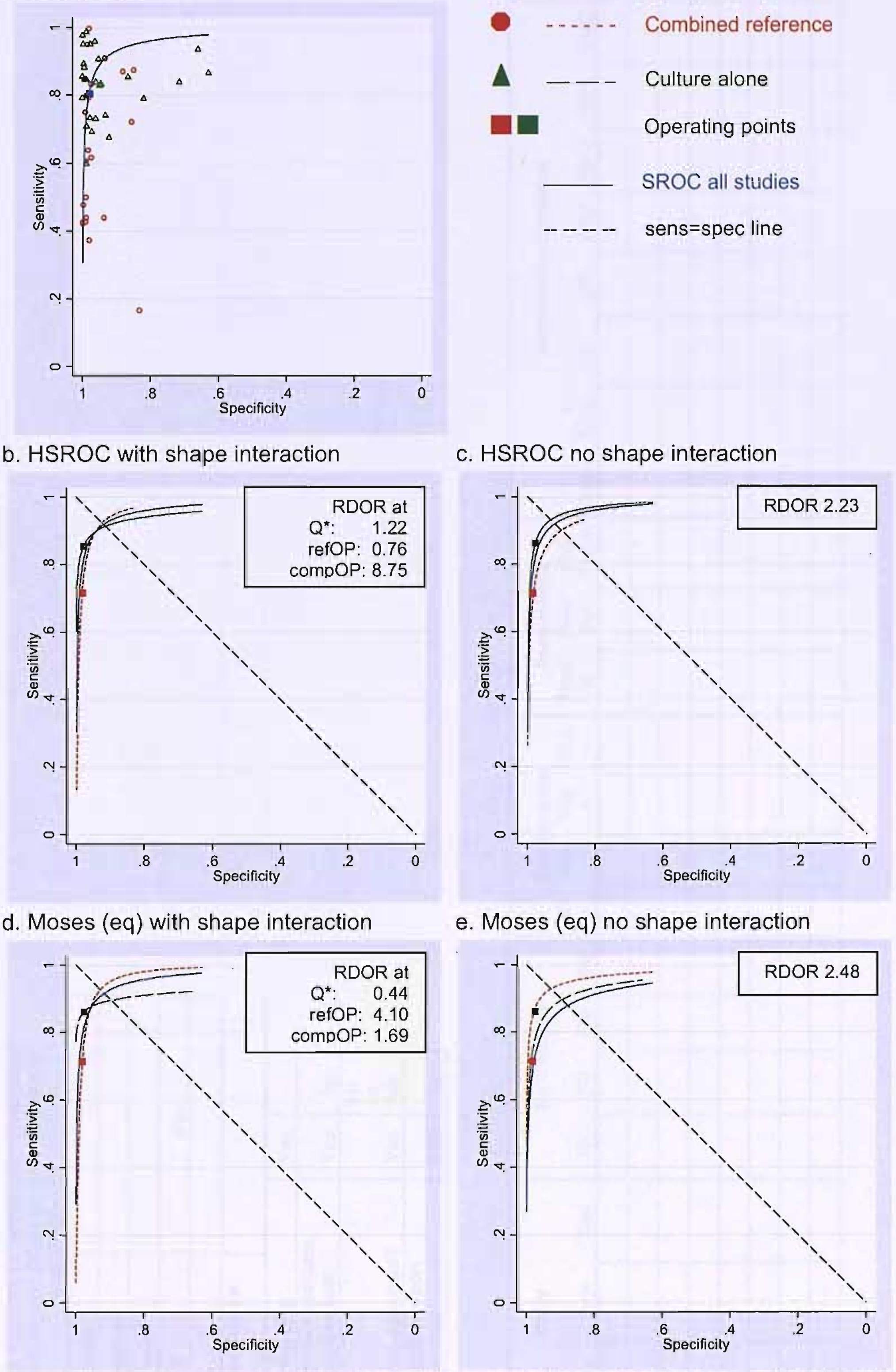
c. HSROC no shape interaction

d. Moses (eq) with shape interaction

e. Moses (eq) no shape interaction

Moses (eq) – unweighted Moses model; RDOR – relative diagnostic odds ratio (studies of Amplicor test are reference case (denominator); Q* - point where sensitivity=specificity (denoted by diagonal line); refOP – operating point estimated at average threshold in reference group; compOP - operating point estimated at average threshold in comparator group

Appendix 17 Plots according to reference test used (combined reference test as reference case)



a. All studies

b. HSROC with shape interaction

c. HSROC no shape interaction

d. Moses (eq) with shape interaction

e. Moses (eq) no shape interaction

Moses (eq) – unweighted Moses model; RDOR – relative diagnostic odds ratio (combined reference test is reference case (denominator); Q* - point where sensitivity=specificity (denoted by diagonal line); refOP – operating point estimated at average threshold in reference group; compOP - operating point estimated at average threshold in comparator group

Appendix 18 Data extraction form

NAME FILE **"YYYY SURNAME1STAUTHOR"**
ENTER EACH DISEASE IN A SEPARATE WORKSHEET

Review code	
Review author	
Review year	
Extractor	Jac
Disease	
Cochrane Review Group	

Needs checking	Yes	
Some studies evaluate more than one test	Yes	No
Covariates extracted?	Yes	No or n/a
Some studies report subgroup data	Yes	No

Study			**2 x 2 counts**				**Sample sizes**				**Prevalences**			**Performance Statistics**								**Test**	**Spec trum**	**Test**
Author	Year	Test	TP	FP	FN	TN	Dis +	Dis -	Test +	Test -	Total N	Prev Dis+	Prev Test+	Sens	Spec	LR+	LR-	PPV	NPV	DOR	AUC	detail	detail	detail

Appendix 19 Primary analysis details

		DORs at average threshold			beta, P-values			theta, P-values	Ratio of DORs (RORs)			Subgroups by key characteristics[a]				
id	Review	Moses (eq)	Moses (w)	HSROC	Moses (eq)	Moses (w)	HSROC	HSROC	Moses (eq) v HSROC	Moses (w) v HSROC	Moses (w) v Moses (eq)	DOR	Range in 'S'	Zero cells	Asym metry	Thresh old diffs
1	Balk 2001[180]	38	15	34	0.29	0.06	0.12	0.00	1.11	0.45	0.41	1	3	2	2	1
2	Bricker 2000[179]	705	688	872	0.96	0.41	0.52	0.01	0.81	0.79	0.98	3	1	1	3	1
3	Buchanan 2001[181]	2	3	1	0.13	0.20	0.13	0.07	4.81	5.51	1.15	1	3	1	2	1
4	Chapell 2002[182]	80	61	117	0.74	0.04	0.18	0.01	0.68	0.52	0.76	3	1	3	2	1
5	Delgado 2003[178]	19	19	25	0.30	0.30	0.42	0.38	0.75	0.75	1.00	1	2	3	3	3
6	Deville 2000[123]	4	3	.	0.82	0.07	.	.	.	.	0.86		3	1	.	
7	Dijkhuizen 2000[100]	823	481	18220	0.54	0.13	0.34	0.10	0.05	0.03	0.58	3	2	3	3	2
8	Eden 2001[183]	12	11	14	0.88	0.10	0.36	0.00	0.82	0.75	0.91	1	1	3	3	1
9	Flemons 2003[184]	57	31	54	0.02	0.01	0.02	0.00	1.05	0.57	0.54	2	3	2	1	1
10	Flobbe 2002[185]	30	26	30	0.07	0.00	0.01	0.00	1.01	0.88	0.87	1	1	1	1	1
11	Gifford 2000[186]	4	4	4	0.03	0.02	0.12	0.48	0.99	0.95	0.96	1	2	2	2	3
12	Glas 2003[80]	29	16	32	0.02	0.09	0.01	0.00	0.92	0.49	0.53	1	2	2	1	1
13	Gould 2001[107]	72	69	107	0.05	0.52	0.71	0.06	0.68	0.65	0.96	3	2	3	3	1
14	Gould 2003[187]	52	34	50	0.96	0.94	0.72	0.79	1.04	0.69	0.66	2	1	2	3	3
15	Gray 2000[188]	36	22	44	0.40	0.69	0.35	0.11	0.83	0.50	0.60	2	2	2	3	2
16	Ioannidis[243]	96	35	83	0.09	0.05	0.08	0.40	1.17	0.42	0.36	2	3	2	1	3
17	Kittler 2002[104]	69	37	69	0.00	0.11	0.01	0.07	1.00	0.53	0.54	2	2	1	1	1
18	Koelemay 2001[190]	228	91	384	0.15	0.03	0.08	0.28	0.59	0.24	0.40	3	2	3	1	2
19	Lysakowski 2001[191]	2	1	.	0.71	0.84	.	.	.	.	0.68		1	3	.	
20	MSAC 2002[192]	46	25	368	0.46	0.39	0.22	0.31	0.12	0.07	0.55	3	3	3	2	2
21	Nallamothu 2001[193]	16	12	88	0.92	0.68	0.62	0.00	0.18	0.13	0.76	2	1	1	3	1
22	Patwardhan 2004[194]	33	21	28	0.03	0.06	0.07	0.01	1.16	0.73	0.63	1	3	2	1	1
23	Romagnuolo 2003[195]	285	201	438	0.35	0.09	0.13	0.92	0.65	0.46	0.71	3	2	3	2	3
24	Sauerland 2004[196]	91	72	95	0.87	0.73	0.71	0.56	0.95	0.75	0.79	2	2	2	3	3

25	Sotiriadis 2003[197]	6	7	7	0.04	0.39	0.36	0.00	0.88	1.12	1.27	1	2	1	3	1
26	Varonen 2000[198]	16	12	15	0.91	0.95	0.99	0.78	1.03	0.78	0.75	1	1	1	3	3
27	Visser 2000[119]	164	83	156	0.03	0.00	0.01	0.20	1.05	0.53	0.51	3	2	1	1	2
28	Whitsel 2000[114]	6	2	5	0.19	0.03	0.08	0.00	1.20	0.49	0.41	1	2	1	1	1
29	Wiese 2000[129]	757	764	.	0.83	0.72	.	.	.	.	1.01		2	3	.	

[a] The stratification by DOR is based on the HSROC overall pooled estimate; where the HSROC model did not run, it is based on the unweighted Moses model result.
The stratification by range in 'S' is based on values for 'S' from Moses model
The stratification by zero cells number of zero false positive or false negative cells as a percentage of the total number of cells per analysis
The stratification by degree of asymmetry based on P-value associated with shape term from HSROC model
The stratification by threshold differences based on P-value associated with threshold term from HSROC model

Appendix 20 Heterogeneity investigations: comparison of relative diagnostic odds ratios between models (Ratio of RDORs)

					Parallel curves			Crossing curves									Parallel v Crossing curves		
								Q*			ref threshold			comp threshold					
id	**Review**	**Comparator**	**Reference**	**n**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**M(eq)**	**M(w)**	**H**
1.1	Balk 2001[180]	hospitalised patients	emergency department patients with symptoms/pain	14v18	0.6	1.2	0.7	0.5	1.6	0.8	0.5	0.2	0.4	0.5	1.2	0.6	1.5	2.9	2.1
2.1[a]	Bricker 2000[179]	tertiary care	primary/secondary care	4v7	0.7	-	-	0.5	76.8	35.6	0.7	0.2	0.1	0.7	2.7	1.8	0.0	0.0	
2.2[b]	Bricker 2000[179]	2nd trimester	1st, 2nd and 3rd trimester	6v5	0.9	1.4	1.3	0.1			0.8			1.2			0.0	0.0	
2.3	Bricker 2000[179]	low risk	unselected	4v7	0.9	0.8	0.8	0.1	51.0	5.8	0.7	0.2	0.1	0.6	2.5	1.6	0.0	0.0	0.0
3.1	Buchanan 2001[181]	prison release	community/hospital discharges	8v13	0.3	1.0	0.3	0.9	1.0	0.9	0.7	1.5	1.1	0.7	1.0	0.7	1.0	0.4	1.0
3.2	Buchanan 2001[181]	time at risk <=20mos	>20 mos	10v8	1.0	0.9	0.9	1.0	1.0	1.0	1.1	1.3	1.4	1.1	1.0	1.1	1.1	1.6	1.1
4.1	Chapell 2002[182]	possible age bias	no bias or not reported	4v9	1.9	0.7	1.3	2.0	11.3	22.5	0.8	0.5	0.6	1.2	0.5	0.4	0.0	0.1	0.5
4.2	Chapell 2002[182]	possible bias to easy cases	no bias to easy cases	5v8	1.1	0.4	0.4	6.0	0.6	3.4	0.8	0.7	0.6	0.9	2.1	1.9	0.2	0.4	0.2
4.3	Chapell[182]	symptoms/presented cases	unspecified diagnosis	8v5	0.7	0.6	0.9	7.5	0.4	0.1	0.8	1.5	1.2	0.7	0.5	0.7	0.3	0.3	0.8
5.1	Delgado 2003[178]	unknown primary tumours	other	8v7	1.2	0.4	0.4	1.0	0.3	0.3	1.0	0.4	0.4	1.0	0.2	0.2	0.9	1.1	0.1
6.1	Deville 2000[123]	previous surgery	no previous surgery	8v9	0.6	1.6	1.0	0.6	0.6	1.1	0.5	1.8	0.9	0.5	1.8	0.9	1.3	1.0	0.8
6.2[b]	Deville 2000[123]	bilateral excluded	bilateral not excluded	3v14	0.7	1.2	0.9	0.6			0.7			0.7			1.4	0.6	
6.3	Deville 2000[123]	<=60% men	>60% men	10v4	1.1	0.9	1.0	1.1	0.9	1.0	0.9	1.0	0.9	1.1	1.0	1.1	0.8	0.8	0.8
7.1[a] [b]	Dijkhuizen 2000[100]	pre and post-menopausal women	post-menopausal women only	22v7	1.0			0.5			1.0			1.3			0.3	0.7	
7.2[a] [b]	Dijkhuizen 2000[100]	asymptomatic or both	symptomatic only	20v13	0.6			1.8			0.6			0.6			0.2	0.1	
8.1	Eden 2001[183]	environmental exposure	medical/not exposed	3v4	0.5	0.5	1.0	1.3	1.9	2.6	1.0	0.9	0.9	1.0	0.7	0.7	0.0	0.1	0.2
9.1	Flemons 2003[184]	home setting	sleep laboratory	13v36	1.0	0.9	0.8	0.8	1.0	0.8	0.9	1.9	1.7	0.9	0.9	0.9	1.0	4.0	1.1
9.2	Flemons 2003[184]	<75%men	75-100% men	10v29	1.4	0.9	1.3	1.5	0.5	0.8	0.7	0.6	0.4	1.5	0.5	0.7	0.9	0.8	0.5

					Parallel curves			Crossing curves									Parallel v Crossing curves		
								Q*			ref threshold			comp threshold					
id	Review	Comparator	Reference	n	1	2	3	1	2	3	1	2	3	1	2	3	M(eq)	M(w)	H
9.3	Flemons 2003[184]	mean AHI[14]<=30	Ahl>30	15v17	0.8	0.9	0.7	0.8	0.7	0.6	0.7	0.5	0.8	0.8	0.8	0.7	0.8	0.7	0.6
9.4	Flemons 2003[184]	mean BMI<=30	BMI>30	9v25	0.8	0.8	0.7	1.0	0.3	0.3	1.0	0.4	0.4	1.0	0.1	0.1	0.5	0.4	0.1
10.1	Flobbe 2002[185]	pathology referral	clinical examination or mammography	13v9	1.2	0.9	1.0	1.2	0.7	0.8	1.2	0.6	0.5	1.2	0.6	0.7	1.0	1.8	0.7
11.1	Gifford 2000[186]	age <=70	>70 years	3v8	0.5	0.5	1.0	0.4	0.4	1.0	0.3	0.3	1.0	0.6	0.6	1.0	0.8	1.0	1.0
11.2	Gifford 2000[186]	dementia/memory clinics	other setting	5v6	0.6	0.6	0.4	0.8	0.7	0.5	0.8	1.9	1.5	0.8	0.5	0.4	0.9	0.7	0.9
11.3	Gifford 2000[186]	diagnostic criteria met	referrals	6v5	0.5	0.9	0.5	0.5	0.9	0.4	0.5	1.2	0.6	0.5	0.8	0.4	1.0	2.1	0.9
12.1	Glas 2003[80]	<30% Grade 1 tumours	>=30% Grade 1 tumours	14v6	0.8	1.0	0.7	0.1	3.3	0.3	1.0	3.0	2.9	0.6	1.1	0.6	0.3	2.7	0.9
12.2	Glas 2003[80]	<30% Grade 2 tumours	>=30% Grade 2 tumours	6v14	3.0	0.5	1.5	1.4	0.8	1.1	1.6	0.2	0.3	1.6	0.2	0.3	0.1	0.1	0.1
12.3	Glas 2003[80]	<30% Grade 3 tumours	>=30% Grade 3 tumours	8v12	0.8	1.0	0.8	0.7	3.1	2.1	0.6	1.1	0.7	0.6	1.2	0.8	0.3	0.3	0.8
12.4[b]	Glas 2003[80]	100% urological	rest		1.2	1.0	1.2	0.0			0.7			0.5			0.2	0.2	
13.1	Gould 2001[107]	>=70% men	<70% men	14v14	0.8	1.4	1.1	0.9	0.2	0.2	0.9	0.1	0.1	1.3	0.3	0.3	0.4	0.3	0.1
13.2	Gould 2001[107]	<60years	>=60 years	7v17	0.7	1.4	0.9	0.5	0.6	0.3	1.6	0.0	0.1	0.6	0.1	0.1	0.3	0.3	0.1
14.1	Gould 2003[187]	>=70% men	<70% men	12v10	0.7	0.3	0.5	0.6	0.2	0.4	0.6	0.6	0.4	0.6	0.4	0.7	0.8	0.9	0.9
14.2[a b]	Gould 2003[187]	<60 years	>=60 years	4v21	1.1			1.3			1.4			1.6			2.6	2.2	
15.1	Gray 2000[188]	suspicion/lesions	cancer history	10v4	0.1	1.6	0.2	0.1	2.5	0.2	0.1	0.3	1.9	0.3	1.0	0.3	0.5	1.0	0.8
16.1	Ioannidis[243]	symptoms suggestive of ACI	pts with chest pain	4v6	0.7	0.9	0.6	0.6	2.6	1.6	0.5	4.7	2.3	0.3	0.8	0.2	0.3	1.1	0.8
16.2	Ioannidis[243]	<65 years	>=65 years	3v4	1.1	0.8	0.9	1.1	0.6	0.5	1.0	0.9	0.8	1.4	0.8	1.2	0.5	0.3	1.4
16.3[b]	Ioannidis[243]	<65% men	>=65% men	3v4	1.7	0.7	1.3	1.0			0.9			0.5			0.1	0.6	
17.1	Kittler 2002[104]	non-melanocytic lesions excluded	non-melanocytic lesions included	4v9	2.2	0.3	0.6	2.7	0.0	0.0	2.6	0.5	0.2	2.3	0.0	0.0	1.7	1.4	0.0
18.1	Koelemay 2001[190]	<65 years	>=65 years	9v7	2.7	0.7	0.3	3.9	0.0	0.0	1.5	0.7	0.5	2.3	18.2	42.1	0.0	0.0	1.0
18.2	Koelemay 2001[190]	<70% men	>=70% men	7v11	2.1	0.9	1.9	0.4	68.8	27.5	1.2	0.0	0.0	1.7	0.5	0.8	0.0	0.1	0.9

[14] mean apnoea-hypopnea index

					Parallel curves			Crossing curves									Parallel v Crossing curves		
								Q*			ref threshold			comp threshold					
id	Review	Comparator	Reference	n	1	2	3	1	2	3	1	2	3	1	2	3	M(eq)	M(w)	H
18.3	Koelemay 2001[190]	<65% with intermittent claudication	>=65% with intermittent claudication	5v10	1.9	0.6	1.1	1.3	1.3	1.7	1.6	1.8	2.9	1.9	0.5	0.9	0.4	0.7	1.1
19.1 a b	Lysakowski 2001[191]	heterogeneous population	homogenous population	4v3	0.7						0.7			0.7					
20.1	MSAC 2002[192]	<50% men	>=50% men	6v6	2.6	0.0	0.1	1.8	0.0	0.0	2.1	0.7	1.5	2.1	0.0	0.0	1.3	19.0	1.3
20.2	MSAC 2002[192]	families /pedigree	definite/suspected/prenatal	8v4	4.3	0.3	1.2	25.9	0.4	11.0	8.0	1.5	11.8	7.4	0.2	1.7	0.8	0.1	1.2
21.1	Nallamothu 2001[193]	<55 years	>=55 years	5v9	0.8	0.6	0.5	0.4	2.1	0.8	0.8	0.7	0.6	0.7	0.6	0.5	0.4	0.5	1.3
21.2	Nallamothu 2001[193]	<65%men	>=65% men	7v7	0.4	0.6	1.3	2.3	0.1	0.0	0.6	0.4	0.6	0.6	1.3	0.8	0.1	0.1	0.4
22.1	Patwardhan 2004[194]	<70 years	>=70 years	11v5	0.4	1.3	0.5	2.0	0.5	0.3	0.4	2.1	0.9	0.5	0.6	0.3	0.4	0.4	1.1
22.2	Patwardhan 2004[194]	healthy controls	diseased controls	13v6	0.7	1.1	0.8	2.4	0.4	0.9	0.8	0.3	0.4	1.3	0.5	0.6	1.0	0.3	0.3
23.1	Romagnuolo 2003[195]	wide variety of possible diagnoses	stones or cancer diagnoses	11v35	1.0	0.7	0.8	1.5	0.5	0.7	1.1	0.3	0.3	1.0	0.4	0.4	0.8	0.6	0.6
24.1	Sauerland 2004[196]	adults	children	10v3	0.6	1.9	1.2	0.6	1.2	0.8	0.6	0.1	0.2	0.6	2.4	1.5	1.0	1.0	0.6
25.1	Sotiriadis 2003[197]	<=30 years	>30 years	4v8	0.9	0.7	0.6	5.3	0.1	0.7	0.7	1.1	0.8	0.7	1.0	0.7	1.0	1.8	0.2
25.2	Sotiriadis 2003[197]	high risk	low risk/routine	7v5	1.3	0.4	0.5	0.9	0.0	0.0	1.4	0.5	0.4	1.3	0.3	0.4	0.6	0.8	0.0
26.1	Varonen 2000[198]	ENT clinic	general clinic	3v4	0.8	1.2	1.0	0.9	1.6	1.5	0.9	3.4	3.2	0.8	1.8	1.4	1.1	1.1	1.6
27.1	Visser 2000[119]	<=60% men	>60% men	8v8	0.1	0.9	0.1	0.0	2.0	0.1	0.4	1.0	0.4	0.1	2.4	0.2	0.5	0.5	1.2
27.2	Visser 2000[119]	<=65 years	>65 years	8v8	1.0	0.8	0.8	5.8	0.5	3.0	0.9	0.7	0.6	0.7	0.5	0.3	2.0	0.3	1.2
27.3	Visser 2000[119]	N America	other country	14v7	1.6	0.7	1.2	4.3	0.7	3.2	2.6	0.7	1.8	1.6	0.8	1.2	1.0	0.4	1.0
28.1	Whitsel 2000[114]	<=40 years	> 40 years	8v8	0.3	1.7	0.5	0.8	0.1	0.1	0.3	0.0	0.0	0.5	0.0	0.1	0.1	0.5	0.0
28.2	Whitsel 2000[114]	<=50% men	>50% men	5v11	0.7	0.4	0.6	0.8	0.2	0.2	0.5	0.4	0.8	0.7	0.3	0.4	0.7	0.6	0.6
28.3	Whitsel 2000[114]	<=50% type 1 diabetes	50-100%	5v10	1.1	1.0	1.1	0.2	0.0	0.0	1.2	0.0	0.0	1.3	0.0	0.0	0.1	0.4	0.0
28.4	Whitsel 2000[114]	mean duration <=10 years	>10 years	10v4	0.4	1.2	0.5	0.5	2.1	1.0	0.5	1.9	0.8	0.4	1.4	0.6	0.8	0.7	1.4
29.1 a b	Wiese 2000[129]	STD clinic	speciality/general clinic	14v16	1.1			1.0			1.1			1.1			0.1	0.3	

1 - Moses (w) versus Moses (eq) model comparison; 2 – Moses (eq) versus HSROC model comparison; 3 - Moses (w) versus HSROC model comparison
a denotes covariates for which the parallel curve HSROC analysis could not be completed b denotes covariates for which the crossing curve HSROC analysis could not be completed

Appendix 21 Heterogeneity investigations – P-values for RDORs per model

								Crossing curve								
					Parallel curve models			Q*			ref threshold			comp threshold		
id	study	Comparator group	Reference group	n	M(eq)	M(w)	H	M(eq)	M(w)	H	M(eq)	M(w)	H	M(eq)	M(w)	H
1.1	Balk 2001[180]	hospitalised patients	emergency department patients with symptoms/pain	14v18	0.03	0.06	0.03	0.20	0.92	0.64	0.02	0.04	0.90	0.07	0.49	0.20
2.1[a]	Bricker 2000[179]	tertiary care	primary/secondary care	4v7	0.48	0.48		0.18	0.10	0.26	0.54	0.72	0.19	0.27	0.21	0.93
2.2[b]	Bricker 2000[179]	2nd trimester	1st, 2nd and 3rd trimester	6v5	0.75	0.70	0.98	0.17	0.10		0.79	0.98		0.62	0.31	
2.3	Bricker 2000[179]	low risk	unselected	4v7	0.43	0.39	0.24	0.20	0.17	0.24	0.47	0.67	0.17	0.28	0.40	0.92
3.1	Buchanan 2001[181]	prison release	community/hospital discharges	8v13	0.12	0.19	0.12	0.20	0.16	0.16	0.13	0.31	0.17	0.13	0.34	0.12
3.2	Buchanan 2001[181]	time at risk <=20mos	>20 mos	10v8	0.43	0.43	0.33	0.60	0.52	0.48	0.45	0.22	0.53	0.43	0.19	0.34
4.1	Chapell 2002[182]	possible age bias	no bias or not reported	4v9	0.48	0.09	0.41	0.03	0.00	0.36	0.98	0.76	0.36	0.24	0.07	0.96
4.2	Chapell 2002[182]	possible bias to easy cases	no bias to easy cases	5v8	0.09	0.10	0.04	0.11	0.01	0.10	0.09	0.14	0.10	0.08	0.11	0.93
4.3	Chapell[182]	symptoms/presented cases	unspecified diagnosis	8v5	0.50	0.97	0.89	0.71	0.29	0.99	0.48	0.74	0.96	0.60	0.96	0.85
5.1	Delgado 2003[178]	unknown primary tumours	other	8v7	0.80	0.58	0.71	0.72	0.72	0.63	0.86	0.86	0.31	0.73	0.73	0.74
6.1	Deville 2000[123]	previous surgery	no previous surgery	8v9	0.08	0.53	0.36	0.21	0.80	0.97	0.05	0.83	0.30	0.06	0.87	0.63
6.2[b]	Deville 2000[123]	bilateral excluded	bilateral not excluded	3v14	0.37	0.75	0.47	0.83	0.74		0.47	0.96		0.36	0.77	
6.3	Deville 2000[123]	<=60% men	>60% men	10v4	0.99	0.72	0.71	0.73	0.39	0.34	0.97	0.86	0.87	0.88	0.59	0.78
7.1[a b]	Dijkhuizen 2000[100]	pre and post-menopausal women	post-menopausal women only	22v7	0.56	0.66		0.17	0.33		0.37	0.37		0.88	0.65	
7.2[a b]	Dijkhuizen 2000[100]	asymptomatic or both	symptomatic only	20v13	0.32	0.91		0.05	0.06		0.44	0.78		0.27	0.92	
8.1	Eden 2001[183]	environmental exposure	medical/not exposed	3v4	0.95	0.42	0.36	0.42	0.34	0.53	0.82	0.86	0.44	0.79	0.87	0.92
9.1	Flemons 2003[184]	home setting	sleep laboratory	13v36	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.00
9.2	Flemons 2003[184]	<75%men	75-100% men	10v29	0.97	0.37	0.87	0.77	0.21	0.34	0.87	0.53	0.14	0.92	0.30	0.48
9.3	Flemons 2003[184]	mean AHI[15]<=30	Ahl>30	15v17	0.14	0.25	0.07	0.11	0.11	0.06	0.38	0.77	0.81	0.11	0.14	0.08
9.4	Flemons 2003[184]	mean BMI<=30	BMI>30	9v25	0.61	0.94	0.48	0.09	0.18	0.12	0.48	0.55	0.08	0.61	0.67	0.27
10.1	Flobbe 2002[185]	pathology referral	clinical examination or mammography	13v9	0.30	0.04	0.07	0.28	0.04	0.12	0.28	0.04	1.00	0.36	0.07	0.06
11.1	Gifford 2000[186]	age <=70	>70 years	3v8	0.57	0.64	0.84	0.51	0.67	0.84	0.52	0.63	0.74	0.66	0.92	0.92
11.2	Gifford 2000[186]	dementia/memory clinics	other setting	5v6	0.28	0.74	0.18	0.20	0.41	0.26	0.18	0.38	0.61	0.28	0.52	0.28

[15] mean apnoea-hypopnea index

								Crossing curve								
					Parallel curve models			Q*			ref threshold			comp threshold		
id	study	Comparator group	Reference group	n	M(eq)	M(w)	H	M(eq)	M(w)	H	M(eq)	M(w)	H	M(eq)	M(w)	H
11.3	Gifford 2000[186]	diagnostic criteria met	referrals	6v5	0.01	0.07	0.06	0.02	0.09	0.07	0.02	0.10	0.21	0.02	0.10	0.11
12.1	Glas 2003[80]	<30% Grade 1 tumours	>=30% Grade 1 tumours	14v6	0.02	0.05	0.03	0.24	0.86	0.16	0.03	0.06	0.05	0.03	0.06	0.07
12.2	Glas 2003[80]	<30% Grade 2 tumours	>=30% Grade 2 tumours	6v14	0.11	0.00	0.01	0.00	0.00	0.02	0.04	0.01	0.19	0.08	0.02	0.20
12.3	Glas 2003[80]	<30% Grade 3 tumours	>=30% Grade 3 tumours	8v12	0.48	0.62	0.46	0.25	0.30	0.38	0.60	0.91	0.37	0.46	0.92	0.74
12.4[b]	Glas 2003[80]	100% urological	rest		0.83	0.52	0.84	0.41	0.53		0.55	0.81		0.96	0.33	
13.1	Gould 2001[107]	>=70% men	<70% men	14v14	0.46	0.72	0.91	0.46	0.51	0.31	0.40	0.48	0.27	0.00	0.54	0.54
13.2	Gould 2001[107]	<60years	>=60 years	7v17	0.60	0.94	0.96	0.30	0.52	0.57	0.95	0.53	0.51	0.57	0.77	0.65
14.1	Gould 2003[187]	>=70% men	<70% men	12 v10	0.40	0.62	0.34	0.30	0.50	0.44	0.43	0.93	0.50	0.38	0.82	0.75
14.2[a] [b]	Gould 2003[187]	<60 years	>=60 years	4v21	0.09	0.11		0.69	0.51		0.26	0.14		0.08	0.05	
15.1	Gray 2000[188]	suspicion/lesions	cancer history	10v4	0.55	0.13	0.76	0.43	0.15	0.66	0.53	0.38	0.83	0.89	0.68	0.84
16.1	Ioannidis[243]	symptoms suggestive of ACI	pts with chest pain	4v6	0.03	0.03	0.01	0.01	0.01	0.00	0.01	0.01	0.02	0.03	0.57	0.07
16.2	Ioannidis[243]	<65 years	>=65 years	3v4	0.71	0.62	0.49	0.93	0.93	0.80	0.99	1.00	0.83	0.74	0.68	0.54
16.3[b]	Ioannidis[243]	<65% men	>=65% men	3v4	0.49	0.06	0.17	0.04	0.01		0.05	0.01		0.81	0.46	
17.1	Kittler 2002[104]	non-melanocytic lesions excluded	non-melanocytic lesions included	4v9	0.52	0.16	0.10	0.86	0.28	0.18	0.66	0.18	0.34	0.24	0.09	0.39
18.1	Koelemay 2001[190]	<65 years	>=65 years	9v7	0.81	0.15	0.91	0.08	0.05	0.89	0.88	0.53	0.75	0.10	0.03	0.87
18.2	Koelemay 2001[190]	<70% men	>=70% men	7v11	0.17	0.01	0.25	0.14	0.15	0.38	0.08	0.03	0.58	0.18	0.03	0.26
18.3	Koelemay 2001[190]	<65% with intermittent claudication	>=65% with intermittent claudication	5v10	0.60	0.16	0.34	0.77	0.72	0.40	0.68	0.45	0.80	0.62	0.21	0.43
19.1[a] [b]	Lysakowski 2001[191]	heterogeneous population	homogenous population	4v3	0.16	0.29		0.25	0.22		0.17	0.27		0.19	0.29	
20.1	MSAC 2002[192]	<50% men	>=50% men	6v6	0.24	0.16	0.04	0.40	0.45	0.08	0.25	0.24	0.29	0.28	0.30	0.08
20.2	MSAC 2002[192]	families /pedigree	definite/suspected/prenatal	8v4	0.51	0.19	0.34	0.49	0.08	0.41	0.53	0.11	0.59	0.54	0.12	0.34
21.1	Nallamothu 2001[193]	<55 years	>=55 years	5v9	0.48	0.62	0.37	0.35	0.71	0.58	0.54	0.76	0.37	0.44	0.69	0.37
21.2	Nallamothu 2001[193]	<65%men	>=65% men	7v7	0.43	0.57	0.98	0.27	0.06	0.71	0.45	0.94	0.64	0.38	0.79	0.85
22.1	Patwardhan 2004[194]	<70 years	>=70 years	11v5	0.34	0.96	0.42	0.99	0.62	0.67	0.31	0.80	0.72	0.90	0.65	0.43
22.2	Patwardhan 2004[194]	healthy controls	diseased controls	13v6	0.34	0.49	0.39	0.69	0.32	0.12	0.36	0.41	0.48	0.51	0.30	0.26
23.1	Romagnuolo 2003[195]	wide variety of possible diagnoses	stones or cancer diagnoses	11v35	0.16	0.11	0.10	0.17	0.04	0.21	0.15	0.06	0.41	0.18	0.12	0.36
24.1	Sauerland 2004[196]	adults	children	10v3	0.32	0.61	0.77	0.35	0.62	0.40	0.35	0.64	0.51	0.35	0.60	0.92

								Crossing curve								
					Parallel curve models			Q*			ref threshold			comp threshold		
id	study	Comparator group	Reference group	n	M(eq)	M(w)	H	M(eq)	M(w)	H	M(eq)	M(w)	H	M(eq)	M(w)	H
25.1	Sotiriadis 2003[197]	<=30 years	>30 years	4v8	0.04	0.55	0.27	0.49	0.31	0.19	0.06	0.70	0.25	0.06	0.74	0.60
25.2	Sotiriadis 2003[197]	high risk	low risk/routine	7v5	0.63	0.65	0.28	0.50	0.75	0.21	0.89	0.69	0.68	0.61	0.66	0.49
26.1	Varonen 2000[198]	ENT clinic	general clinic	3v4	0.01	0.02	0.01	0.04	0.06	0.11	0.06	0.10	0.22	0.03	0.05	0.05
27.1	Visser 2000[119]	<=60% men	>60% men	8v8	0.39	0.06	0.34	0.23	0.01	0.47	0.52	0.78	0.25	0.26	0.03	0.84
27.2	Visser 2000[119]	<=65 years	>65 years	8v8	0.09	0.10	0.05	0.66	0.02	0.12	0.02	0.03	0.16	0.12	0.32	0.16
27.3	Visser 2000[119]	N America	other country	14v7	0.69	0.37	0.41	0.84	0.41	0.44	0.76	0.34	0.55	0.70	0.39	0.50
28.1	Whitsel 2000[114]	<=40 years	> 40 years	8V8	0.23	0.80	0.47	0.66	0.58	0.25	0.21	0.81	0.48	0.85	0.59	0.57
28.2	Whitsel 2000[114]	<=50% men	>50% men	5v11	0.66	0.96	0.65	0.57	0.61	0.63	0.98	0.65	0.60	0.61	0.77	0.63
28.3	Whitsel 2000[114]	<=50% type 1 diabetes	50-100%	5v10	0.66	0.45	0.58	0.10	0.34	0.14	0.61	0.37	0.40	0.69	0.39	0.42
28.4	Whitsel 2000[114]	mean duration <=10 years	>10 years	10v4	0.45	0.85	0.52	0.57	0.90	0.93	0.46	1.00	0.93	0.51	0.84	0.83
29.1 [a] [b]	Wiese 2000[129]	STD clinic	speciality/general clinic	14v16	0.05	0.03		0.15	0.15		0.03	0.02		0.07	0.04	

[a] denotes covariates for which the parallel curve HSROC analysis could not be completed

[b] denotes covariates for which the crossing curve HSROC analysis could not be completed

Appendix 22 Heterogeneity investigations – P-values for differences in slope and threshold (crossing curve models only)

					difference in slope P-values			difference in theta P-values	
id	Review	Comparator group	Reference group	n	M(eq)	M(w)	H	Hp	Hx
1.1	Balk 2001[180]	hospitalised patients	emergency department patients with symptoms/pain	14v18	0.27	0.22	0.16	0.07	0.74
2.1[a]	Bricker 2000[179]	tertiary care	primary/secondary care	4v7	0.21	0.11	0.35		0.27
2.2[b]	Bricker 2000[179]	2nd trimester	1st, 2nd and 3rd trimester	6v5	0.15	0.09		0.43	
2.3	Bricker 2000[179]	low risk	unselected	4v7	0.23	0.20	0.32	0.36	0.27
3.1	Buchanan 2001[181]	prison release	community/hospital discharges	8v13	0.89	0.32	0.88	0.97	0.95
3.2	Buchanan 2001[181]	time at risk <=20mos	>20 mos	10v8	0.71	0.50	0.57	0.81	0.73
4.1	Chapell 2002[182]	possible age bias	no bias or not reported	4v9	0.04	0.01	0.50	0.29	0.29
4.2	Chapell 2002[182]	possible bias to easy cases	no bias to easy cases	5v8	0.24	0.03	0.31	0.45	0.27
4.3	Chapell[182]	symptoms/presented cases	unspecified diagnosis	8v5	0.56	0.28	0.89	0.28	0.64
5.1	Delgado 2003[178]	unknown primary tumours	other	8v7	0.63	0.63	0.47	0.37	0.70
6.1	Deville 2000[123]	previous surgery	no previous surgery	8v9	0.07	0.28	0.22	0.93	0.93
6.2[b]	Deville 2000[123]	bilateral excluded	bilateral not excluded	3v14	0.37	0.54		0.53	
6.3	Deville 2000[123]	<=60% men	>60% men	10v4	0.44	0.19	0.21	0.62	0.66
7.1[a] [b]	Dijkhuizen 2000[100]	pre and post-menopausal women	post-menopausal women only	22v7	0.21	0.05			
7.2[a] [b]	Dijkhuizen 2000[100]	asymptomatic or both	symptomatic only	20v13	0.09	0.04			
8.1	Eden 2001[183]	environmental exposure	medical/not exposed	3v4	0.40	0.42	0.64	0.91	0.90
9.1	Flemons 2003[184]	home setting	sleep laboratory	13v36	0.98	0.55	0.69	0.59	0.51
9.2	Flemons 2003[184]	<75%men	75-100% men	10v29	0.35	0.26	0.14	0.58	0.48
9.3	Flemons 2003[184]	mean AHI[16]<=30	Ahl>30	15v17	0.48	0.24	0.33	0.04	0.68
9.4	Flemons 2003[184]	mean BMI<=30	BMI>30	9v25	0.02	0.05	0.04	0.64	0.09
10.1	Flobbe 2002[185]	pathology referral	clinical examination or mammography	13v9	0.67	0.54	0.53	0.74	0.72

[16] mean apnoea-hypopnea index

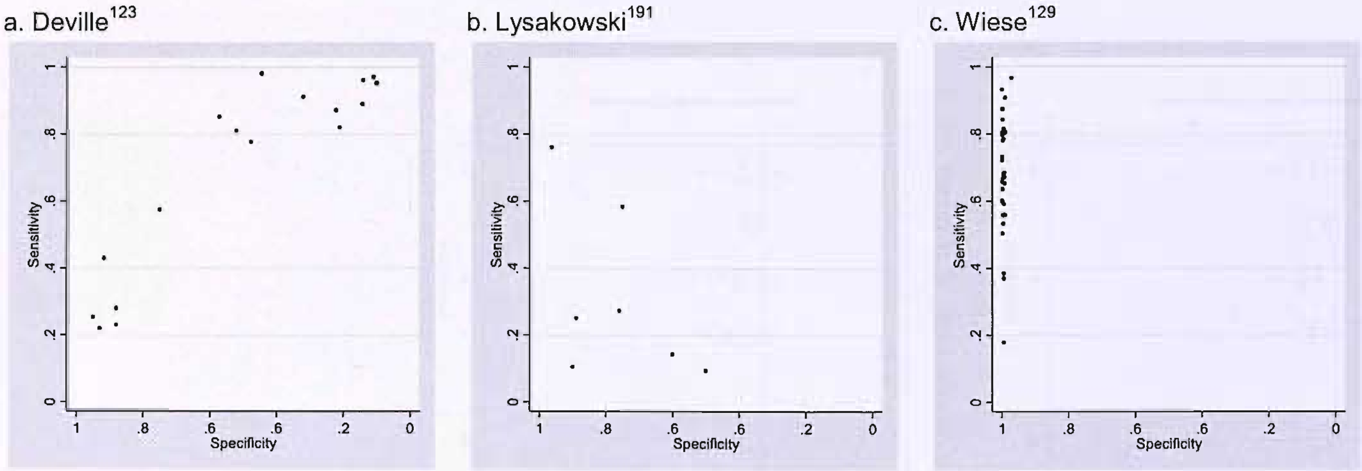
					difference in slope P-values			difference in theta P-values	
id	Review	Comparator group	Reference group	n	M(eq)	M(w)	H	Hp	Hx
11.1	Gifford 2000[186]	age <=70	>70 years	3v8	0.70	0.84	0.61	0.11	0.20
11.2	Gifford 2000[186]	dementia/memory clinics	other setting	5v6	0.22	0.28	0.17	0.74	0.37
11.3	Gifford 2000[186]	diagnostic criteria met	referrals	6v5	0.90	0.77	0.75	0.90	0.98
12.1	Glas 2003[80]	<30% Grade 1 tumours	>=30% Grade 1 tumours	14v6	0.58	0.64	0.95	0.17	0.29
12.2	Glas 2003[80]	<30% Grade 2 tumours	>=30% Grade 2 tumours	6v14	0.02	0.02	0.02	0.96	0.11
12.3	Glas 2003[80]	<30% Grade 3 tumours	>=30% Grade 3 tumours	8v12	0.34	0.36	0.65	0.75	0.59
12.4[b]	Glas 2003[80]	100% urological	rest		0.43	0.42		0.46	
13.1	Gould 2001[107]	>=70% men	<70% men	14v14	0.17	0.30	0.12	0.85	0.17
13.2	Gould 2001[107]	<60years	>=60 years	7v17	0.37	0.44	0.33	0.32	0.33
14.1	Gould 2003[187]	>=70% men	<70% men	12 v10	0.52	0.41	0.30	0.93	0.34
14.2[a b]	Gould 2003[187]	<60 years	>=60 years	4v21	0.26	0.20			
15.1	Gray 2000[188]	suspicion/lesions	cancer history	10v4	0.58	0.90	0.60	0.36	0.81
16.1	Ioannidis[243]	symptoms suggestive of ACI	pts with chest pain	4v6	0.12	0.08	0.17	0.12	0.38
16.2	Ioannidis[243]	<65 years	>=65 years	3v4	0.80	0.84	0.66	0.41	0.70
16.3[b]	Ioannidis[243]	<65% men	>=65% men	3v4	0.04	0.03		0.11	
17.1	Kittler 2002[104]	non-melanocytic lesions excluded	non-melanocytic lesions included	4v9	0.24	0.30	0.03	0.78	0.16
18.1	Koelemay 2001[190]	<65 years	>=65 years	9v7	0.08	0.08	0.77	0.25	0.46
18.2	Koelemay 2001[190]	<70% men	>=70% men	7v11	0.22	0.35	0.36	0.15	0.24
18.3	Koelemay 2001[190]	<65% with intermittent claudication	>=65% with intermittent claudication	5v10	0.86	0.93	0.90	0.47	0.86
19.1[a b]	Lysakowski 2001[191]	heterogeneous population	homogenous population	4v3	0.38	0.33			
20.1	MSAC 2002[192]	<50% men	>=50% men	6v6	0.63	0.72	0.87	0.34	0.54
20.2	MSAC 2002[192]	families /pedigree	definite/suspected/prenatal	8v4	0.77	0.21	0.80	0.81	0.89
21.1	Nallamothu 2001[193]	<55 years	>=55 years	5v9	0.48	0.84	0.72	0.77	0.88
21.2	Nallamothu 2001[193]	<65%men	>=65% men	7v7	0.14	0.07	0.60	0.09	0.32

					difference in slope P-values			difference in theta P-values	
id	**Review**	**Comparator group**	**Reference group**	**n**	**M(eq)**	**M(w)**	**H**	**Hp**	**Hx**
22.1	Patwardhan 2004[194]	<70 years	>=70 years	11v5	0.53	0.58	0.95	0.08	0.38
22.2	Patwardhan 2004[194]	healthy controls	diseased controls	13v6	0.98	0.42	0.26	0.45	0.51
23.1	Romagnuolo 2003[195]	wide variety of possible diagnoses	stones or cancer diagnoses	11v35	0.63	0.19	0.38	0.37	0.29
24.1	Sauerland 2004[196]	adults	children	10v3	1.00	0.69	0.45	0.01	0.86
25.1	Sotiriadis 2003[197]	<=30 years	>30 years	4v8	0.96	0.38	0.33	0.80	0.88
25.2	Sotiriadis 2003[197]	high risk	low risk/routine	7v5	0.57	0.90	0.12	0.26	0.37
26.1	Varonen 2000[198]	ENT clinic	general clinic	3v4	0.69	0.89	0.66	0.20	0.83
27.1	Visser 2000[119]	<=60% men	>60% men	8v8	0.35	0.08	0.48	0.12	0.20
27.2	Visser 2000[119]	<=65 years	>65 years	8v8	0.05	0.07	0.11	0.68	0.11
27.3	Visser 2000[119]	N America	other country	14v7	1.00	0.62	0.63	0.42	0.49
28.1	Whitsel 2000[114]	<=40 years	> 40 years	8V8	0.38	0.61	0.07	0.28	0.36
28.2	Whitsel 2000[114]	<=50% men	>50% men	5v11	0.70	0.57	0.76	0.02	0.23
28.3	Whitsel 2000[114]	<=50% type 1 diabetes	50-100%	5v10	0.10	0.48	0.03	0.62	0.17
28.4	Whitsel 2000[114]	mean duration <=10 years	>10 years	10v4	0.88	0.81	0.85	0.34	0.51
29.1 [a b]	Wiese 2000[129]	STD clinic	speciality/general clinic	14v16	0.30	0.31			

[a] denotes covariates for which the parallel curve HSROC analysis could not be completed

[b] denotes covariates for which the crossing curve HSROC analysis could not be completed

Appendix 23 ROC plots for reviews for which HSROC analyses would not complete



a. Deville[123]

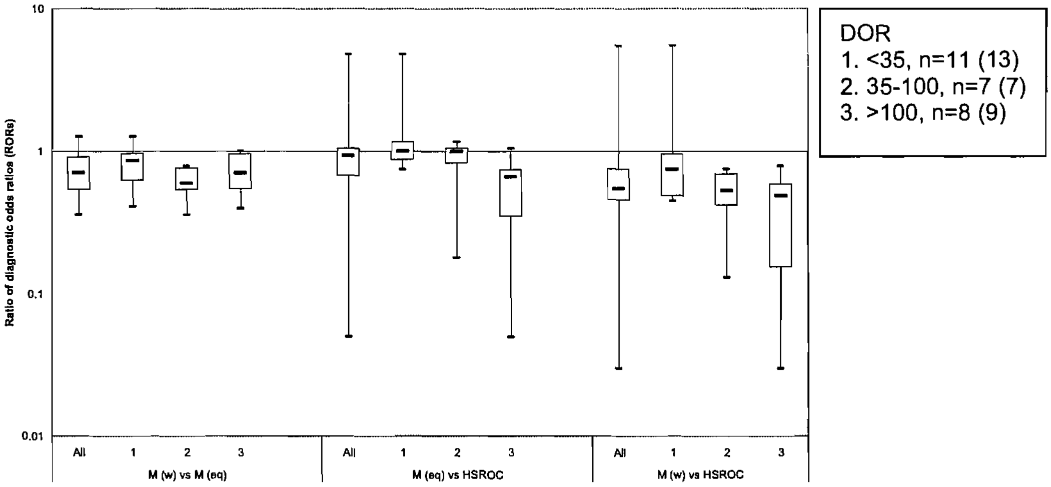
b. Lysakowski[191]

c. Wiese[129]

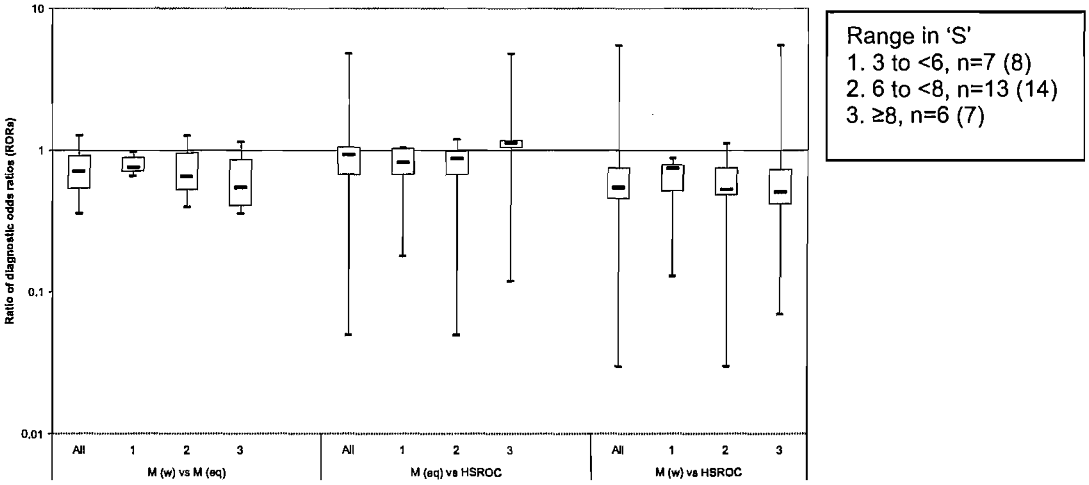
Appendix 24 Box and whisker plots for stratified analyses comparing DORs between models

Box and whisker plots showing ratio of DORs (RORs) at mean threshold between models for stratified analyses: median, interquartile range (box) and range (whiskers). Weighted Moses model is compared to the unweighted Moses model (denominator) and each Moses model is compared to the HSROC model (denominator)

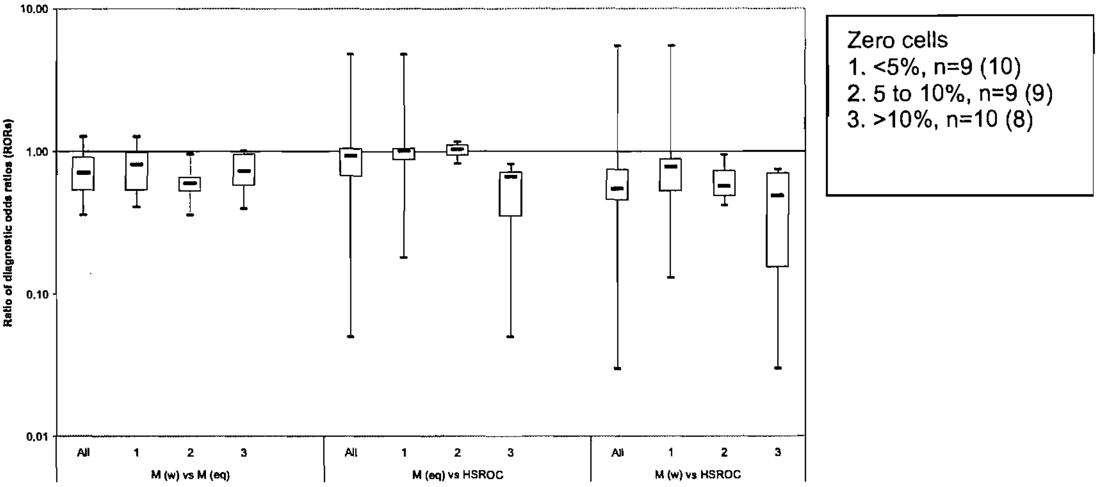
i. size of DOR



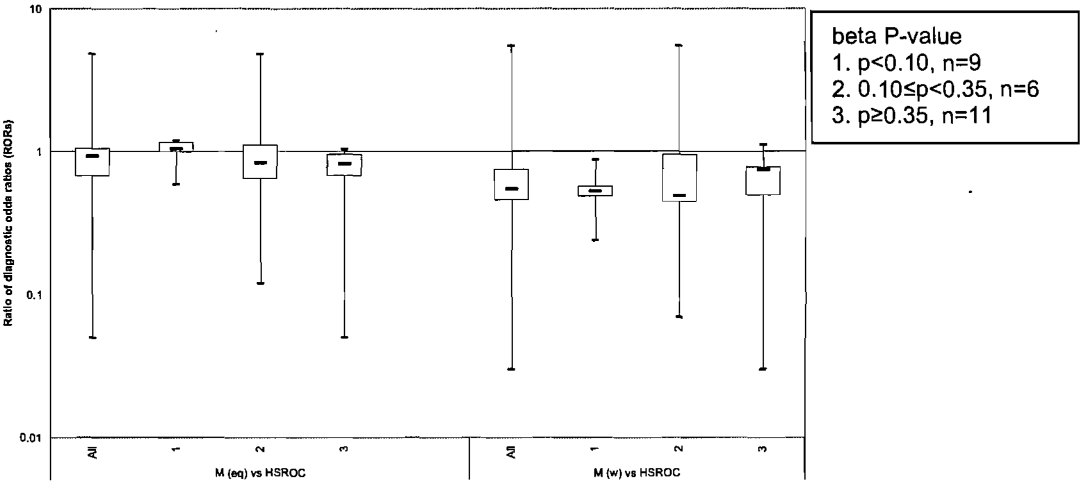
ii. range in 'S' (from Moses model)



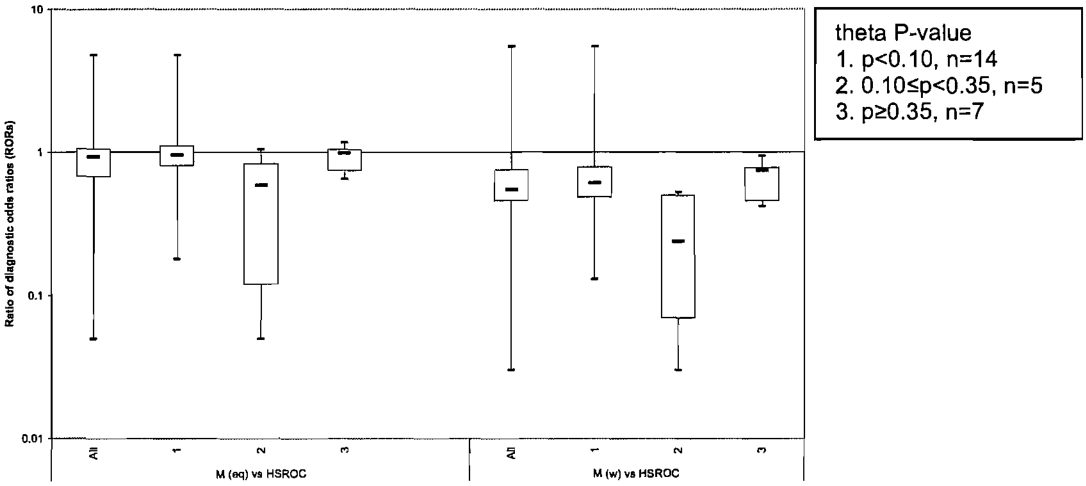
iii. percentage of 2x2 cells with zero values [(no. FP + no. FN)/(no. studies x4)]



iv. degree of asymmetry (based on beta P-value from HSROC model)



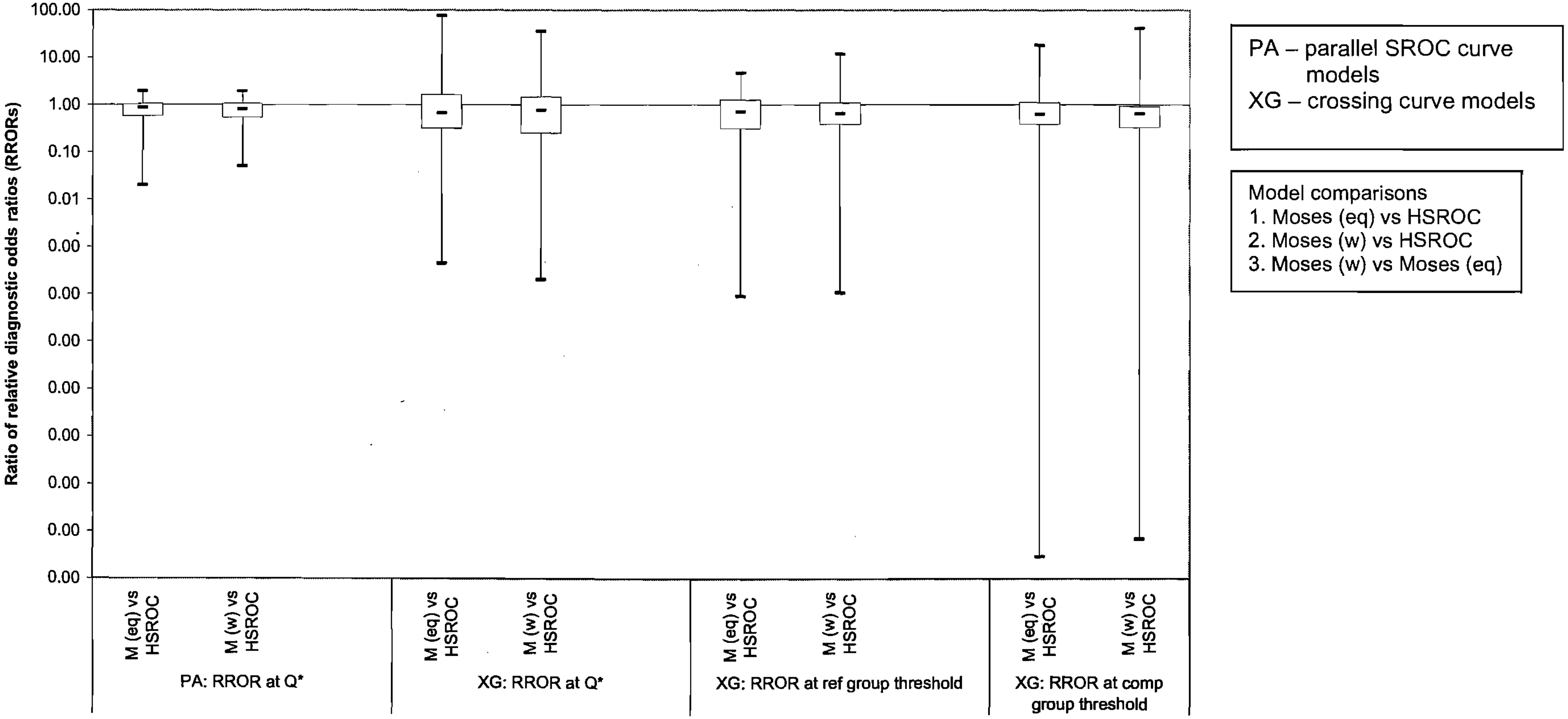
v. importance of threshold (based on theta P-value from HSROC mode)



ROR – ratio of diagnostic odds ratios; Moses (w) – weighted Moses model; Moses (eq) – unweighted Moses model; HSROC – hierarchical SROC model; Q* - point where sensitivity=specificity; mean threshold – operating point estimated using mean threshold across studies

Appendix 25 Box and whisker plots comparing RDORs between models

Box and whisker plots showing ratio of RDORs (RRORs) between models at Q* and at the mean threshold of the reference and comparator groups: median, interquartile range (box) and range (whiskers). Each Moses model is compared against the HSROC model results (denominator) for both the parallel (PA) and crossing (XG) curve versions of the models.



RROR – ratio of relative diagnostic ods ratios between models; Moses (w) – weighted Moses model; Moses (eq) – unweighted Moses model; HSROC – hierarchical SROC model; RROR – ratio of RDORs between models; Q* - point where sensitivity=specificity; ref group threshold – operating point estimated using mean threshold of reference group; comp group threshold – operating point estimated using mean threshold of comparator group

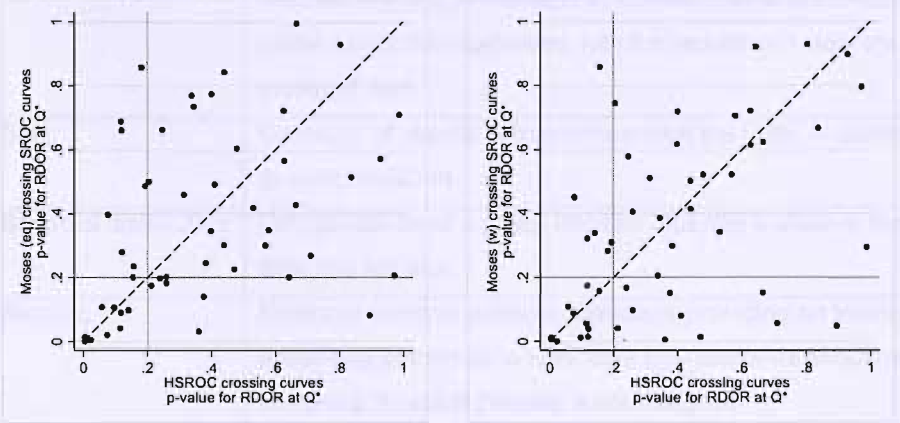
NB: The very extreme ranges, especially for the far right comparison have occurred in reviews with very small numbers of studies in one of the comparator groups leading to very big differences in RDORs between models.

Appendix 26 Comparison of P-values for RDORs between crossing curve models

RDOR at Q*

a. Moses (eq) versus HSROC

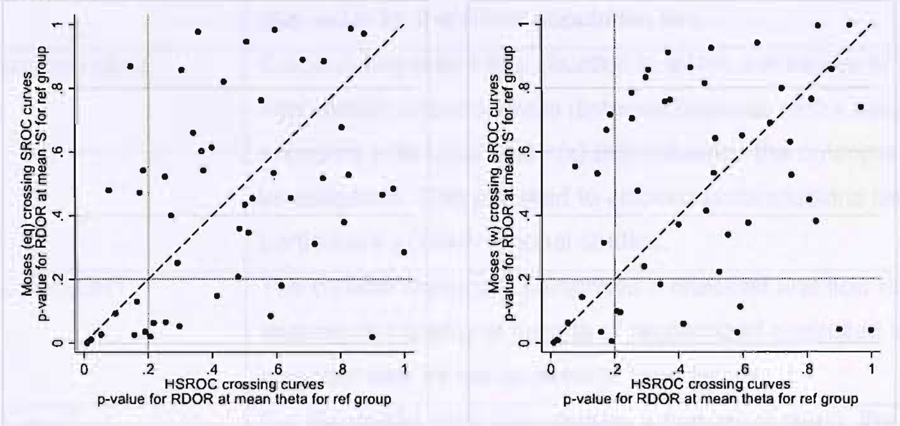
b. Moses (w) versus HSROC



RDOR at average reference group threshold

c. Moses (eq) versus HSROC

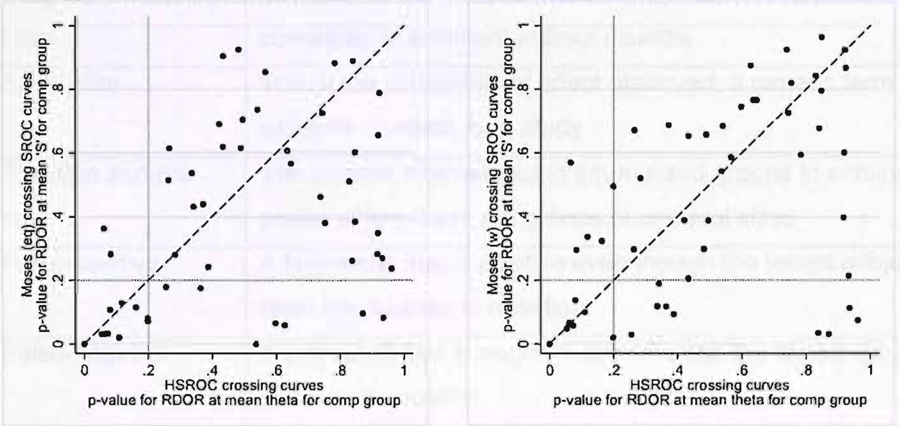
d. Moses (w) versus HSROC



RDOR at average comparator group threshold

e. Moses (eq) versus HSROC

f. Moses (w) versus HSROC



Moses (w) – weighted Moses model; Moses (eq) – unweighted Moses model; HSROC – hierarchical SROC model; RDOR – relative diagnostic odds ratio; Q* - point where sensitivity=specificity; ref group – reference group; comp group – comparator group

Glossary

Accuracy	A general term to describe the discriminative ability of the test or alternatively, the percentage of correct results obtained by a test under evaluation compared with the results of a reference or 'gold standard' test.
Bias	Deviation of results or inferences from the truth, or processes leading to such deviation
Binomial distribution	Categorization of a group into two mutually exclusive subgroups, e.g. Sick and not sick.
Blinding	Refers to whether patients, clinicians providing an intervention, people assessing outcomes, and/or data analysts were aware or unaware of the group to which patients were assigned
Confidence interval	Quantifies the uncertainty in measurement; usually reported as 95% ci, which is the range of values within which we can be 95% sure that the true value for the whole population lies.
Confounding	Confounding refers to a situation in which a measure of the effect of an intervention or exposure is distorted because of the association of exposure with other factor(s) that influence the outcome under investigation. This can lead to erroneous conclusions being drawn, particularly in observational studies.
CONSORT	The consort statement comprises a checklist and flow diagram to help improve the quality of reports of randomized controlled trials. It offers a standard way for researchers to report trials.
Cut-off	For diagnostic tests that produce a numerical result, the point above which test results are classified as positive is called the cut-off.
Diagnostic odds ratio	The ratio of the odds of a positive test result in a patient with disease compared to a patient without disease
Effect size	This is the standardised effect observed; a generic term for the estimate of effect for a study
Effective sample size	The sample size needed in equal-sized groups to achieve the available power where there are groups of unequal sizes
False-positive	A test result that is positive even though the tested subject does not have the disease in question
False-negative	A test result that is negative even though the tested subject has the disease in question
Fixed effect model	A meta-analytic model where only within-study variation is taken to influence the uncertainty of results (as reflected in the confidence interval). Variation between the estimates of effect from each study does not affect the confidence interval.

Heterogeneity	Variability or differences between studies in the estimates of effects.
Individual patient data	The availability of raw data for each study participant in each included study
Likelihood Ratio	The likelihood that a given test result would be expected in a patient with a disease compared to the likelihood that the same result would be expected in a patient without that disease.
Meta-analysis	A method for combining the results of several independent studies that measure the same outcomes so that an overall summary statistic can be calculated.
Odds ratio	Describes the odds of a patient in the experimental group having an event divided by the odds of a patient in the control group having the event
P-value	The probability (ranging from zero to one) that the results observed in a study (or results more extreme) could have occurred by chance.
Polymerase chain reaction	A laboratory technique that can amplify the amount of dna from a tiny sample to a large amount within just a few hours
Predictive value	The probability that a positive/negative result accurately indicates the presence/absence of disease.
Prevalence	The proportion of a given population with a target disorder at a given time
Primary care	Medical care provided by the clinician of first contact for the patient. Typically, the primary care physician is a general practitioner.
Q*	Point on the sroc curve at which sensitivity=specificity
Random effect model	A meta-analytic model in which both within-study sampling error (variance) and between-study variation are included in the assessment of the uncertainty (confidence interval) of the results of a meta-analysis.
Randomized controlled trial	Experiment in which subjects are randomly allocated to receive or not receive an experimental preventive, therapeutic, or diagnostic procedure and then followed to determine the effect.
Reference standard	A method having established or widely accepted accuracy for determining a diagnosis, providing a standard to which a new screening or diagnostic test can be compared. The method need not be a single or simple procedure but could include follow-up of patients to observe the evolution of their conditions or the consensus of an expert panel of clinicians, as is frequently used in the study of psychiatric conditions.
Relative diagnostic odds ratio	Estimate of relative difference in accuracy between two groups of studies

Sensitivity	The sensitivity of a diagnostic or screening test is the proportion of people with a designated disorder who are so identified by the test
Specificity	The specificity of a diagnostic or screening test is the proportion of people free of a designated disorder who are so identified by the text.
Standard deviation	A measure of variability; quantifies how much values vary from each other.
Standard error	A measure of variability; quantifies how accurately the true population mean is known.
STARD	The stard statement comprises a checklist and flow diagram to help improve the quality of reports of diagnostic accuracy studies.
Tests	Any method for obtaining additional information regarding a patient's health status.
Variance	A measure of the average distance between each of a set of data points and their mean value; equal to the sum of the squares of the deviation from the mean value. Describes the spread of a distribution

Sources of definitions

http://www.nature.com/nrmicro/journal/v5/n11_supp/glossary/nrmicro1523.html

http://www.jr2.ox.ac.uk/bandolier/glossary.html

http://www.elsevier.com/framework_products/promis_misc/apmrglossary.pdf

Reference list

1. Bossuyt PM, Irwig L, Craig JC, Glasziou PP. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ* 2006;332:1089-92.

2. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making* 1991;11:88-94.

3. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. *Lancet* 2000;356:1844-7.

4. Griner PF, Mayewski RJ, Mushlin AI, Greenland P. Selection and interpretation of diagnostic tests and procedures. Principles and applications. *Ann Intern Med* 1981;94:557-92.

5. Sackett DL, Haynes RB, Guyatt GH, Tugwell T. *Clinical epidemiology. A basic science for clinical medicine*. London: Little, Brown and Company, 1991.

6. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;299:926-30.

7. Hlatky MA, Pryor DB, Harrell-FE J, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. *Am J Med* 1984;77:64-71.

8. Feinstein AR. *Clinical epidemiology: The architecture of clinical research*. Philadelphia, PA: WB Saunders Co, 1985.

9. Hlatky MA, Daniel MB, Harrell-FE J, Leek KL, Califf RM, Pryor DB. Rethinking sensitivity and specificity. *Am J Cardiol* 1987;59:1195-8.

10. Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. *Clin Chem* 2005;51:1335-41.

11. Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. In: Egger M, Davey Smith G, Altman D, editors. Systematic reviews in health care: Meta analysis in context London: BMJ Books 2001.

12. Mol BW, Bossuyt PMM. Evaluating the effectiveness of diagnostic tests. Tubal subfertility and ectopic pregnancy: evaluating the effectiveness of diagnostic tests 1999.

13. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. *BMJ* 2002;32:669-71.

14. Knottnerus JA, Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. *J Clin Epidemiol* 1992;45:1143-54.

15. Rozanski A, Diamond GA, Berman D, Forrester JS, Morris D, Swan HJ. The declining specificity of exercise radionuclide ventriculography. *N Engl J Med* 1983;309:518-22.

16. Philbrick JT, Horwitz RI, Feinstein AR, Langou RA, Chandler JP. The limited spectrum of patients studied in exercise test research. Analyzing the tip of the iceberg. *JAMA* 1982;248:2467-70.

17. Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, SCHWARTZ JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection . *Ann Intern Med* 1992;117:135-40.

18. Moons KG, van Es GA, Deckers JW, Habbema JD, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and bayes' theorem in assessing diagnostic probabilities: a clinical example . *Epidemiology* 1997;8:12-7.

19. Goehring C, Perrier A, Morabia A. Spectrum bias: a quantitative and graphical analysis of the variability of medical diagnostic test performance. *Stat Med* 2004;23:125-35.

20. Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation.. *Ann Intern Med* 2002;137:598-602.

21. Whiting P, Rutjes A, Dinnes J, Reitsma J, Bossuyt P, and Kleijnen J Development and validation of methods for assessing the quality and reporting of diagnostic studies. Health Technol Assess 8[25], 1-234. 2004.

22. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet* 2005;365:176-86.

23. Cook DI, Gebski VJ, Keech AC. Subgroup analysis in clinical trials. *Medical Journal of Australia* 2004;180:289-91.

24. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med* 2002;21:2917-30.

25. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials.. *Lancet* 2000;355:1064-9.

26. Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, CHALMERS TC *et al.* Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med* 1994;120:667-76.

27. Jaeschke R, Guyatt G, Sackett DL. Users guides to the medical literature .3. How to use an article about a diagnostic-test .a. Are the results of the study valid. *JAMA* 1994;271:389-91.

28. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. *JAMA* 1995;274:645-51.

29. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM *et al.* Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ* 2003;326:41-4.

30. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM *et al.* The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem* 2003;49:7-18.

31. Moher D. CONSORT: an evolving tool to help improve the quality of reports of randomized controlled trials. *JAMA* 1998;279:1489-91.

32. Ginsberg JS, Wells PS, Kearon C, Anderson D, Crowther M, Weitz JI *et al.* Sensitivity and specificity of a rapid whole-blood assay for D-dimer in the diagnosis of pulmonary embolism . *Ann Intern Med* 1998;129:1006-11.

33. Morise AP, Diamond GA. Comparison of the sensitivity and specificity of exercise electrocardiography in biased and unbiased populations of men and women. *Am Heart J* 1995;130:741-7.

34. Weiner DA, Ryan TJ, McCabe CH, Kennedy JW, Schloss M, Tristani F *et al.* Exercise stress testing. Correlations among history of angina, ST-segment response and prevalence of coronary-artery disease in the Coronary Artery Aurgery Study (CASS). *N Engl J Med* 1979;301:230-5.

35. Miller WC, Hoffman IF, Owen-O'Dowd J, McPherson JT, Privette A, Schmitz JL *et al.* Selective screening for chlamydial infection: which criteria to use? *Am J Prev Med* 2000;18:115-22.

36. Banks E. Hormone replacement therapy and the sensitivity and specificity of breast cancer screening: a review. *J Med Screen* 2001;8:29-35.

37. Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 [updated September 2006]. In: Higgins JP, Green S, editors. The Cochrane Library Chichester, UK: John Wiley & Sons, Ltd. 2006.

38. Haynes RB, Wilczynski NL. Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. *BMJ* 2004;328:1040.

39. Doust JA, Pietrzak E, Sanders S, Glasziou PP. Identifying studies for systematic reviews of diagnostic tests was difficult due to the poor sensitivity and precision of methodologic filters and the lack of information in the abstract. *J Clin Epidemiol* 2005;58:444-9.

40. Deville WL, Buntinx F. Guidelines for conducting systematic reviews of studies evaluating the accuracy of diagnostic tests. In: Knottnerus JA, ed. The Evidence Base of Clinical Diagnosis London: BMJ Books 2002.

41. Glasziou P, Irwig L, Bain C, Colditz G. Diagnostic tests. Systematic reviews in health care: A practical guide Cambridge: Cambridge University Press 2001.

42. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. *J Clin Epidemiol* 1995;48:119-30.

43. van der Schouw YT, Verbeek AL, Ruijs SH. Guidelines for the assessment of new diagnostic tests. *Invest Radiol* 1995;30:334-40.

44. Shapiro DE. Issues in combining independent estimates of the sensitivity and specificity of a diagnostic test. *Acad Radiol* 1995;2.

45. Simel D, Samsa G, Matchar D. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol* 1991;44:763.

46. Sheps SB, Schechter MT. The assessment of diagnostic tests: a survey of current medical research. 1984;252:2418-22.

47. Arrol BA, Schechter MT, Sheps SB. The assessment of diagnostic tests: a comparison of medical literature in 1982 and 1985. *J Gen Intern Med* 1995;3:443-7.

48. Siddiqui MA, Azuara-Blanco A, Burr J. The quality and reporting of diagnostic accuracy studies published in opthalmic journals. *Br J Opthalmol* 2005;89:261-5.

49. Heffner JE, Feinstein D, Barbieri C. Methodologic standards for diagnostic test research in pulmonary medicine . *Chest* 1998;114:877-85.

50. Harper R, Reeves B. Compliance with methodological standards when evaluating ophthalmic diagnostic tests. *Invest Ophthalmol Visual Sci* 1999;40:1650-7.

51. Rothwell PM, Pendlebury ST, Wardlaw J, Warlow CP. Critical appraisal of the design and reporting of studies of imaging and measurement of carotid stenosis. *Stroke* 2000;31:1444-50.

52. Lumbreras-Lacarra B, Ramos-Rincon JM, Hernandez-Aguado I. Methodology in diagnostic laboratory test research in *Clinical Chemistry* and *Clinical Chemistry and Laboratory Medicine. Clin Chem* 2004;50:530-6.

53. Lambert PC, Sutton AJ, Abrams KR, Jones DR. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *J Clin Epidemiol* 2002;55:86-94.

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

55. Rutter CM, Gatsonis CA. Regression methods for meta-analysis of diagnostic test data. *Acad Radiol* 1995;2:s48-s56.

56. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med* 2001;20:2865-84.

57. van Houwelingen HC, Zwinderman KH, Stijnen T. A Bivariate Approach to Metaanalysis. *Stat Med* 1993;12:2273-84.

58. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat Med* 2002;21:589-624.

59. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology* 2005;58:982-90.

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

61. Walter SD, Irwig L, Glasziou PP. Meta-analysis of diagnostic tests with imperfect reference standards. *J Clin Epidemiol* 1999;52:943-51.

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

63. Lijmer JG, Bossuyt PM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Stat Med* 2002;21:1525-37.

64. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JHP *et al.* Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282:1061-6.

65. Rutjes A, Reitsma J, Di Nisio M, Smidt N, Zwinderman AH, Rijn JC *et al.* Bias in diagnostic accuracy studies due to shortcomings in design and conduct. 2003. Xth Annual Cochrane Colloquium, Barcelona.

66. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med* 2004;140:189-202.

67. Begg CB, McNeil BJ. Assessment of radiologic tests: control of bias and other design considerations. *Radiology* 1988;167:565-9.

68. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ* 1998;317:1185-90.

69. Fletcher RH, Fletcher SW, Wagner EH. Studying cases. Clinical Epidemiology: The Essentials London: Williams & Wilkins 1996. p.208-27.

70. Begg CB. Biases in the assessment of diagnostic tests. *Stat Med* 1987;6:411-23.

71. Diamond GA. Reverend Bayes' silent majority. An alternative factor affecting sensitivity and specificity of exercise electrocardiography. *Am J Cardiol* 1986;57:1175-80.

72. Heffner JE. Evaluating diagnostic tests in the pleural space. Differentiating transudates from exudates as a model. *Clin Chest Med* 1998;19:277-93.

73. Macaskill P. Interpretation of S, the proxy for test threshold in the Moses model. Contributions to the analysis and meta-analysis of diagnostic test comparisons. School of Public Health, University of Sydney; 2003. p. 163-71.

74. Dinnes J, Deeks JJ, Kirby J, Roderick P. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. *Health Technol Assess* 2005;9.

75. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Med Decis Making* 1993;13:313-21.

76. Scheidler J, Hricak H, Yu KK, Subak L, Segal MR. Radiological evaluation of lymph node metastases in patients with cervical cancer. A meta-analysis. *JAMA* 1997;278:1096-101.

77. Deville W, Yzermans N, Bouter LM, Bezemer PD, and van der Windt DA Heterogeneity in systematic reviews of diagnostic studies. 1999 Oxford 2nd Symposium on Systematic Reviews: Beyond the Basics.

78. Dinnes J, Deeks J, Kunst H, Gibson A, Cummins E, Waugh N, *et al.* A systematic review of diagnostic tests for the detection of tuberculosis. Health Technol Assess 2007;11(3).

79. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics* 2007;8:239-51.

80. Glas AS, Roos D, Deutekom M, Zwinderman AH, Bossuyt PM, Kurth KH. Tumor markers in the diagnosis of primary bladder cancer: a systematic review. *J Urol* 2003;169:1975-82.

81. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol* 2005;58:882-93.

82. Badgett RG, Lucey CR, Mulrow CD. Can the clinical examination diagnose left-sided heart failure in adults? *JAMA* 1997;277:1712-9.

83. Anand SS, Wells PS, Hunt D, Brill-Edwards P, Cook D, Ginsberg JS. Does this patient have deep vein thrombosis? *JAMA* 1998;279:1094-9.

84. D'Arcy CA, McGee S. Does this patient have carpal tunnel syndrome? *JAMA* 2000;283:3110-7.

85. Koumans EH, Johnson RE, Knapp JS, Louis ME. Laboratory testing for neisseria gonorrhoeae by recently introduced nonculture tests: a performance review with clinical and public health considerations. *Clin Infect Dis* 1998;27:1171-80.

86. Nelson H, Helfand M. Screening for chlamydial infection. *Am J Prev Med* 2001;20:95-41.

87. Owens DK, Holodniy M, Garber AM, Scott J, Sonnad S, Moses L *et al.* Polymerase chain reaction for the diagnosis of HIV infection in adults. A meta-analysis with recommendations for clinical practice and study design. *Ann Intern Med* 1996;124:803-15.

88. Adams E. Positron emission tomography: systematic review. PET as a diagnostic test in lung cancer. Technology Assessment Program. Boston (MA): Veterans Affairs Medical Center, Health Services Research and Development Service; Technology Assessment Program PET Report; A6 1996.

89. US Preventive Services Task Force. *Guide to Clinical Preventive Services: Report of the US Preventive Services Task Force.* Washington DC: Office of Disease Prevention and Health Promotion, U.S. Government Printing Offfice, 1996.

90. Badgett RG, Mulrow CD, Otto PM, Ramirez G. How well can the chest radiograph diagnose left ventricular dysfunction. *J Gen Intern Med* 1996;11:625-34.

91. de Vries SO, Hunink MG, Polak JF. Summary receiver operating characteristic curves as a technique for meta-analysis of the diagnostic performance of duplex ultrasonography in peripheral arterial disease. *Acad Radiol* 1996;3:361-9.

92. Oosterhuis WP, Niessen RW, Bossuyt PM, Sanders GT, Sturk A. Diagnostic value of the mean corpuscular volume in the detection of vitamin B12 deficiency. *Scand J Clin Lab Invest* 2000;60:9-18.

93. de Bruyn G, Graviss E. A systematic review of the diagnostic accuracy of physical examination for the detection of cirrhosis. *BMC Medical Informatics and Decision Making* 2001;1-11.

94. De Bernardinis M, Violi V, Roncoroni L, Boselli AS, Giunta A, Peracchia A. Discriminant power and information content of Ranson's prognostic signs in acute pancreatitis: a meta-analytic study. *Crit Care Med* 1999;27:2272-83.

95. Kwok Y, Kim C, Grady D, Segal M, Redberg R. Meta-analysis of exercise testing to detect coronary artery disease in women. *Am J Cardiol* 1999;83:660-6.

96. Berry E, Kelly S, Westwood ME, Davies LM, Gough MJ, Bamford JM *et al.* The cost-effectiveness of magnetic resonance angiography for carotid artery stenosis and peripheral vascular disease: a systematic review. *Health Technology Assessment* 2002;6:1-155.

97. Leitich H, Egarter C, Kaider A, Hohlagschwandtner M, Berghammer P, Husslein P. Cervicovaginal fetal fibronectin as a marker for preterm delivery: a meta-analysis. *Am J Obstet Gynecol* 1999;180:1169-76.

98. Cher D, Conwell J, Mandell J. MRI for detecting silicone breast implant rupture: Meta-analysis and implications. *Annals of Plastic Surgery* 2001;47:367-80.

99. Chien PFW, Khan KS, Ogston S, Owen P. The diagnostic accuracy of cervico-vaginal fetal fibronectin in predicting preterm delivery: an overview. *British Journal of Obstetrics and Gynaecology* 1997;104:436-44.

100. Dijkhuizen FP, Mol BW, Brolmann HA, Heintz AP. The accuracy of endometrial sampling in the diagnosis of patients with endometrial carcinoma and hyperplasia: a meta-analysis. *Cancer* 2000;89:1765-72.

101. Kearon C, Julian JA, Newman TE, Ginsberg JS. Noninvasive diagnosis of deep venous thrombosis. *Ann Intern Med* 1998;128:663-77.

102. Kowalski J, Tu XM, Jia G, Pagano M. A comparative meta-analysis on the variability in test performance among FDA-licensed enzyme immunosorbent assays for HIV antibody testing. *J Clin Epidemiol* 2001;54:448-61.

103. Carlson KJ, Skates SJ, Singer DE. Screening for ovarian cancer. *Ann Intern Med* 1994;121:124-32.

104. Kittler H, Pehamberger H, Wolff K, Binder M. Diagnostic accuracy of dermoscopy. *Lancet Oncol* 2002;3:159-65.

105. Tugwell P, Dennis DT, Weinstein A, Wells G, Shea B, Nichol G *et al.* Laboratory evaluation in the diagnosis of Lyme disease: clinical guideline, part 2. *Ann Intern Med* 1997;127:1109-23.

106. Berger M, Velden JJIM, Lijmer J, de K, Prins A, Bohnen A. Abdominal symptoms: Do they predict gallstones? A systematic review. *Scand J Gastroenterol* 2000;35:70-6.

107. Gould M, Maclean C, Kuschner W, Rydzak C, Owens D. Accuracy of positron emission tomography for diagnosis of pulmonary nodules and mass lesions: a meta-analysis. *JAMA* 2001;285:914-24.

108. Buntinx F, Wauters H. The diagnostic value of macroscopic haematuria in diagnosing urological cancers: a meta-analysis. *Fam Pract* 1997;14:63-8.

109. Fleischmann KE, Hunink MG, Kuntz KM, Douglas PS. Exercise echocardiography of exercise SPECT imaging: a meta analysis of diagnostic test performance. *JAMA* 1998;280:913-20.

110. Law J, Boyle J, Harris F, Harkness A, Nye C. Screening for speech and language delay: a systematic review of the literature. *Health Technol Assess* 1998;2(9):1-184.

111. Orr RK, Porter D, Hartman D. Ultrasonography to evaluate adults for appendicitis: decision making based on meta-analysis and probabilistic reasoning. *Acad Emerg Med* 1995;2:644-50.

112. Scouller K, Conigrave KM, Macaskill P, Irwig L, Whitfield JB. Should we use carbohydrate-deficient transferrin instead of gamma-glutamyltransferase for detecting problem drinkers? A systematic review and metaanalysis. *Clin Chem* 2000;46:1894-902.

113. Gianrossi R, Detrano R, Colombo A, Froelicher V. Cardiac fluoroscopy for the diagnosis of coronary artery disease: a meta analytic review. *Am Heart J* 1990;120:1179-88.

114. Whitsel EA, Boyko EJ, Siscovick DS. Reassessing the role of QTc in the diagnosis of autonomic failure among patients with diabetes: a meta-analysis. *Diabetes Care* 2000;23:241-7.

115. Peters AL, Davidson MB, Schriger DL, Hasselblad V. A clinical approach for the diagnosis of diabetes mellitus: an analysis using glycosylated hemoglobin levels. Meta-analysis Research Group on the Diagnosis of Diabetes Using Glycated Hemoglobin Levels [published erratum appears in JAMA 1997 Apr 9;277(14):1125]. *JAMA* 1996;276:1246-52.

116. Hoffman RM, Clanon DL, Littenberg B, Frank JJ, Peirce JC. Using the free-to-total prostate-specific antigen ratio to detect prostate cancer in men with nonspecific elevations of prostate-specific antigen levels. *J Gen Intern Med* 2000;15:739-48.

117. Kinkel K, Hricak H, Lu Y, Tsuda K, Filly RA. US characterization of ovarian masses: a meta-analysis. *Radiology* 2000;217:803-11.

118. Loy CT, Irwig LM, Katelaris PH, Talley NJ. Do commercial serological kits for Helicobacter pylori infection differ in accuracy? A meta-analysis. *Am J Gastroenterol* 1996;91:1138-44.

119. Visser K, Hunink MG. Peripheral arterial disease: gadolinium-enhanced MR angiography versus color-guided duplex US--a meta-analysis. *Radiology* 2000;216:67-77.

120. Mol BW, Bairam N, Lijmer JG, Wiegerinck MA, Bongers MY, van Der Veen F *et al.* The accuracy of CA-125 in the diagnosis of endometriosis: a meta-analysis. *Fertil Steril* 1998;70:1101-8.

121. Kim C, Kwok Y, Heagerty P, Redberg R. Pharmacologic stress testing for coronary disease diagnosis: A meta-analysis. *Am Heart J* 2001;142:934-44.

122. McCrory D, Matchar D, Bastian L, Datta S, Hasselblad V, Hickey J et al. Evaluation of cervical cytology. Rockville, MD: Agency for Healthcare Research and Quality; 1999. Evidence Report/Technology Assessment No. 36

123. Deville WL, van der Windt DA, Dzaferagic A, Bezemer PD, Bouter LM. The test of Lasegue: systematic review of the accuracy in diagnosing herniated discs. *Spine* 2000;25:1140-7.

124. Di Fabio RP. Meta-analysis of the sensitivity and specificity of platform posturography. *Arch Otolaryngol Head Neck Surg* 1996;122:150-6.

125. Fahey MT, Irwig L, Macaskill P. Meta-analysis of Pap test accuracy. *A J Epidemiol* 1995;141:680-9.

126. Hurley JC. Concordance of endotoxemia with gram-negative bacteremia. A meta-analysis using receiver operating characteristic curves. *Arch Pathol Lab Med* 2000;124:1157-64.

127. Nanda K, McCrory DC, Myers ER, Bastian LA, Hasselblad V, Hickey JD *et al.* Accuracy of the Papanicolaou test in screening for and follow-up of cervical cytologic abnormalities: a systematic review. *Ann Intern Med* 2000;132:810-9.

128. Owens DK, Holodniy M, McDonald TW, Scott J, Sonnad S. A meta-analytic evaluation of the polymerase chain reaction for the diagnosis of HIV infection in infants. *JAMA* 1996;275:1342-8.

129. Wiese W, Patel SR, Patel SC, Ohl CA, Estrada CA. A meta-analysis of the Papanicolaou smear and wet mount for the diagnosis of vaginal trichomoniasis. *Am J Med* 2000;108:301-8.

130. Shaheen AAM, Myers RP. Diagnostic accuracy of the aspartate aminotransferaseto-platelet ratio index for the prediction of hepatitis c-related fibrosis: A systematic review. *Hepatology* 2007;46:912-21.

131. Thangaratinam S, Daniels J, Ewer AK, Zamora J, Khan KS. Accuracy of pulse oximetry in screening for congenital heart disease in asymptomatic newborns: a systematic review. *Archives of Disease in Childhood Fetal & Neonatal Edition* 2007;92:F176-80.

132. Bipat S, Glas AS, van d, V, Zwinderman AH, Bossuyt PM, Stoker J. Computed tomography and magnetic resonance imaging in staging of uterine cervical carcinoma: a systematic review. *Gynecologic Oncology* 2003;91:59-66.

133. Bipat S, Glas AS, Slors FJM, Zwinderman AH, Bossuyt PMM, Stoker J. Rectal cancer: Local staging and assessment of lymph node involvement with endoluminal US, CT, and MR imaging - A meta-analysis. *Radiology* 2004;232:773-83.

134. Bipat S, van Leeuwen MS, Comans EFI, Pijl MEJ, Bossuyt PMM, Zwinderman AH *et al.* Colorectal liver metastases: CT, MR imaging, and PET for diagnosis - Meta-analysis. *Radiology* 2005;237:123-31.

135. van Westreenen HL, Westerterp M, Bossuyt PMM, Pruim J, Sloof GW, van Lanschot JJB *et al.* Systematic review of the staging performance of F-18-fluorodeoxyglucose positron emission tomography in esophageal cancer. *J Clin Oncol* 2004;22:3805-12.

136. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology* 2003;3:25.

137. Bipat S, Phoa SS, van Delden OM, Bossuyt PM, Gouma DJ, Lameris JS *et al.* Ultrasonography, computed tomography and magnetic resonance imaging for diagnosis and determining resectability of pancreatic adenocarcinoma: a meta-analysis. *Journal of Computer Assisted Tomography* 2005;29:438-45.

138. Koelemay MJ, Nederkoorn PJ, Reitsma JB, Majoie CB. Systematic review of computed tomographic angiography for assessment of carotid artery disease. *Stroke* 2004;35:2306-12.

139. Whiting P, Harbord R, Main C, Deeks JJ, Filippini G, Egger M *et al.* Accuracy of magnetic resonance imaging for the diagnosis of multiple sclerosis: systematic review. *British Medical Journal* 2006;332:875-8.

140. Williams GJ, Macaskill P, Chan SF, Karplus TE, Yung W, Hodson EM *et al.* Comparative accuracy of renal duplex sonographic parameters in the diagnosis of renal artery stenosis: paired and unpaired analysis. *AJR* 2007;188:798-811.

141. Halligan S, Altman DG, Taylor SA, Mallett S, Deeks JJ, Bartram CI *et al.* CT colonography in the detection of colorectal polyps and cancer: Systematic review meta-analysis and proposed minimum data set for study level reporting. *Radiology* 2005;237:893-904.

142. Kwee TC, Kwee RM. MR angiography in the follow-up of intracranial aneurysms treated with Guglielmi detachable coils: systematic review and meta-analysis. *Neuroradiology* 2007;49:703-13.

143. Steurer J, Fischer JE, Bachmann LM, Koller M, ter Riet G. Communicating accuracy of tests to general practitioners: a controlled study. *BMJ* 2002;324:824-6.

144. Deeks JJ, Higgins JPT, Altman DG. Analysing and presenting results. In: Alderson P, Higgins JPT, Altman DG, editors. Cochrane Reviewers' Handbook 4.2.2 [updated November 2004] http://www.cochrane.org/resources/handbook/hbook.htm (accessed April 2005) 2004.

145. Dickersin K, Min YI, Meinert CL. Factors influencing publication of research results. Follow-up of applications submitted to two institutional review boards. *JAMA* 1992;267:374-8.

146. Deeks JJ, Altman DG, Bradburn MJ. Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In: Egger M, Davey Smith G, Altman D, editors. Systematic reviews in health care: Meta analysis in context London: BMJ Books 2001. p.285-312.

147. Deville WL, Buntinx F. Guidelines for conducting systematic reviews of studies evaluating the accuracy of diagnostic tests. In: Knottnerus JA, ed. The Evidence Base of Clinical Diagnosis London: BMJ Books 2002.

148. Harbord R, Whiting P, Sterne JAC, Egger M, Deeks JJ, Shang A *et al.* An empirical comparison of methods for meta-analysis of diagnostic accuracy studies. *J Clin Epidemiol* 2008;in press.

149. Sterne JA, Egger M, Smith GD. Systematic reviews in health care: Investigating and dealing with publication and other biases in meta-analysis. *BMJ* 2001;323:101-5.

150. Sutton AJ, Higgins JPT. Recent developments in meta-analysis. *Stat Med* 2008;27:625-50.

151. Devallois A, Legrand E, Rastogi N. Evaluation of Amplicor MTB test as adjunct to smears and culture for direct detection of Mycobacterium tuberculosis in the French Caribbean. *J Clin Microbiol* 1996;34:1065-8.

152. Chedore P, Jamieson FB. Routine use of the Gen-Probe MTD2 amplification test for detection of Mycobacterium tuberculosis in clinical specimens in a large public health mycobacteriology laboratory. *Diagn Microbiol Infect Dis* 1999;35:185-91.

153. Wang SX, Tay L. Evaluation of three nucleic acid amplification methods for direct detection of Mycobacterium tuberculosis complex in respiratory specimens. *J Clin Microbiol* 1999;37:1932-4.

154. Smith MB, Bergmann JS, Onoroto M, Mathews G, Woods GL. Evaluation of the enhanced amplified Mycobacterium tuberculosis direct test for direct detection of Mycobacterium tuberculosis complex in respiratory specimens. *Arch Pathol Lab Med* 1999;123:1101-3.

155. Abu-Amero KK. Potential for the use of Polymerase Chain Reaction (PCR) in the detection and identification of Mycobacterium tuberculosis complex in sputum samples. *Mol Biol Today* 2002;3:39-42.

156. Yam WC, Yuen KY, Seto WH. Direct detection of Mycobacterium tuberculosis in respiratory specimens using an automated DNA amplification assay and a single tube nested polymerase chain reaction (PCR). *Clin Chem Lab Med* 1998;36:597-9.

157. Eing BR, Becker A, Sohns A, Ringelmann R. Comparison of Roche Cobas Amplicor Mycobacterium tuberculosis assay with in-house PCR and culture for detection of M. tuberculosis. *J Clin Microbiol* 1998;36:2023-9.

158. Larocco M, Wanger A, Macias E. Evaluation of the Gen-Probe amplified Mycobacterium tuberculosis direct test. *33rd Interscience Conference on*

Antimicrobial Agents and Chemotherapy 1993;33rd Interscience Conference on Antimicrobial Agents and Chemotherapy, New Orleans, Louisiana, USA, October#17-20, #1993; Program and Abstracts of the Interscience Conference on Antimicrobial Agents and Chemotherapy#33:180.

159. Hoffner SE, Cristea M, Klintz L, Petrini B, Kallenius G. RNA amplification for direct detection of mycobacterium tuberculosis in respiratory samples. *Scand J Infect Dis* 1996;28:59-61.

160. Piersimoni C, Scarparo C, Piccoli P, Rigon A, Ruggiero G, Nista D *et al.* Performance assessment of two commercial amplification assays for direct detection of Mycobacterium tuberculosis complex from respiratory and extrapulmonary specimens. *J Clin Microbiol* 2002;40:4138-42.

161. Al Zahrani K, Al Jahdali H, Poirier L, Rene P, Gennaro ML, Menzies D. Accuracy and utility of commercially available amplification and serologic tests for the diagnosis of minimal pulmonary tuberculosis. *Am J Respir Crit Care Med* 2000;162:1323-9.

162. Vuorinen P, Miettinen A, Vuento R, Hallstrom O. Direct detection of Mycobacterium tuberculosis complex in respiratory specimens by Gen-Probe Amplified Mycobacterium Tuberculosis Direct Test and Roche Amplicor Mycobacterium Tuberculosis Test. *J Clin Microbiol* 1995;33:1856-9.

163. dos Anjos Filho L, Oelemann W, Barreto CE, Kritski AL, de Souza Fonseca L. Sensitivity of AMPLICOR MTB on direct detection of Mycobacterium tuberculosis in smear-negative specimens from outpatients in Rio de Janeiro. *Braz J Microbiol* 2002;33:163-5.

164. Sato K, Tomioka H, Kawahara S, Shishido S. Evaluation of two commercial diagnostic kits for Mycobacterium tuberculosis completely based on bacterial DNA and rRNA amplification for direct detection of tubercle bacilli in sputum specimens. *Kansenshogaku Zasshi* 1998;72:504-11.

165. Gomez-Pastrana D, Torronteras R, Caro P, Anguita ML, Lopez-Barrio AM, Andres A *et al.* Comparison of amplicor, in-house polymerase chain reaction, and conventional culture for the diagnosis of tuberculosis in children. *Clin Infect Dis* 2001;32:17-22.

166. Osumi M, Toyoda T, Kawashiro T, Aoyagi T. Detection of Mycobacterium tuberculosis in clinical specimens other than sputum by a specific DNA probe with amplification of the ribosomal RNA. *Kansenshogaku Zasshi* 1995;69:1376-82.

167. Middleton AM, Cullinan P, Wilson R, Kerr JR, Chadwick MV. Interpreting the results of the amplified Mycobacterium tuberculosis direct test for detection of M. tuberculosis rRNA. *J Clin Microbiol* 2003;41:2741-3743.

168. La Rocco MT, Wanger A, Ocera H, Macias E. Evaluation of a commercial rRNA amplification assay for direct detection of Mycobacterium tuberculosis in processed sputum. *Eur J Clin Microbiol Infect Dis* 1994;13:726-31.

169. Deeks JJ. Using evaluation of diagnostic tests: Understanding their limitations and making the most of available evidence. *Ann Oncol* 1999;10:761-8.

170. Alcala L, Ruiz-Serrano MJ, Hernangomez S, Marin M, de Viedma DG, San Juan R *et al.* Evaluation of the upgraded amplified Mycobacterium tuberculosis direct test (gen-probe) for direct detection of Mycobacterium tuberculosis in respiratory and non-respiratory specimens. *Diagn Microbiol Infect Dis* 2001;41:51-6.

171. Kambashi B, Mbulo G, McNerney R, Tembwe R, Kambashi A, Tihon V *et al.* Utility of nucleic acid amplification techniques for the diagnosis of pulmonary tuberculosis in sub-Saharan Africa. *Int J Tuberc Lung Dis* 2001;5:364-9.

172. Mitarai S, Tanoue S, Sugita C, Sugihara E, Tamura A, Nagono Y *et al.* Potential use of Amplicor PCR kit in diagnosing pulmonary tuberculosis from gastric aspirate. *J Microbiol Methods* 2001;47:339-44.

173. Abe C, Hirano K, Wada M, Kazumi Y, Takahashi M, Fukasawa Y *et al.* Detection of Mycobacterium tuberculosis in clinical specimens by polymerase chain reaction and Gen-Probe Amplified Mycobacterium Tuberculosis Direct Test. *J Clin Microbiol* 1993;31:3270-4.

174. Bemer-Melchior P., Boudigueux V, Drugeon HB. Clinical validity of an automated DNA amplification system for diagnosis of pulmonary tuberculosis 5401. *Medecines et Maladies Infectieuses* 2000;30:253-61.

175. Cavusoglu C, Guneri S, Suntur M, Bilgic A. Clinical evaluation of the FASTPlaqueTB for the rapid diagnosis of pulmonary tuberculosis. *Turkish Journal of Medical Sciences* 2002;32:487-92.

176. Kang EY, Choi JA, Seo BK, Oh YW, Lee CK, Shim JJ. Utility of polymerase chain reaction for detecting Mycobacterium tuberculosis in specimens from percutaneous transthoracic needle aspiration. *Radiology* 2002;225:205-9.

177. Neu N. Diagnosis of pediatric tuberculosis in the modern era. *Pediatr Infect Dis J* 1999;18:122-6.

178. Delgado-Bolton RC, Fernandez-Perez C, Gonzalez-Mate A, Carreras JL. Meta-analysis of the performance of 18F-FDG PET in primary tumor detection in unknown primary tumors. *J Nucl Med* 2003;44:1301-14.

179. Bricker L, Garcia J, Henderson J, Mugford M, Neilson J, Roberts T *et al.* Ultrasound screening in pregnancy: a systematic review of the clinical effectiveness, cost-effectiveness and women's views. *Health Technol Assess* 2000;4(16):1-193.

180. Balk E, Ioannidis J, Salem D, Chew P, Lau J. Accuracy of biomarkers to diagnose acute cardiac ischemia in the emergency department: A meta-analysis. *Ann Emerg Med* 2001;37:478-94.

181. Buchanan A, Leese M. Detention of people with dangerous severe personality disorders: a systematic review. *The Lancet* 2001;358:1955-9.

182. Chapell R, Bruening W, Mitchell MD, Reston JT, Treadwell JR. Diagnosis and treatment of worker-related musculoskeletal disorders of the upper extremity. 2002.

183. Eden K, Mahon S, Helfand M. Screening high-risk populations for thyroid cancer. *Medical & Pediatric Oncology* 2001;36:583-91.

184. Flemons WW, Littner MR, Rowley JA, Gay P, Anderson WM, Hudgel DW *et al.* Home diagnosis of sleep apnea: a systematic review of the literature. An evidence review cosponsored by the American Academy of Sleep Medicine, the American College of Chest Physicians, and the American Thoracic Society. *Chest* 2003;124:1543-79.

185. Flobbe K, Nelemans PJ, Kessels AG, Beets GL, von Meyenfeldt MF. The role of ultrasonography as an adjunct to mammography in the detection of breast cancer: a systematic review. *Eur J Cancer* 2002;38:1044-50.

186. Gifford DR, Holloway RG, Vickrey BG. Systematic review of clinical prediction rules for neuroimaging in the evaluation of dementia. *Arch Intern Med* 2000;160:2855-62.

187. Gould MK, Kuschner WG, Rydzak CE, Maclean CC, Demas AN, Chan JK *et al.* Test performance of positron emission tomography and computed tomography for mediastinal staging in patients with non-small-cell lung cancer: a meta-analysis. *Ann Intern Med* 2003;139:879-92.

188. Gray M, Gold L, Burls A, Elley K. The effectiveness of toluidine blue dye as an adjunct to oral cancer screening in general dental practice. 2000.

189. Ioannidis J, Salem D, Chew P, Lau J. Accuracy and clinical effect of out-of-hospital electrocardiography in the diagnosis of acute cardiac ischemia: A meta-analysis. *Ann Emerg Med* 2001;37:461-70.

190. Koelemay M, Lijmer J, Stoker J, Legemate D, Bossuyt P. Magnetic resonance angiography for the evaluation of lower extremity arterial disease: a meta-analysis. *JAMA* 2001;285:1338-45.

191. Lysakowski C, Walder B, Costanza M, Tramer M. Transcranial Doppler versus angiography in patients with vasospasm due to a ruptured cerebral aneurysm - A systematic review. *Stroke* 2001;32:2292-8.

192. Medical Services Advisory Committee. Genetic test for fragile X syndrome. Assessment report 2002; MSAC application 1035

193. Nallamothu B, Saint S, Bielak L, Sonnad S, Peyser P, Rubenfire M *et al.* Electron-beam computed tomography in the diagnosis of coronary artery disease. *Arch Intern Med* 2001;161:833-8.

194. Patwardhan MB, McCrory DC, Matchar DB, Samsa GP, Rutschmann OT. Alzheimer disease: operating characteristics of PET. A meta-analysis. *Radiology* 2004;231:73-80.

195. Romagnuolo J, Bardou M, Rahme E, Joseph L, Reinhold C, Barkun AN. Magnetic resonance cholangiopancreatography: a meta-analysis of test performance in suspected biliary disease. *Ann Intern Med* 2003;139:547-57.

196. Sauerland S, Bouillon B, Rixen D, Raum MR, Koy T, Neugebauer EA. The reliability of clinical examination in detecting pelvic fractures in blunt trauma patients: a meta-analysis. *Archives of Orthopaedic and Trauma Surgery* 2004;124:123-8.

197. Sotiriadis A, Makrydimas G, Ioannidis JP. Diagnostic performance of intracardiac echogenic foci for Down syndrome: a meta-analysis. *Obstet Gynecol* 2003;101:1009-16.

198. Varonen H, Makela M, Savolainen S, Laara E, Hilden J. Comparison of ultrasound, radiography, and clinical examination in the diagnosis of acute maxillary sinusitis: a systematic review. *J Clin Epidemiol* 2000;53:940-8.

199. Rutjes AWS, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PMM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 2006;174.

200. Simmonds MC, Higgins JPT. Covariate heterogeneity in meta-analysis: Criteria for deciding between meta-regression and individual patient data. *Stat Med* 2007;26:2982-99.

201. Cochrane Handbook for Systematic Reviews of Diagnostic Tests. The Cochrane Library Chichester, UK: John Wiley & Sons, Ltd. 2008.

202. Stewart LA, Tierney JF. To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Eval Health Prof* 2002;25:76-97.

203. Koopman L, Van der Heiden G, Glasziou PP, Grobbee DE, Rovers MM. A systematic review of analytical methods used to study subgroups in (individual patient data) meta-analyses. *J Clin Epidemiol* 2007;60:1002-9.

204. Khan KS, Dinnes J, Kleijnen J. Systematic reviews to evaluate diagnostic tests. *Eur J Obstet Gyn* 2001;95:6-11.

205. Adams E, Flynn K. Positron emission tomography: descriptive analysis of experience with PET in VA. Boston, MA, USA: Health Services Research & Development Services; 1998. No. 55

206. Attia J, Margetts P, Guyatt G. Diagnosis of thyroid disease in hospitalized patients: a systematic review. *Arch Intern Med* 1999;159:658-65.

207. Bader J, Shugars D, Rozier G, Lohr K, Bonito A, Nelson J. Diagnosis and management of dental caries. Rockville MD: Agency for Healthcare Research and Quality; 2001. Evidence Report/Technology Assessment No. 36

208. Bafounta M, Beauchet A, Aegerter P, Saiag P. Is dermoscopy (epiluminescence microscopy) useful for the diagnosis of melanoma? Results of a meta-analysis using techniques adapted to the evaluation of diagnostic tests. *Arch Dermatol* 2001;137:1343-50.

209. Lau J, Ioannidis J, Balk E, Milch C, Terrin N, Chew P *et al.* Diagnosing acute cardiac ischemia in the emergency department: A systematic review of the accuracy and clinical effect of current technologies. *Ann Emerg Med* 2001;37:453-60.

210. Bastian LA, Nanda K, Hasselblad V, Simel DL. Diagnostic efficiency of home pregnancy test kits: a meta-analysis. *Arch Fam Med* 1998;7:465-9.

211. Bastian LA, Piscitelli JT. Is this patient pregnant? Can you reliably rule in or rule out early pregnancy by clinical examination? *JAMA* 1997;278:586-91.

212. Becker D, Philbrick J, Bachhuber T, Humphries J. D-dimer testing and acute venous thromboembolism. *Arch Intern Med* 1996;156:939-46.

213. Bell R, Petticrew M, Luengo S, Sheldon TA. Screening for ovarian cancer: a systematic review. *Health Technol Assess* 1998;2(2):1-84.

214. Berry E, Kelly S, Hutton J, Harris K, Roderick P, Boyce J *et al.* A systematic literature review of spiral and electron beam computed tomography: with particular reference to clinical applications in hepatic lesions, pulmonary embolus and coronary artery disease. *Health Technol Assess* 1999;3:1-118.

215. Kelly S, Berry E, Roderick P, Harris KM, Cullingworth J, Gathercole L *et al.* The identification of bias in studies of the diagnostic performance of imaging modalities. *Br J Radiol* 1997;70:1028-35.

216. Blakeley DD, Oddone EZ, Hasselblad V, Simel DL, Matchar DB. Noninvasive carotid artery testing. A meta-analytic review . *Ann Intern Med* 1995;122:360-7.

217. Bonis PA, Ioannidis JP, Cappelleri JC, Kaplan MM, Lau J. Correlation of biochemical response to interferon alfa with histological improvement in hepatitis C: a meta-analysis of diagnostic test characteristics. *Hepatology* 1997;26:1035-44.

218. Bradley KA, Boyd-Wickizer J, Powell SH, Burman ML. Alcohol screening questionnaires in women: a critical review. *JAMA* 1998;280:166-71.

219. Cabana MD, Alavi A, Berlin JA, Shea JA, Kim CK, Williams SV. Morphine-augmented hepatobiliary scintigraphy: a meta-analysis. *Nucl Med Commun* 1995;16:1068-71.

220. Campens D, Buntinx F. Selecting the best renal function tests. A meta-analysis of diagnostic studies. *Int J Technol Assess Health Care* 1997;13:343-56.

221. Conde-Agudelo A, Kafury-Goeta AC. Triple-marker test as screening for Down syndrome: a meta-analysis. *Obstet Gynecol Surv* 1998;53:369-76.

222. Cuzick J, Sasieni P, Davies P, Adams J, Normand C, Frater A *et al.* A systematic review of the role of human papillomavirus testing within a cervical screening programme. *Health Technol Assess* 1999;3:i-iv.

223. Da Silva O, Ohlsson A, Kenyon C. Accuracy of leukocyte indices and c-reactive protein for diagnosis of neonatal sepsis: a critical review. *Pediatr Infect Dis J* 1995;14:362-6.

224. Devous MD, Sr., Thisted RA, Morgan GF, Leroy RF, Rowe CC. SPECT brain imaging in epilepsy: a meta-analysis. *J Nucl Med* 1998;39:285-93.

225. Dinnes J, Moss S, Melia J, Blanks R, Song F, Kleijnen J. Effectiveness and cost-effectiveness of double reading of mammograms in breast cancer screening: findings of a systematic review. *Breast* 2001;10:455-63.

226. Divakaran T, Waugh J, Clark T, Khan K, Whittle M, Kilby M. Noninvasive techniques to detect fetal anemia due to red blood cell alloimmunization: A systematic review. *Obstet Gynecol* 2001;98:509-17.

227. Ebell MH, Flewelling D, Flynn CA. A systematic review of troponin T and I for diagnosing acute myocardial infarction. *J Fam Pract* 2000;49:550-6.

228. Fiellin D, Reid MC, O'Connor PG. Screening for alcohol problems in primary care. *Arch Intern Med* 2000;160:1977-89.

229. Fischer B, Mortensen J, Hojgaard L. Positron emission tomography in the diagnosis and staging of lung cancer: a systematic, quantitative review. *Lancet Oncology* 2001;2:659-66.

230. Fowlie PW, Schmidt B. Diagnostic tests for bacterial infection from birth to 90 days - a systematic review. *Arch Dis Child* 1998;78:F92-F98.

231. Garzon P, Eisenberg MJ. Functional testing for the detection of restenosis after percutaneous transluminal coronary angioplasty: a meta-analysis. *Can J Cardiol* 2001;17:41-8.

232. Gottlieb RH, Widjaja J, Tian L, Rubens DJ, Voci SL. Calf sonography for detecting deep venous thrombosis in symptomatic patients: experience and review of the literature. *J Clin Ultrasound* 1999;27:415-20.

233. Gronseth GS, Ashman EJ. Practice parameter: the usefulness of evoked potentials in identifying clinically silent lesions in patients with suspected multiple sclerosis (an evidence-based review): Report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology* 2000;54:1720-5.

234. Hallan S, Asberg A. The accuracy of C-reactive protein in diagnosing acute appendicitis. *Scand J Clin Lab Invest* 1997;57:373-80.

235. Heffner JE, Brown LK, Barbieri C, Deleo JM. Pleural fluid chemical-analysis in parapneumonic effusions: a metaanalysis. *A J Respir CritiCare Med* 1995;151:1700-8.

236. Heffner JE, Brown LK, Barbieri CA. Diagnostic value of tests that discriminate between exudative and transudative pleural effusions. *Chest* 1997;111:970-80.

237. Hider P, Nicholas B. The early detection and diagnosis of breast cancer: an update. *New Zealand Health Technol Assess Report* 1999;2:1-150.

238. Hobbs FD, Delaney BC, Fitzmaurice DA, Wilson S, Hyde CJ, Thorpe GH *et al.* A review of near patient testing in primary care. *Health Technol Assess* 1997;1:1-231.

239. Hooft L, Hoekstra O, Deville W, Lips P, Teule G, Boers M *et al.* Diagnostic accuracy of 18F-fluorodeoxyglucose positron emission tomography in the follow-up of papillary or follicular thyroid cancer. *J Clin Endocrinol Metab* 2001;86:3779-86.

240. Hrung JM, Sonad SS, Schwartz JS, Langlotz CP. Accuracy of MR imagining in the work-up of suspicious breast lesions: a diagnostic meta-analysis. *Acad Radiol* 1999;6:387-97.

241. Huicho L, Campos-Sanchez M, Alamo C. Metaanalysis of urine screening tests for determining the risk of urinary tract infection in children. *Pediatr Infect Dis J* 2002;21:1-11,88.

242. Huicho L, Campos M, Rivera J, Guerrant RL. Fecal screening tests in the approach to acute infectious diarrhea: a scientific overview. *Pediatr Infect Dis J* 1996;15:486-94.

243. Ioannidis J, Salem D, Chew P, Lau J. Accuracy of imaging technologies in the diagnosis of acute cardiac ischemia in the emergency department: A meta-analysis. *Ann Emerg Med* 2001;37:471-7.

244. Kinkel K, Kaji Y, Yu KK, Segal MR, Lu Y, Powell CB *et al.* Radiologic staging in patients with endometrial cancer: a meta-analysis. *Radiology* 1999;212:711-8.

245. Klompas M. Does this patient have an acute thoracic aortic dissection? *JAMA* 2002;287:2262-72.

246. Koelemay MJ, Denhartog D, Prins MH, Kromhout JG, Legemate DA, Jacobs MJ. Diagnosis of arterial disease of the lower extremities with duplex ultrasonography. *B J Surg* 1996;83:404-9.

247. Lacasse Y, Wong E, Guyatt GH, Cook DJ. Transthoracic needle aspiration biopsy for the diagnosis of localised pulmonary lesions: a meta-analysis. *Thorax* 1999;54:884-93.

248. Lau J, Zucker D, Engels E, Balk E, Barza M, Terrin N et al. Diagnosis and treatment of acute bacterial rhinosinusitis. Rockville, MD: Agency for Health Care Policy and Research; 1999. Evidence Report/Technology Assessment_ No. 9

249. Lederle FA, Simel DL. Does this patient have abdominal aortic aneurysm? *JAMA* 1999;281:77-82.

250. Liedberg J, Panmekiate S, Petersson A, Rohlin M. Evidence-based evaluation of three imaging methods for the temporomandibular disc. *Dentomaxillofacial Radiology* 1996;25:234-41.

251. Lindbaek M, Hjortdahl P. The clinical diagnosis of acute purulent sinusitis in general practice--a review. *Br J Gen Pract* 2002;52:491-5.

252. Littenberg B, Siegel A, Tosteson ANA, Mead T. Clinical efficacy of SPECT bone imaging for low back pain. *J Nucl Med* 1995;36:1707-13.

253. Mackenzie R, Palmer CR, Lomas DJ, Dixon AK. Magnetic resonance imaging of the knee: diagnostic performance statistics. *Clin Radiol* 1996;51:251-7.

254. Mango LJ, Radensky PW. Interactive neural-network-assisted screening: a clinical assessment. *Acta Cytologica* 1998;42:233-45.

255. Mayer J. Systematic review of the diagnostic accuracy of dermatoscopy in detecting malignant melanoma. *Med J Aust* 1997;167:206-10.

256. McGee S, Abernethy WB, Simel DL. Is this patient hypovolemic? *JAMA* 1999;281:1022-9.

257. McNaughton CM, MacDonald R, Wilt TJ. Diagnosis and treatment of chronic abacterial prostatitis: a systematic review. *Ann Intern Med* 2000;133:367-81.

258. Metlay JP, Kapoor WN, Fine MJ. Does this patient have community-acquired pneumonia? Diagnosing pneumonia by history and physical examination. *JAMA* 1997;278:1440-5.

259. Mol BW, Dijkman B, Wertheim P, Lijmer J, van der Veen F, Bossuyt PM. The accuracy of serum chlamydial antibodies in the diagnosis of tubal pathology: a meta-analysis. *Fertil Steril* 1997;67:1031-7.

260. Mol BW, Lijmer JG, Ankum WM, van der Veen F, Bossuyt PM. The accuracy of single serum progesterone measurement in the diagnosis of ectopic pregnancy: a meta-analysis. *Hum Reprod* 1998;13:3220-7.

261. Mol BW, Lijmer JG, van der Meulen J, Pajkrt E, Bilardo CM, Bossuyt PM. Effect of study design on the association between nuchal translucency measurement and Down syndrome. *Obstet Gynecol* 1999;94:864-9.

262. Australasian Cochrane Centre. Oto-acoustic emission audiometry. Canberra, Aus: Medicare Services Advisory Committee; 1999.

263. Mullins MD, Becker DM, Hagspiel KD, Philbrick J. The role of spiral volumetric computed tomography in the diagnosis of pulmonary embolism. *Arch Intern Med* 2000;160:293-8.

264. Muris JWM, Starmans R, Pop P, Crebolder HFJM, Knottnerus JA. Discriminant value of symptoms in patients with dyspepsia. *J Fam Pract* 1994;38:139-43.

265. Muris JW, Starmans R, Pop P, Crebolder HF, Knottnerus JA. The diagnostic value of symptoms for the identification of patients with an increased risk of colorectal disease. A criteria-based analysis. *Fam Pract* 1992;9:415-20.

266. Mustafa BO, Rathbun SW, Whitsett TL, Raskob GE. Sensitivity and specificity of ultrasonography in the diagnosis of upper extremity deep vein thrombosis: a systematic review. *Arch Intern Med* 2002;162:401-4.

267. Nuovo J, Melnikow J, Hutchison B, Paliescheskey M. Is cervicography a useful diagnostic test? a systematic overview of the literature. *J Am Board Fam Pract* 1997;10:390-7.

268. Pasternack I, Tuovinen E, Lohman M, Vehmas T, Malmivaara A. MR findings in humeral epicondylitis: a systematic review. *Acta Radiologica* 2001;42:434-40.

269. Patel SR, Wiese W, Patel SC, Ohl C, Byrd JC, Estrada CA. Systematic review of diagnostic tests for vaginal trichomoniasis. *Infect Dis Obstet Gynecol* 2000;8:248-57.

270. Pearl WS, Todd KH. Ultrasonography for the initial evaluation of blunt abdominal trauma: a review of prospective trials. *Ann Emerg Med* 1996;27:353-61.

271. Rao JK, Weinberger M, Oddone EZ, Allen NB, Landsman P, Feussner JR. The role of antineutrophil cytoplasmic antibody (c-ANCA) testing in the diagnosis of Wegener granulomatosis. A literature review and meta-analysis. *Ann Intern Med* 1995;123:925-32.

272. Rao G. Diagnostic yield of screening for type 2 diabetes in high-risk patients: a systematic review. *J Fam Pract* 1999;48:805-10.

273. Rathbun SW, Raskob GE, Whitsett TL. Sensitivity and specificity of helical computed tomography in the diagnosis of pulmonary embolism: a systematic review. *Ann Intern Med* 2000;132:227-32.

274. Reed WW, Byrd GS, Gates RH, Jr., Howard RS, Weaver MJ. Sputum gram's stain in community-acquired pneumococcal pneumonia. A meta-analysis. *West J Med* 1996;165:197-204.

275. Ross SD, Allen IE, Harrison KJ, Kvasz M, Connelly J, Sheinhait IA. Systematic review of the literature regarding the diagnosis of sleep apnea. Rockville, MD: Agency for Health Care Policy and Research; 1999. Evidence Report/Technology Assessment No. 1

276. Scheid D, McCarthy L, Lawler F, Hamm R, Reilly K. Screening for microalbuminuria to prevent nephropathy in patients with diabetes: a systematic review of the evidence. *J Fam Pract* 2001;50:661-8.

277. Schwimmer J, Essner R, Patel A, Jahan SA, Shepherd JE, Park K *et al.* A review of the literature for whole-body FDG PET in the management of patients with melanoma. *Q J Nucl Med* 2000;44:153-67.

278. Smith-Bindman R, Hosmer W, Feldstein V, Deeks J, Goldberg J. Second-trimester ultrasound to detect fetuses with down syndrome: a meta-analysis. *JAMA* 2001;285:1044-55.

279. Solomon DH, Simel DL, Bates DW, Katz JN, Schaffer JL. The rational clinical examination. Does this patient have a torn meniscus or ligament of the knee? Value of the physical examination. *JAMA* 2001;286:1610-20.

280. Sonnad SS, Langlotz CP, Schwartz JS. Accuracy of MR imaging for staging prostate cancer: a meta-analysis to examine the effect of technologic change. *Acad Radiol* 2001;8:149-57.

281. Spencer-Green G, Alter D, Welch HG. Test performance in systemic sclerosis: anti-centromere and Anti-Scl-70 antibodies . *Am J Med* 1997;103:242-8.

282. Stengel D, Bauwens K, Sehouli J, Porzsolt F, Rademacher G, Mutze S *et al.* Systematic review and meta-analysis of emergency ultrasonography for blunt abdominal trauma. *Br J Surg* 2001;88:901-12.

283. Swart P, Mol BW, van der Veen F, van Beurden M, Redekop WK, Bossuyt PM. The accuracy of hysterosalpingography in the diagnosis of tubal pathology: a meta-analysis. *Fertil Steril* 1995;64:486-91.

284. Taylor-Weetman K, Wake B, Hyde C. Comparison of panoramic and bitewing radiography for the detection of dental caries: a systematic review of diagnostic tests. Birmingham: University of Birmingham, Department of Public Health and Epidemiology; 2002.

285. van Beek EJ, Brouwers EM, Song B, Bongaerts AH, Oudkerk M. Lung scintigraphy and helical computed tomography for the diagnosis of pulmonary embolism: a meta-analysis. *Clin Appl Thromb Hemost* 2001;7:87-92.

286. van den Hoogen HM, Koes BW, van Eijk JT, Bouter LM. On the accuracy of history, physical examination, and erythrocyte sedimentation rate in diagnosing low back pain in general practice. *Spine* 1995;20:318-27.

287. van der Wurff P, Meyne W, Hagmeijer RH. Clinical tests of the sacroiliac joint. *Man Ther* 2000;5:89-96.

288. Vasbinder GB, Nelemans PJ, Kessels AG, Kroon AA, de Leeuw PW, van Engelshoven JM. Diagnostic tests for renal artery stenosis in patients suspected of having renovascular hypertension: a meta-analysis. *Ann Intern Med* 2001;135:401-11.

289. Vroomen PC, de Krom MC, Knottnerus JA. Diagnostic value of history and physical examination in patients suspected of sciatica due to disc herniation: a systematic review. *J Neurol* 1999;246:899-906.

290. Watson EJ, Templeton A, Russell I, Paavonen J, Mardh PA, Stary A *et al*. The accuracy and efficacy of screening tests for Chlamydia trachomatis: a systematic review. *J Med Microbiol* 2002;51:1021-31.

291. Wells PS, Lensing AW, Davidson BL, Prins MH, Hirsh J. Accuracy of ultrasound for the diagnosis of deep venous thrombosis in asymptomatic patients after orthopedic surgery. A meta-analysis. *Ann Intern Med* 1995;122:47-53.

292. White PM, Wardlaw JM, Easton V. Can noninvasive imaging accurately depict intracranial aneurysms? A systematic review. *Radiology* 2000;217:361-70.

293. Whited JD, Grichnik JM. Does this patient have a mole or a melanoma? *JAMA* 1998;279:696-701.

294. Wijnberger LD, Huisjes AJ, Voorbij HA, Franx A, Bruinse HW, Mol BW. The accuracy of lamellar body count and lecithin/sphingomyelin ratio in the prediction of neonatal respiratory distress syndrome: a meta-analysis. *Br J Obstet Gynecol* 2001;108:583-8.

295. Williams JWJ, Noel PH, Cordes JA, Ramirez G, Pignone M. Is this patient clinically depressed? *JAMA* 2002;287:1160-70.

296. Faron G, Boulvain M, Irion O, Barnard PM, Fraser WD. Prediction of preterm delivery by fetal fibronectin: a meta-analysis. *Obstet Gynecol* 1998;92:153-8.

297. Hofman PA, Nelemans P, Kemerink GJ, Wilmink JT. Value of radiological diagnosis of skull fracture in the management of mild head injury: meta-analysis. *J Neurol Neurosurg Psychiatry* 2000;68:416-22.

298. Revah A, Hannah ME, Sue AQ. Fetal fibronectin as a predictor of preterm birth: an overview. *Am J Perinatol* 1998;15:613-21.

299. Smith-Bindman R, Kerlikowske K, Feldstein VA, Subak L, Scheidler J, Segal M *et al*. Endovaginal ultrasound to exclude endometrial cancer and other endometrial abnormalities. *JAMA* 1998;280:1510-7.

300. Shingadia D, Novelli V. Diagnosis and treatment of tuberculosis in children.[erratum appears in Lancet Infect Dis. 2004 Apr;4(4):251 Note: Correction of dosage error.]. *Lancet Infect Dis* 2003;3:624-32.

301. Nelson LJ, Wells CD. Global epidemiology of childhood tuberculosis. *Int J Tuberc Lung Dis* 2004;8:636-47.

302. Davies PDO. *Clinical tuberculosis*. London: Arnold, 2003.

303. Marras TK, Daley CL. Epidemiology of human pulmonary infection with nontuberculous mycobacteria. *Clin Chest Med* 2002;23:553-67.

304. Henry MT, Inamdar L, O'Riordain D, Schweiger M, Watson JP. Nontuberculous mycobacteria in non-HIV patients: epidemiology, treatment and response. *Eur Respir J* 2004;23:741-6.

305. Shafer RW, Edlin BR. Tuberculosis in patients infected with human immunodeficiency virus: perspective on the past decade. *Clin Infect Dis* 1996;22:683-704.

306. Corbett EL, Watt CJ, Walker N, Maher D, Williams BG, Raviglione MC *et al.* The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. *Arch Intern Med* 2003;163:1009-21.

307. Narain JP, Lo YR. Epidemiology of HIV-TB in Asia. *Indian J Med Res* 2004;120:277-89.

308. Pelly T, Moore DA, Gilman R, Evans C. Recent tuberculosis advances in Latin America. *Curr Opin Infect Dis* 2004;17:397-403.

309. Feleke Y, Abdulkadir J, Aderaye G. Prevalence and clinical features of tuberculosis in Ethiopian diabetic patients. *East Afr Med J* 1999;76:361-4.

310. John GT, Shankar V, Abraham AM, Mukundan U, THomas PP, Jacob CK. Risk factors for post-transplant tuberculosis. *Kidney Int* 2001;60:1148-53.

311. CDC. Tuberculosis associated with blocking agents against tumor necrosis factor-alpha--California, 2002-2003. *MMWR Morb Mortal Wkly Rep* 2004;53:683-6.

312. American Thoracic Society Workshop. Rapid diagnostic tests for tuberculosis. What is the appropriate use? *Am J Respir Crit Care Med* 1997;155:1804-14.

313. Nucleic acid amplification tests for tuberculosis. *MMWR Morb Mortal Wkly Rep* 1996;45:950-2.

314. Miller WC. Bias in discrepant analysis: when two wrongs don't make a right. *J Clin Epidemiol* 1998;51:219-31.

315. Arimura M, Ohuchi T, Suzuki Y, Hishinuma A, Oikawa S, Sato J *et al.* Clinical significance of direct detection of Mycobacterium tuberculosis in respiratory specimens by polymerase chain reaction. *Dokkyo Journal of Medical Sciences* 1996;22:143-8.

316. Bennedsen J, Thomsen VO, Pfyffer GE, Funke G, Feldmann K, Beneke A *et al.* Utility of PCR in diagnosing pulmonary tuberculosis. *J Clin Microbiol* 1996;34:1407-11.

317. Bergmann JS, Woods GL. Clinical evaluation of the Roche AMPLICOR PCR Mycobacterium tuberculosis test for detection of M. tuberculosis in respiratory specimens. *J Clin Microbiol* 1996;34:1083-5.

318. Bergmann JS, Yuoh G, Fish G, Woods GL. Clinical evaluation of the enhanced Gen-Probe Amplified Mycobacterium Tuberculosis Direct Test for rapid diagnosis of tuberculosis in prison inmates. *J Clin Microbiol* 1999;37:1419-25.

319. Cartuyvels R, De Ridder C, Jonckheere S, Verbist L, Van Eldere J. Prospective clinical evaluation of Amplicor Mycobacterium tuberculosis PCR test as a screening method in a low-prevalence population. *J Clin Microbiol* 1996;34:2001-3.

320. Catanzaro A, Perry S, Clarridge JE, Dunbar S, Goodnight-White S, LoBue PA *et al.* The role of clinical suspicion in evaluating a new diagnostic test for active tuberculosis: results of a multicenter prospective trial. *JAMA* 2000;283:639-45.

321. Chin DP, Yajko DM, Hadley WK, Sanders CA, Nassos PS, Madej JJ *et al.* Clinical utility of a commercial test based on the polymerase chain reaction for detecting Mycobacterium tuberculosis in respiratory specimens. *Am J Respir Crit Care Med* 1995;151:1872-7.

322. Cohen RA, Muzaffar S, Schwartz D, Bashir S, Luke S, McGartland LP *et al.* Diagnosis of pulmonary tuberculosis using PCR assays on sputum collected within 24 hours of hospital admission. *Am J Respir Crit Care Med* 1998;157:156-61.

323. Gleason Beavis K, Lichty MB, Jungkind DL, Giger O. Evaluation of Amplicor PCR for direct detection of Mycobacterium tuberculosis from sputum specimens. *J Clin Microbiol* 1995;33:2582-6.

324. Lim TK, Gough A, Chin NK, Kumarasinghe G. Relationship between estimated pretest probability and accuracy of automated Mycobacterium tuberculosis assay in smear-negative pulmonary tuberculosis. *Chest* 2000;118:641-7.

325. Lim TK, Zhu D, Gough A, Lee KH, Kumarasinghe G. What is the optimal approach for using a direct amplification test in the routine diagnosis of pulmonary tuberculosis? A preliminary assessment. *Respirology* 2002;7:351-7.

326. Lockman S, Hone N, Kenyon TA, Mwasekaga M, Villauthapillai M, Creek T *et al.* Etiology of pulmonary infections in predominantly HIV-infected adults with suspected tuberculosis, Botswana. *Int J Tuberc Lung Dis* 2003;7:714-23.

327. Mitarai S, Kurashima A, Tamura A, Nagai H, Shishido H. Clinical evaluation of Amplicor Mycobacterium detection system for the diagnosis of pulmonary mycobacterial infection using sputum. *Tuberculosis (Edinb)* 2001;81:319-25.

328. Piersimoni C, Zitti P, Cimarelli ME, Nista D, De Sio G. Clinical utility of the Gen-Probe amplified Mycobacterium tuberculosis direct test compared with smear and culture for the diagnosis of pulmonary tuberculosis
5179. *Clin Microbiol Infect* 1998;4:442-6.

329. Reischl U, Lehn N, Wolf H, Naumann L. Clinical evaluation of the automated COBAS AMPLICOR MTB assay for testing respiratory and nonrespiratory specimens. *J Clin Microbiol* 1998;36:2853-60.

330. Se Thoe SY, Tay L, Sng EH. Evaluation of Amplicor- and IS6110-PCR for direct detection of Mycobacterium tuberculosis complex in Singapore. *Trop Med Int Health* 1997;2:1095-101.

331. Shim TS, Chi HS, Lee SD, Koh Y, Kim WS, Kim DS *et al.* Adequately washed bronchoscope does not induce false-positive amplification tests on bronchial aspirates in the diagnosis of pulmonary tuberculosis. *Chest* 2002;121:774-81.

332. Yee YC, Gough A, Kumarasinghe G, Lim TK. The pattern of utilisation and accuracy of a commercial nucleic acid amplification test for the rapid diagnosis of Mycobacterium tuberculosis in routine clinical practice. *Singapore Med J* 2002;43:415-20.

Related publications

Published

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

2. Dinnes J, Deeks JJ, Kunst H, Gibson A, Cummins E, Waugh N *et al.* A systematic review of diagnostic tests for the detection of tuberculosis. *Health Technol Assess* 2007;11.

In preparation

1. Dinnes J, Deeks JJ, Roderick P. The application of advanced methods of meta-analysis of diagnostic test accuracy: a case study

2. Dinnes J, Deeks JJ, Roderick P. An empirical comparison of methods of meta-analysis of diagnostic test accuracy – how biased are 'standard' SROC methods?